Estimating the Lyapunov Exponent of a Chaotic System With Nonparametric Regression

DANIEL F. McCAFFREY, STEPHEN ELLNER, A. RONALD GALLANT, and DOUGLAS W. NYCHKA*

We discuss procedures based on nonparametric regression for estimating the dominant Lyapunov Exponent λ_1 from time series data generated by a nonlinear autoregressive system with additive noise. For systems with bounded fluctuations, $\lambda_1 > 0$ is the defining feature of chaos. Thus our procedures can be used to examine time series data for evidence of chaotic dynamics. We show that a consistent estimator of the partial derivatives of the autoregression function can be used to obtain a consistent estimator of λ_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, and projection pursuit)—are presented for two deterministic chaotic systems. Local splines and 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 system suggest that the thin-plate spline and neural net regression methods also provide reliable values of the Lyapunov exponent in this case.

KEY WORDS: Dynamical systems; Neural networks; Nonlinear dynamics; Nonlinear time series models; Projection pursuit regression; Thin plate smoothing splines.

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, called *chaos*, is now recognized as ubiquitous in the nonlinear equations used to model various phenomena including fluid dynamics, chemical reactions, electrical circuits, physiological feedback mechanisms, and disease epidemics (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 data in such diverse fields as physics, geology, astronomy, neurobiology, ecology, and economics (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, Schaffer, Truty, Graser, and Olsen 1988; Kurths and Herzel 1987; Mayer-Kress 1986; Mpitsos, Creech, Cohan, and Mendelson 1988; Mpitsos, Burton, Creech, and Soinila 1988; Ramsay and Yuan 1989; Sugihara and May 1990). The available methods, developed over the last decade in theoretical physics (Schuster 1988; Mayer-Kress 1986), are based on calculating a few key quantities that characterize the dynamics, in particular fractal dimensions and Lyapunov exponents. These methods give reliable results if the data are abundant (thousands or tens of thousands of values), if measurement error is near 0, and if the data really come from a deterministic system. With limited data or a system subject to nonnegligible stochastic perturbations, the results may be incorrect or ambiguous (Ruelle 1990).

We present 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 that 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

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

with $x_t \in \mathbb{R}^1$ and $\{e_t\}$ a sequence of iid random variables. It is useful to express this system in terms of a state vector $\mathbf{X}_t = (x_t, x_{t-1}, \ldots, x_{t-d+1})'$, an error vector $\varepsilon_t = (e_t, 0, 0, \ldots, 0)'$ in $\mathbb{R}^d$, and a function $F: \mathbb{R}^d \mapsto \mathbb{R}^d$ such that

$$\mathbf{X}_t = F(\mathbf{X}_{t-1}) + \sigma\varepsilon_t. \tag{1.2}$$

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. The dominant Lyapunov exponent is defined as

$$\lambda_1 = \lim_{t\to\infty} \frac{1}{t} \log\|\mathbf{J}_{t-1} \cdot \mathbf{J}_{t-2} \cdots \mathbf{J}_0\|,$$

where $\mathbf{J}_t$ is the matrix of partial derivatives of the map F from (1.2) evaluated at $\mathbf{X}_t$ (i.e., $\mathbf{J}_t = DF(\mathbf{X}_t)$) and $\|\cdot\|$ denotes some $d \times d$ matrix norm. A rigorous development of this definition is given in Section 2. 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" that 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 infinites-

* Daniel F. McCaffrey is Associate Statistician, RAND Corporation, Santa Monica, CA 90407-2138. A. Ronald Gallant is Professor and Douglas W. Nychka is Associate Professor, Department of Statistics, and Stephen Ellner is Associate Professor, Biomathematics Program and Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. The authors acknowledge the support of National Science Foundation Grants DMS-8715756 and SES-8808015, North Carolina Agricultural Experiment Station Project NCO-6134, and the Institute for Advanced Studies, Hebrew University, Jerusalem.

Journal of the American Statistical Association
September 1992, Vol. 87, No. 419, Theory and Methods

imally small perturbations) is given by the dominant Lyapunov exponent λ_1. If λ_1 is negative rather than positive, then nearby trajectories are converging rather than diverging. Thus a common definition of chaos for deterministic systems is bounded fluctuations with $\lambda_1 > 0$ (Eckmann and Ruelle 1985). In systems with stochastic components $\lambda_1 > 0$ still indicates a sensitive dependence on initial conditions, meaning that the nonlinearity is contributing to the system's unpredictability.

Traditionally, chaos has referred only to purely deterministic systems and has been considered a distinct alternative to stochastic modeling (Farmer and Sidorwich 1988a). Eckmann and Ruelle (1985), however, define a system as chaotic if it exhibits sensitive dependence on initial conditions for all initial conditions. Only sensitive dependence on initial conditions distinguishes chaotic systems from nonchaotic systems, and any system with bounded fluctuations and sensitive dependence on initial conditions is chaotic, whether or not it is purely deterministic. Under this general definition of chaos, systems like (1.1) or (1.2) may be chaotic. For a detailed discussion on the relationship between chaos and randomness, see Wegman (1988).

In Section 2 we review the definition and properties of Lyapunov exponents for (1.1) and (1.2). In Section 3 we discuss estimation of λ_1 via nonparametric regression estimates of f, including consistency and rates of convergence. This establishes a framework for applying statistical analytic tools to an estimation problem that has been of concern for nearly a decade in various scientific disciplines. In Section 4 we survey some nonparametric regression methods with potential value for estimating f, and in Section 5 we report the results of a numerical study that applies our procedures to data from chaotic systems. We discuss these numerical results in Section 6.

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. Here we also review the methods currently used to estimate Lyapunov exponents and motivate our approach to the problem.

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

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

or the linear autoregressive model

$$x_t = \sum_{k=1}^{d} a_k x_{t-k} + e_t, \tag{2.1b}$$

where in both instances $x_0, x_1, \ldots, x_{d-1}$ are given. In state-space form

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \sigma\varepsilon_t, \tag{2.2}$$

where $\mathbf{A}$ is the $d \times d$ matrix with $[a_1, a_2, \ldots, a_d]$ in the top row, 1's on the subdiagonal, and 0's elsewhere. The characteristic polynomial for (2.2) is

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

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

If $\sigma = 0$, then the qualitative behavior of solutions to (2.2) is related to the magnitudes of the eigenvalues. Suppose for simplicity that the eigenvalues of $\mathbf{A}$ are distinct. Then for initial vector $\mathbf{X}_0 = \sum a_i v_i$, where v_i is an eigenvector corresponding to γ_i, the solution to (2.2) is $\mathbf{X}_t = \mathbf{A}^t\mathbf{X}_0 = \sum a_i \gamma_i^t v_i$. Hence

$$\|\mathbf{X}_t\|^{1/t} = \|\mathbf{A}^t\mathbf{X}_0\|^{1/t} \to \max\{|\gamma_i| : a_i \neq 0\} \quad \text{as } t \to \infty. \tag{2.4}$$

Because all norms on a finite-dimensional vector space are equivalent, (2.4) holds in any norm; hereafter, $\|\cdot\|$ denotes any matrix or vector norm that corresponds to the dimension of the matrix or vector. The Lyapunov exponents for (2.2) are the numbers $\lambda_i = \log|\gamma_i|$, arranged so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. In terms of the Lyapunov exponents, (2.4) says that

$$\frac{1}{t}\log\|\mathbf{X}_t\| = \frac{1}{t}\log\|\mathbf{A}^t\mathbf{X}_0\| \to \max\{\lambda_i : a_i \neq 0\} \quad \text{as } t \to \infty. \tag{2.5}$$

The Lyapunov exponents give the possible asymptotic rates of exponential increase (or decrease) for solutions of (2.2) with $\sigma = 0$. The asymptotic growth rate will be λ_1 for almost all initial vectors $\mathbf{X}_0$ (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 λ_1 also gives the growth rate for $\|\mathbf{A}^t\|$.

Consequently, the behavior of typical solutions depends on whether the dominant exponent λ_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 $\lambda_1 < 0$, then 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 $\lambda_1 < 0$, then solutions settle into stationary oscillations with bounded variance; whereas if $\lambda_1 > 0$, then 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, $\{\mathbf{A}_t\}_{t=0}^{\infty}$, and the behavior of the corresponding linear system, $\mathbf{X}_{t-1} = \mathbf{A}_t\mathbf{X}_t$.

Definition 2.1 (Lyapunov Exponents for the Linear System $\boldsymbol{X}_{t+1} = \boldsymbol{A}_t\boldsymbol{X}_t$). Let $\mathbf{H}_t = \mathbf{A}_{t-1} \cdots \mathbf{A}_0$, so that solutions to the system $\mathbf{X}_{t+1} = \mathbf{A}_t\mathbf{X}_t$ take the form $\mathbf{X}_t = \mathbf{H}_t\mathbf{X}_0$, for some $\mathbf{X}_0$. The Lyapunov exponent of a solution $\mathbf{X}_t$

$= \mathbf{H}_t\mathbf{X}_0$ is defined as

$$\lambda(\mathbf{X}_0) = \lim_{t\to\infty} \frac{1}{t} \log\|\mathbf{H}_t\mathbf{X}_0\|. \tag{2.6}$$

(See Arnold and Wihstutz 1986 for a more general discussion on this definition of Lyapunov exponents.) The values $\lambda(\mathbf{X}_0)$ generally will depend on both the initial vector $\mathbf{X}_0$ and the realization of the process $\{\mathbf{A}_t\}$. Given any realization of the $\{\mathbf{A}_t\}$, however, there exists at most $p \le d$ distinct values for $\lambda(\mathbf{X}_0)$. Furthermore, under very general conditions Oseledec's multiplicative ergodic theorem (Eckmann and Ruelle 1985) states: If the process $\{\mathbf{A}_t\}$ is ergodic, then, outside a set of measure 0, the set of possible values for $\lambda(\mathbf{X}_0)$ over all $\mathbf{X}_0$ is the same for all realizations of the sequence of matrices. Thus under these conditions we can define the spectrum of Lyapunov exponents $\lambda_p < \cdots < \lambda_1$ and with probability 1 $\lambda(\mathbf{X}_0) = \lambda_i$ for some $i = 1, \ldots, p$ and any $\mathbf{X}_0$.

Oseledec's multiplicative ergodic theorem also provides that the sets $E_i = \{\mathbf{X}_0 \in \mathbb{R}^d: \lambda(\mathbf{X}_0) \le \lambda_i\}$, $i = 1, \ldots, p$, form subspaces with

$$0 = E_{p+1} \subset E_p \subset \cdots \subset E_1 = \mathbb{R}^d \tag{2.7}$$

and $\lambda(\mathbf{X}_0) = \lambda_i$ iff $\mathbf{X}_0 \in E_i \backslash E_{i+1}$. As with the systems defined by (2.1) or (2.2), λ_1 gives the growth rate of "typical" solutions; that is $\lambda(\mathbf{X}_0) = \lambda_1$ for all $\mathbf{X}_0$ lying outside a $(d-1)$ or smaller dimensional subspace of $\mathbb{R}^d$.

We are, therefore, mainly concerned with the dominant exponent λ_1. The existence of λ_1 was first proved by Furstenberg and Kesten (1960), who provided an alternative formulation of the definition of λ_1: If $\{\mathbf{A}_t\}$ is stationary and ergodic and $\log^+\|\mathbf{A}_1\|$ has finite expectation ($\log^+ x = \max\{\log x, 0\}$), then

$$\lambda_1 \overset{\text{def}}{=} \lim_{t\to\infty} \frac{1}{t} \log\|\mathbf{H}_t\| \tag{2.8}$$

exists (as a random variable) and is constant with probability 1. Because this definition involves only the matrix $\mathbf{H}_t$, it is useful for computing λ_1.

The Furstenberg–Kesten result is now an immediate consequence of Kingman's subadditive ergodic theorem (Hall and Heyde 1980) applied to the subadditive process $x_{st} = \log\|\mathbf{A}_{t-1}\mathbf{A}_{t-2}\cdots\mathbf{A}_s\|$, $s < t$. Three axioms define a subadditive process; only one is relevant to our discussion: a subadditive process x_{st}, $s < t$, and s and t nonnegative integers satisfies

$$x_{su} \le x_{st} + x_{tu}, \tag{2.9}$$

whenever $s < t < u$. If equality holds in (2.9), then the process is additive. (For a complete discussion of additive and subadditive processes see Hall and Heyde 1980.)

For nonlinear dynamical systems the Lyapunov exponents are defined as follows.

Definition 2.2 (Lyapunov Exponents for the Nonlinear Dynamical System $\mathbf{X}_t = F(\mathbf{X}_{t-1}) + \sigma\varepsilon_t$). Let $\mathbf{J}_t = DF(\mathbf{X}_t)$, the Jacobian matrix of F at $\mathbf{X}_t$. The Lyapunov exponents for the nonlinear dynamical system, $\mathbf{X}_t = F(\mathbf{X}_{t-1}) + \sigma\varepsilon_t$, $\mathbf{X}_0$ given and $\mathbf{X}_t \in Q \subseteq \mathbb{R}^d$, are defined as

$$\lambda(\mathbf{X}_0, \mathbf{Y}_0) = \lim_{t\to\infty} \frac{1}{t} \log\|\mathbf{T}_t\mathbf{Y}_0\|, \tag{2.10}$$

where $\mathbf{T}_t\mathbf{Y}_0$ are solutions to the linear system

$$\mathbf{Y}_t = \mathbf{J}_t\mathbf{Y}_{t-1}. \tag{2.11}$$

That is, $\mathbf{T}_t = J_{t-1}\mathbf{J}_{t-2}\cdots\mathbf{J}_0$ or $\mathbf{T}_t = DF(\mathbf{X}_{t-1}) \times DF(\mathbf{X}_{t-2})\cdots DF(\mathbf{X}_0)$.

Because the Lyapunov exponents for the nonlinear system are defined to be the Lyapunov exponents of the tangent map system (2.11), the discussion following Definition 2.1 applies here as well. Under very general assumptions the process $\{\mathbf{J}_t\}$ satisfies the multiplicative ergodic theorem and $\lambda(\mathbf{X}_0, \mathbf{Y}_0)$ equals a constant, λ_1, for almost all $\mathbf{Y}_0$ (with respect to the Lebesgue measure on $\mathbb{R}^d$) with probability 1. Furthermore, under the same assumptions, the Furstenberg–Kesten type definition holds and $\lambda_1 = \lim_{t\to\infty}(1/t)\log\|\mathbf{T}_t\|$. The required assumptions are that the initial vector $\mathbf{X}_0$ is chosen according to an invariant measure for the Markov chain (1.2) and that the expectation of $\log^+\|DF(\mathbf{X})\|$, taken with respect to this invariant measure, is finite. See Chapter I of Kifer (1986) for a precise definition of invariance, and see Chapter III for the extension of Oseledec's multiplicative ergodic theorem to the tangent map system of a nonlinear dynamical system. Note that Kifer's discussion considers the general case where

$$\mathbf{X}_t = F_{t-1}(\mathbf{X}_{t-1}) \tag{2.12}$$

and $\{F_t\}_{t=0}^{\infty}$ is a sequence of random variables taking values in the space of differentiable functions from $Q \mapsto Q$.

The significance of Lyapunov exponents for nonlinear systems is their relationship to sensitive dependence on initial conditions, the defining characteristic of chaos. Let $\mathbf{X}_t^{(1)}$ and $\mathbf{X}_t^{(2)}$ be solutions of (1.2) that differ only in their initial values; that is,

$$\mathbf{X}_t^{(i)} = F(\mathbf{X}_{t-1}^{(i)}) + \sigma\varepsilon_t, \qquad \mathbf{X}_0^{(i)} \text{ given}, \quad \text{for } i = 1, 2. \tag{2.13}$$

If the initial perturbation $\mathbf{Y}_0 = (\mathbf{X}_0^{(1)} - \mathbf{X}_0^{(2)})$ is sufficiently small, then the subsequent separation $\mathbf{Y}_t = (\mathbf{X}_t^{(1)} - \mathbf{X}_t^{(2)})$ can be approximated by linearization,

$$\mathbf{Y}_t = F_{t-1}(\mathbf{X}_{t-1}^{(1)}) - F_{t-1}(\mathbf{X}_{t-1}^{(2)}) \approx \mathbf{J}_{t-1}\mathbf{Y}_{t-1}. \tag{2.14}$$

Thus the time evolution of a small perturbation to the state of the nonlinear system (1.2) is approximated by the linear system (2.11). In the formal limit of infinitesimally small $\mathbf{Y}_0$, the approximation becomes exact. Thus the Lyapunov exponents of the nonlinear system give the growth rate of local perturbations to the system. If $\lambda_1 > 0$, then almost all initially small perturbations grow exponentially over time and $\lambda_1 > 0$ indicates "sensitive dependence on initial conditions" in system (1.2). As we noted previously, this is a key indicator of chaotic dynamics and is responsible for the extreme unpredictability of chaotic systems.

3. ESTIMATING LYAPUNOV EXPONENTS

3.1 Methods for Estimating Lyapunov Exponents

There are two classes of methods in use for estimating λ_1 from experimental or observational data. "Direct" methods, proposed by Guckenheimer (1982) and successfully implemented by Wolf, Swift, Swinney, and Vastano (1985), are based on the assumption that $\|\mathbf{Y}_t\|$ in (2.14) will grow exponentially at rate λ_1. $\mathbf{Y}_0$ as defined previously cannot be observed, however, because our time series of data is only a single realization. Thus the growth rate of an infinitesimal perturbation is estimated by tracking the evolution of the difference between initially close observed data points. The available time series of data, $\{\mathbf{X}_t\}_{t=1}^N$, are searched for pairs of times (t_1, t_2) where $\|\mathbf{X}_{t_1} - \mathbf{X}_{t_2}\|$ is sufficiently small. Then the growth of $\tilde{\mathbf{Y}}_\tau = \|\mathbf{X}_{t_1+\tau} - \mathbf{X}_{t_2+\tau}\|$ is recorded until the trajectories diverge beyond some preset limit (or, for some preset time). The data are then searched for a replacement point $\mathbf{X}_{t_3}$ that is sufficiently near $\mathbf{X}_{t_1+\tau}$ and also sufficiently near the line connecting $\mathbf{X}_{t_1+\tau}$ and $\mathbf{X}_{t_2+\tau}$. The process is repeated, and the average divergence rate over the entire data set is the estimate of λ_1.

In "Jacobian" methods the data are used to estimate the Jacobians $DF(\mathbf{X}_t)$, and λ_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, Kamphorst, Ruelle, and Ciliberto (1986). To estimate $DF(\mathbf{X}_t)$, Eckmann et al. (1986) identified all data points $\mathbf{X}_s$ lying within some specified distance of $\mathbf{X}_t$. The approximation

$$\mathbf{X}_{s+1} - \mathbf{X}_{t+1} = F(\mathbf{X}_s) - F(\mathbf{X}_t) \approx DF(\mathbf{X}_t)(\mathbf{X}_s - \mathbf{X}_t)$$

justifies a linear model $\mathbf{X}_{s+1} - \mathbf{X}_{t+1} \approx \mathbf{A}_t(\mathbf{X}_s - \mathbf{X}_t)$. Assuming that sufficiently many $\mathbf{X}_s$ exist, then $\mathbf{A}_t$ can be determined from a least squares fit to differences. This least squares solution provides an estimate for $DF(\mathbf{X}_t)$, which is then substituted into equation (2.10). Under the conditions of Oseledec's multiplicative ergodic theorem, the Lyapunov spectrum corresponds to the square roots of the ordered eigenvalues of $\mathbf{T}_t'\mathbf{T}_t$ (Eckmann and Ruelle 1985); T' denotes the transpose of T. Eckmann et al. (1986) used iterative decompositions of the matrix $\mathbf{T}_t$ to calculate the estimated exponents as $\hat{\lambda}_i = (1/2N)\log(\nu_i)$, where ν_i is the ith largest eigenvalue of $\mathbf{T}_t'\mathbf{T}_t$.

In tests with simulated data from low-dimensional systems, the direct and linear Jacobian methods both appear 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-Kress 1986). Because the correct dimension is unknown, a common strategy is to increase the dimension until the estimates of λ_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) claimed 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" exponents is essentially *ad hoc* and requires much trial-and-error tuning of the free parameters.

The direct methods are less suited to mixed stochastic/nonlinear dynamics, however. It is easy to see that the direct method will be positively biased when $\sigma > 0$. To eliminate bias, the divergences must be fitted 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 $\mathbf{X}_t$ and $\mathbf{Y}_t$ dependence in the growth of $\mathbf{Y}_t$. Consequently, we believe that Jacobian methods are preferable for stochastic nonlinear systems, but must either incorporate a criterion for choosing the dimension or else use an estimator for $DF(\mathbf{X}_t)$ that is less sensitive to extraneous lags in the model. Procedures for fitting nonlinear time series models to chaotic data have been very successful in generating estimators of F (Abarbanel, Brown, and Kadtke 1989; Casdagli 1989; Farmer and Sidorowich 1987, 1988a, 1988b) despite *ad hoc* fitting methods. These successes are the motivation for our approach, in which the Jacobian estimates are derived from nonlinear estimation of F. Related approaches for deterministic systems were described by Briggs (1990) and Brown, Bryant, and Abarbanel (1991).

3.2 Estimating λ_1 from the Jacobians of the Map

In this section we consider the properties of Jacobian-method estimates for λ_1 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 $\hat{\mathbf{J}}_k$ denote the matrix obtained by substituting estimated for exact partial derivatives, $\hat{\mathbf{J}}_k = D\hat{F}(\mathbf{X}_k)$. We require that the estimate $\hat{\mathbf{J}}_k$ be consistent, as specifically stated in Assumption 1.

Assumption 1. There exists $\beta_N \to 0$ such that $\max_{1 \le t \le N} \|\hat{\mathbf{J}}_t - \mathbf{J}_t\| = \mathcal{O}_p(\beta_N)$ as $N \to \infty$.

From the theory of nonparametric regression, β_N can be expected to converge to 0 at a rate of the form $N^{-\delta}$, where δ depends on the form of the nonparametric estimate and the differentiability of f. For example, if $\hat{f}$ is the mth order thin-plate spline in d dimensions, f has m continuous partial derivatives, and $\sigma > 0$ in (1.1), then $\delta < \frac{1}{2}$. Moreover, in the case that noise is not present ($\sigma = 0$), one can expect that $\delta > 1$. A discussion of these rates is contained in Appendix A.

A second requirement is ergodicity, so that a single realization of the process represents (with probability 1) the "typical" behavior. This is stated formally in the following assumption.

Assumption 2. The Markov chain (1.2) has a unique invariant measure μ, and $\mathbf{X}_0$ is chosen by sampling from this

measure. The stationary process $\{\mathbf{J}_t\} = \{DF(\mathbf{X}_t)\}$ is ergodic, and $\int \log^+\|DF(x)\|\, d\mu(x) < +\infty$.

We also make two further assumptions on the behavior of the process $\{\mathbf{J}_t\}$.

Assumption 3. (a) $\liminf 1/t \log\|\mathbf{J}_t\mathbf{J}_{t-1}\cdots\mathbf{J}_0\| > -\infty$ with probability 1 and (b) $\sup_{t\geq 0}\|\mathbf{J}_t\| < \infty$ with probability 1.

Assumption 3a ensures that the system given in (1.2) has finite Lyapunov exponents. Assumption 3b is satisfied when DF is bounded over the support of $\{\mathbf{X}_t\}$. For example, if F has a bounded domain and continuous first partial derivatives, and if the noise also is bounded, then DF is bounded.

Under these assumptions the sequence of estimated Jacobians $\{\hat{\mathbf{J}}_t\}$, $1 \leq t \leq N$ can be used to obtain consistent estimates of λ_1. To simplify notation, for the remainder of this section we write λ for λ_1. It will be necessary to distinguish between the "sample size" N, which is the number of observations used to estimate $\hat{\mathbf{J}}_t$, and the "block length" M, which is the number of matrices $\hat{\mathbf{J}}_t$ used to estimate λ. Let $\lambda_M = (1/M)\log\|\mathbf{T}_M\|$ and $\gamma_M = \|\mathbf{T}_M\|^{1/M}$, where $\mathbf{T}_M = \mathbf{J}_{M-1}\mathbf{J}_{M-2}\cdots\mathbf{J}_0$. By (2.8) $\lambda = \lim_{M\to\infty}\lambda_M$ with probability 1. Let $\hat{\gamma}_M$, $\hat{\mathbf{T}}_M$ be the estimates obtained by using $\hat{\mathbf{J}}$'s in place of $\mathbf{J}$'s. Then $\hat{\lambda}_M = \log(\hat{\gamma}_M)$ is the Jacobian estimator for λ based on sample size N and block length M. Define γ by $\lambda = \log\gamma$. The main error bounds are summarized by the following theorem.

Theorem 3.1. Under Assumptions 1, 2, and 3,

$$\hat{\lambda}_M - \lambda_M = \mathcal{O}_p(\beta_N^{1/M}) \quad \text{as} \quad N, M \to \infty, \quad \text{such that} \quad M\beta_N \to 0. \tag{3.1}$$

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

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

Then

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

The proof of these results is given in Appendix B.

Although the rate of convergence given by either (3.1) or (3.2) is quite slow, these rates are sufficient to guarantee that the difference $\hat{\lambda}_M - \lambda_M$ converges to 0 in probability. Because λ_M converges almost surely to λ as $M \to \infty$, the implication of Theorem 3.1 is that $\hat{\lambda}_M$ provides a consistent estimate of λ.

For accurate estimation of λ, M should be taken as large as possible. Subadditivity of x_{st} (as defined in Section 2) has the unfortunate implication that $E(\lambda_M) \geq \lambda$ with strict inequality unless x_{st} is actually additive. Additivity is not to be expected in general, because x_{st} is additive only if $\|\mathbf{J}_r\mathbf{J}_{r-1}\cdots\mathbf{J}_k\|$ and $\prod_{i=k}^{r}\|\mathbf{J}_i\|$ 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 to guarantee that $|\hat{\lambda}_M - \lambda_M| \to 0$. In fact, if $\beta_N = \mathcal{O}(N^{-\delta})$ with $\delta < \frac{1}{2}$ as expected, then $M = \mathcal{O}(\log N)$ is necessary. But we conjecture that (3.1) and (3.2) are overly pessimistic and that the actual error is sufficiently small so that $M = o(N^{\delta})$ implies $\hat{\lambda}_M \to \lambda$.

Conjecture 3.1. Under Assumptions 1 and 2,

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

Rationale. The leading-order term in $\mathbf{T}_M - \hat{\mathbf{T}}_M$ (see equations B.1 and B.2 of Appendix B) is the sum of M products with the form

$$\mathbf{J}_M\mathbf{J}_{M-1}\cdots\mathbf{J}_{k+1}(\mathbf{J}_k - \hat{\mathbf{J}}_k)\mathbf{J}_{k-1}\cdots\mathbf{J}_1.$$

Because $\|\mathbf{J}_k - \hat{\mathbf{J}}_k\| = \mathcal{O}_p(\beta_N)$ by assumption and $\|\mathbf{J}_i\| \leq \alpha$ by definition of α, these terms together are $\mathcal{O}_p(M\alpha^{M-1}\beta_N)$, the rate obtained rigorously in the proof of Theorem 3.1. But except for a vanishingly small fraction of such products as $M \to \infty$, both k and $M - k$ will be large. Consequently, the asymptotic growth rates

$$\|\mathbf{J}_M\mathbf{J}_{M-1}\cdots\mathbf{J}_{k+1}\| \sim \gamma^{(M-k)}, \qquad \|\mathbf{J}_{k-1}\cdots\mathbf{J}_1\| \sim \gamma^{(k-1)} \tag{3.4}$$

should be sharper than the crude upper bounds used in the proof of (3.1), which have α in place of γ. Assuming that (3.4) applies simultaneously to all terms in (B.2) and repeating the arguments used to derive (B.4), we obtain

$$|\hat{\lambda}_M - \lambda_M| \leq \frac{\|\hat{\mathbf{T}}_M - \mathbf{T}_M\|/M}{\|\mathbf{T}_M\| - \|\hat{\mathbf{T}}_M - \mathbf{T}_M\|} \tag{3.5}$$

and hence

$$|\hat{\lambda}_M - \lambda_M| = \mathcal{O}_p\left(\frac{\beta_N}{\gamma - M\beta_N}\right). \tag{3.6}$$

The conjecture follows directly.

3.3 Simulations of Estimation Errors

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

$$x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2}, \tag{3.7}$$

and the block size was varied over the range $50 \leq M \leq 1{,}000$. At any given block size M, 1,000 replications were used to simulate the distribution of $|\hat{\lambda}_M - \lambda_M|$. A replicate consisted of selecting a new set of initial conditions for the Hénon system, iterating through transients, and then generating M data points. The corresponding $\mathbf{J}_t$ were calculated using the true derivatives of the generating equation (3.7). To guarantee compliance with Assumption 1 and hasten computation, we created the "estimation error" matrices $\mathbf{U}_t = \hat{\mathbf{J}}_t - \mathbf{J}_t$ rather than estimating the derivatives via a nonparametric regression estimate of the map. The $\hat{\mathbf{J}}_t$ then were calculated as $\mathbf{J}_t + \mathbf{U}_t$, and $\hat{\lambda}_M$ was determined using the "estimated" $\hat{\mathbf{J}}_t$. The $\mathbf{U}_t$ were simulated by choosing all entries independently from the uniform distribution on $[-\beta/2, \beta/2]$, where $\beta = 25/M^2$. Thus $\|\mathbf{U}_t\| \leq \beta$ for the L_2 matrix norm.

The right side of (3.6) is the leading term of an expansion of $|\hat{\lambda}_M - \lambda_M|$ in β_N; therefore, we compared our estimated error to $\beta_N/(\gamma - M\beta_N)$. Because (3.6) is based on conver-

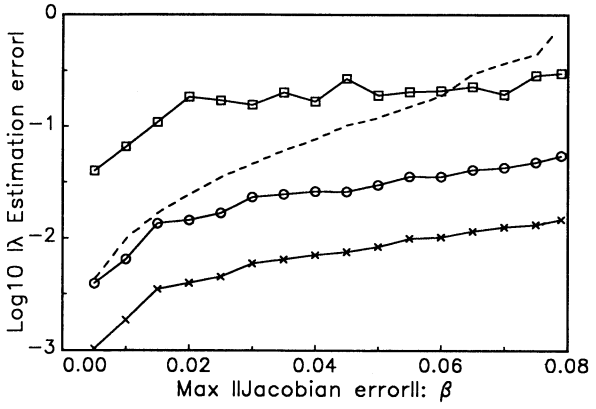

Figure 1. Comparison of the Conjectured Convergence Rate for $\hat{\lambda}_M$ With Simulations of the Hénon Map (3.7). The graph shows the rate from equation (3.6), $\beta_N/(\gamma - M\beta_N)$ (dashed line) and the mean (×), 95th percentile (○), and maximum (□) of the absolute error $|\hat{\lambda}_M - \lambda_M|$ over 1,000 simulations for each (M, β) pair. The block length was $M = 5/\sqrt{\beta}$, so that $\beta \to 0$ corresponds to N, $M \to \infty$ with $M\beta_N = 5\sqrt{\beta} \to 0$. For each replicate, the initial value (x_0, x_1) was chosen at random from a file of 20,000 values on the Hénon map's attractor. Error matrices U_t were chosen independently for each replicate at each (M, β) value. See text for further details.

gence in probability rather than on almost-sure convergence, we compared (3.6) to the mean, the 95th percentile, and the maximum of $|\hat{\lambda}_M - \lambda_M|$ over the 1,000 replicated values of λ_M and $\hat{\lambda}_M$. The bound $\beta_N/(\gamma - M\beta_N)$ is well above the mean error and mostly well above the 95th percentile. But the maximum error was larger than the 95th percentile by an order of magnitude and does not appear to adhere to the bound provided by (3.6), suggesting that the relevant mode of convergence is in probability and not almost sure. Note that the mean, 95th percentile, and maximum all appear to scale as β_N.

These theoretical results suggest that $M \ll N$ may be necessary to obtain a consistent estimate of λ_M. This raises the issue of optimal block length for estimating λ. In practice one presumably would obtain $\hat{\lambda}_M$'s for each of the roughly N/M blocks and use their average as an estimate of λ. Let $\lfloor N/M \rfloor$ denote the integer part of N/M. For $n = 1, \ldots, \lfloor N/M \rfloor$, let $\hat{\lambda}_{M,n}$ denote the value of $\hat{\lambda}_M$ found using the nth block of estimated Jacobian matrices; that is,

$$\hat{\lambda}_{M,n} = (1/M)\log\|\hat{\mathbf{J}}_{nM-1}\hat{\mathbf{J}}_{nM-2}\cdots\hat{\mathbf{J}}_{(n-1)M+1}\hat{\mathbf{J}}_{(n-1)M}\|.$$

The error $(\hat{\lambda}_{M,n} - \lambda)$ for block n consists of

1. block bias: $b(M) = E(\lambda_M) - \lambda$
2. block error: $\nu(n) = \lambda_{M,n} - E(\lambda_M)$
3. estimation error: $\varepsilon(M, n) = \hat{\lambda}_{M,n} - \lambda_{M,n}$,

where $E(\cdot)$ refers to expectation taken with respect to the stationary measure for the true system and $\lambda_{M,n}$ are defined analogously to $\hat{\lambda}_{M,n}$ with the true Jacobians rather than with the estimates. Recall that $\mathbf{J}_t = DF(\mathbf{X}_t)$; hence $\{\mathbf{J}_t\}_{t=0}^{\infty}$, $\lambda_{M,n}$, and $\lambda_M = \lambda_{M,1}$ are all random variables, even when $\sigma = 0$. Also, by the stationarity of Assumption 2, $E(\lambda_{M,n}) = E(\lambda_M)$ for $n = 1, \ldots, \lfloor N/M \rfloor$.

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 for these components can only be conjectured. Under Assumption 2 the most reasonable hypothesis for the block bias is:

Conjecture 3.2. $E(\lambda_M) - \lambda = \mathcal{O}(1/M)$. (3.8)

This is known to hold for positive matrices that satisfy a mixing condition (Heyde 1985), but the Jacobians for a nonlinear system with bounded trajectories cannot be positive everywhere. Under Assumption 2 the only reasonable conjecture for the block error is:

Conjecture 3.3. Std Dev$(\lambda_M) = \mathcal{O}(1/\sqrt{M})$. (3.9)

The rationale for (3.9) is that λ_M can be written as $1/M\sum_{j=1}^{M}\phi(\mathbf{X}_j, S_j)$, where $S_j = \mathbf{T}_{j-1}/\|\mathbf{T}_{j-1}\|$ (Furstenburg and Kesten (1960), and in our case $\phi(\mathbf{X}, S) = \log\|DF(\mathbf{X})S\|$. Because $(\mathbf{X}_j, S_j)$ is a Markov chain, central limit behavior is expected for λ_M. Such results are available for independent random matrices (Bougerol and Lacroix 1986) and for positive matrices under mixing assumptions (Heyde and Cohen 1985). It generally will be reasonable to assume that $\{\mathbf{X}_t\}$ is a Harris-recurrent chain; hence the Jacobian matrices $\{\mathbf{J}_t\}$ will be strong-mixing (Athreya and Pantula 1986).

Both (3.8) and (3.9) are consistent with simulation results for the Hénon map (3.7) and also for the model

$$x_t = x_{t-1} + 10.5\left[\frac{0.2x_{t-5}}{1 + (x_{t-5})^{10}} - 0.1x_{t-1}\right]. \quad (3.10)$$

Equation (3.10) 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 Hénon map are summarized in Figure 2; the results for (3.10) were very similar and thus are not reported. Linear regression fits of the simulated values for $E(\lambda_M) - \lambda$ to $1/M$ and Std Dev(λ_M) to $1/\sqrt{M}$ had $r^2 > .98$ in all cases.

When the block size is smaller than the data series length, several estimates of λ 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 λ 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 presumably would 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 (3.9), it appears that when the nonparametric regression procedure behaves as expected (i.e., $\beta_N \approx N^{-\delta}$, $\delta < \frac{1}{2}$), then either the block bias or block error component of the overall error will have the slowest asymptotic convergence rate, depending on the rate at which $M \to \infty$. 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

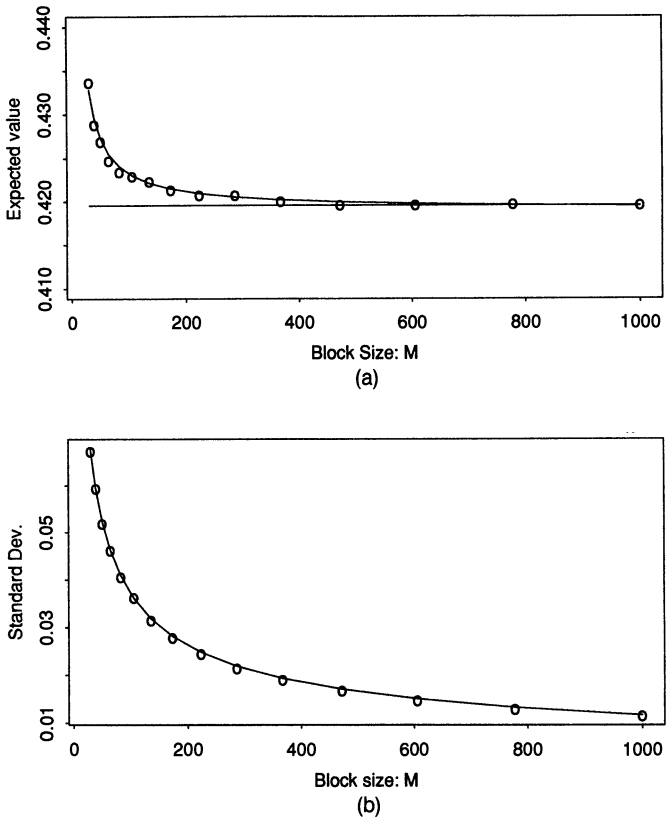

Figure 2. Rates of Convergence of the Block Bias and Block Error for the Hénon Map (3.7). For each block length M, 10,000 separate calculations of λ_M were made. (a) Average λ_M vs. M; the solid line was fit by linear regression of average $\lambda_M - \lambda$ on $1/M$. The actual exponent λ in this case is .418. (b) Standard deviation of λ_M vs. M; the solid line was fit by linear regression of the standard deviation on $1/\sqrt{M}$. For both regressions $r^2 > .98$.

four different methods for estimating the map that were used in a simulation study of our estimator for λ_1. 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 $\mathbf{X}$. 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 ($\sigma = 0$), and 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 as an efficient and accurate way for approximating complicated functions. Splines are defined fundamentally, as the solution to a variational problem. For example, consider the one-dimensional curve fitting problem based on (x_t, y_t), $1 \le t \le N$ where, in analogy to (1.1), $y_t = f(x_t) + \sigma e_t$. For a fixed value of $\mu > 0$, let

$$\mathcal{J}(h) = \int_{\mathbb{R}^1} [h^{(2)}(u)]^2 \, du. \tag{4.1}$$

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

$$\mathcal{L}(h) = \sum_{t=1}^{N} [\mathbf{Y}_t - h(\mathbf{X}_t)]^2 + \mu \mathcal{J}(h) \tag{4.2}$$

over all h such that $\mathcal{J}(h) < \infty$. 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 μ controls the relative weight between good data fitting 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 $\mu \to 0$. This limit also will be the solution to minimizing $\mathcal{J}(h)$ subject to the interpolation constraint $\mathbf{Y}_t = h(x_t)$, $1 \le t \le N$. The quality of the resulting spline estimate depends strongly on the choice of μ, and it is good practice to investigate the sensitivity of the estimate to this parameter.

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

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

Unlike the one-dimensional roughness measure, $\mathcal{J}_{m,d}$ may involve derivatives of higher order than second degree. This generalization is necessary to guarantee that the spline will be consistent for higher dimensions. Although $\mathcal{J}_{m,d}$ has a complicated form, the mixed partials enter in such a way that this curvature measure is invariant to rotation of the coordinate axes. Replacing $\mathcal{J}$ by $\mathcal{J}_{m,d}$ in (4.2), a thin-plate spline is defined as the minimizer of $\mathcal{L}$ over all h such that $\mathcal{J}_{m,d}(h) < \infty$. The same interpolation properties hold when $\mu \to 0$.

Although a spline function is defined abstractly as the solution to a minimization problem, it is readily computable. The solution will be a linear combination of the $\binom{d+m-1}{d}$ monomials of degree less than m and a set of N radial basis functions (Wahba 1990). The coefficients for the spline are found by solving a system of N linear equations. 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 local estimates of the Jacobian of the map. The estimate at any particular point $\mathbf{X}_t$ was calculated using only the L nearest state vectors with respect to Euclidian distance. There are two reasons for considering a local estimate rather than fitting a single spline to the entire time 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. Local estimates, however, do not alleviate the data sparseness problems associated with using multivariate splines on high-dimensional data. In fact local estimates may accentuate these problems, because the effective sample size has been reduced significantly.

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 $\{\phi_j(\mathbf{U})\}$, $1 \le j \le \binom{d+m-1}{d}$, denote the set of all monomials of degree less than m. Let $\{\mathbf{U}_k\}$ for $1 \le k \le K$ be a set of vectors in $\mathbb{R}^d$ and consider the radial basis

$$\psi_k(\mathbf{X}) = \|\mathbf{X} - \mathbf{U}_k\|^{2m-d} \quad \text{for } d \text{ odd}$$
$$\|\mathbf{X} - \mathbf{U}_k\|^{2m-d} \log(\|\mathbf{X} - \mathbf{U}_k\|^{2m-d}) \quad \text{for } d \text{ even.}$$

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

$$\hat{f}(\mathbf{X}) = \sum_j d_j \phi_j(\mathbf{X}) + \sum_k c_k \psi_k(\mathbf{X}), \tag{4.3}$$

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

4.3 Neural Nets

The use of neural nets to approximate the map of a chaotic process was 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 Farber (1987). We used a single hidden-layer, feedforward neural network. The functional form for this configuration is

$$\hat{f}(\mathbf{X}) = \sum_{j=1}^{K} \beta_j G(\gamma_j' \mathbf{X} + \mu_j),$$

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. Compared to the preceding functional approximations, the neural net form is not sensitive to increasing d. Although the lengths of the vectors γ_j increase, 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, but the quality of a neural network approximation is influenced by the number of logistic functions in the sum. The results in Gallant and White (1992), however, suggest that performance may not be sensitive to the choice of K provided that the saturation ratio, $K(d + 2)/N$, is below $1/30$ and $500 \le N < 5{,}000$. For smaller samples with high noise levels, the performance of neural nets in Lyapunov exponent estimation may be sensitive to the choice of K (Nychka, Ellner, McCaffrey, and Gallant 1992).

4.4 Projection Pursuit

Projection pursuit approximation is an ambitious estimation scheme that combines the projective power of neural nets with the flexibility of nonparametric curve 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.

Let $\alpha \in \mathbb{R}^d$ and $\|\alpha\| = 1$; then $(\alpha'\mathbf{X})\alpha$ defines the projection of $\mathbf{X}$ onto the one-dimensional subspace spanned by α. Given K such projections indexed by the vectors $\{\alpha_k\}$, $1 \le k \le K$, the projection pursuit approximation has the form

$$h(\mathbf{X}, \alpha_1, \ldots, \alpha_K) = \sum_{k=1}^{K} G_k(\alpha_k'\mathbf{X}). \tag{4.4}$$

Unlike the neural net representation, the G_k (known as ridge functions) do not follow a rigid parametric form and are estimated. In their original approach Friedman and Stuetzle estimated the ridge functions using a univariate "supersmoother," but we used cubic smoothing splines. Accordingly, the G_k, $k = 1, 2, \ldots, K$ are defined as the solutions to minimizing

$$\sum_{t=1}^{N} [\mathbf{Y}_t - h(\mathbf{X}_t, \alpha_1, \ldots, \alpha_K)]^2 + \mu \sum_{k=1}^{K} \mathcal{J}(G_k) \tag{4.5}$$

over all G_k such that $\mathcal{J}(G_k) < \infty$ and the function $\mathcal{J}$ is given by (4.1). In this manner any given set of projections imply a particular approximant, $\hat{h}$. A best set of projections is estimated by minimizing

$$\sum_{t=1}^{N} [y_t - \hat{h}(\mathbf{X}_t, \alpha_1, \ldots, \alpha_K)]^2 \tag{4.6}$$

over all $\{\alpha_k \in \mathbb{R}^d: 1 \le k \le K\}$.

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 $k \ne j$ are known. Given α_j, one can determine G_j based on fitting a one-dimensional smoothing spline to the regression data $\{\alpha_j'\mathbf{X}_t, y_t - \sum_{k \ne j} G_k(\alpha_k'\mathbf{X}_t)]$, $1 \le t \le N$. 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 Tibshirani 1989), and the resulting limit from backfitting will be the minimizer to (4.6). The complication for computing the projection pursuit approximation, however, is that the projections also are optimized in the backfitting steps. A description of the algorithm used here can be found in Nychka (1989).

4.5 Prediction Error of the Estimated Map

One measure of the accuracy of the map estimate is prediction error. Let $\{x_s\}$, $-d \le s \le S$, 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

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

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

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

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

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 (Eubank 1987; Wahba 1990). One problem with this norm is that it is not strong enough to imply convergence of the Lyapunov exponent estimate; convergence of ISE does not generally imply that $\hat{\mathbf{J}}$ converges to the Jacobian matrix of F. Conditions such that the rate of convergence of ISE to 0 implies the rate of convergence of $\hat{\mathbf{J}}_t$ to $\mathbf{J}_t$ were set forth in Wolfinger (1989).

A practical concern is that "tuning" parameters of the map estimates, if selected by minimizing $\hat{\sigma}^2$, may not yield the best values for estimating λ. For example, the embedding dimension usually is 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 should decrease to 0 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 previous remarks, this may not be the appropriate choice for using the map to estimate λ; one reason for our numerical study was to investigate the relationship between the best choice for d and the dimension that minimizes 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 2,000–2,500 and no measurement error, we evaluated the performance of the four approximation methods described in Section 4 (local spline, radial basis, neural net, and projection pursuit) for two simple chaotic systems (Hénon and Rössler). The Hénon system is quadratic and thus can be fitted exactly by radial basis regression; but because an exact fit would not be encountered in practice, we did not use a radial basis on the Hénon system. 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 $\hat{\lambda}$ and still obtain a consistent estimate. To keep the simulation design simple, this is what was done. The estimates of λ for these cases are summarized in Table 1 by the number of lags (d) used in the approximation.

A second set of simulations was run to investigate the sensitivity of the estimates when a random component is present. Attention was restricted to the Hénon 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 ($N = 2{,}000$, $M = 50, 100, 500, 2{,}000$). The results of these simulations are reported in Table 2. A more useful summary of these results is given in Figure 3, in which boxplots of the estimates of λ 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 in this numerical study. The section ends with an overview of the simulation results.

5.1 Hénon and Rössler Systems

The Hénon map is given at (3.7); it is evident that it has the form of (1.1). The largest Lyapunov exponent is approximately $.418 \pm .001$ (Wolf, Swift, Swinney, and Vastano 1985, p. 289, in which exponents were calculated using $\log_2$ rather than the natural log).

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

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

These equations were numerically integrated by fourth-order Runge–Kutta 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's embedding theorem (Takens 1981) indicates that for d sufficiently large there exists f: $\mathbb{R}^d \to \mathbb{R}$ such that $x_t = f(x_{t-1}, \ldots, x_{t-d-1})$. Thus the Rössler time series also follows (1.1). Although the form of f is unknown, the value of λ is known to be approximately .04505 (see Wolf et al. 1985, p. 289).

To generate a sample from either of these systems, 5,000 values were calculated based on a randomly selected starting value and then discarded. The next 20,000 values are taken

Table 1. Estimated Lyapunov Exponents for the Hénon and Rössler Systems Without Noise

		Hénon system[c]				
		d				
Map estimate	*N*	*1*	*2*	*3*	*4*	*5*
Local spline[a]	2,500	5.7602 (.042)	.4188[e] (.005)	.0750 (.011)	−.0251 (.013)	.0259 (.013)
Neural net	2,000	.1147	.4106	.4227	.4236[e]	—
Projection pursuit	2,000	—	.4163	.4058	.4026[e]	—

		Rössler system[d]						
		d						
Map estimate	*N*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Local spline[b]	2,500	7.1229 (.055)	.0992 (.004)	.0461[e] (.002)	1.7099 (.011)	1.567 (.21)	—	—
Radial basis	2,000	—	.0629	.7778[e]	10.24	10.26	—	—
Neural net	2,000	—	.0010	.1272	.6940	.0482	.0414	.0466[e]
Projection pursuit	2,000	—	—	.0966[e]	.0146	−.2792	−.0640	—

NOTE: The value of m for the local spline estimates was the smallest integer such that $2m > d$. For the radial basis function estimates 200 basis functions were used. N is the length of the data series, and d is the dimension of the model.

[a] Average of five estimates with standard deviation.

[b] Average of ten estimates with standard deviation.

[c] Correct value of λ is approximately .418 (Wolf et al. 1985, p. 289).

[d] Correct value of λ is approximately .04505 (Wolf et al. 1985, p. 289).

[e] Estimate of λ_1 that corresponds to the minimum expected prediction error, $\hat{\sigma}^2$.

to be points on the attractor of the map. To create a sample of size N, the starting value t_0 is randomly selected in the range $[1, 20{,}000 - N]$; the resulting index serves as the starting point of the time series.

For the second simulation study, the random component added to each iteration of the Hénon 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 Hénon system to become unstable. From a qualitative standpoint, the shape of the attractor for this noisy system had the same overall shape as did the strange attractor for the deterministic system. Of course the fractal properties of these two sets are very different, and because the invariant measure for the noisy system differs from the deterministic case the Lyapunov exponents also will differ. For the noisy system a simulation using 200 samples of 20,000 noisy Hénon data points yielded $\lambda = .408 \pm .009$. That is, for each sample the true Jacobian was evaluated at each noisy data point and λ was calculated using these matrices; the average of the exponent from these 200 trials was .408, with a standard error of .009.

Table 2. Estimated Lyapunov Exponents as a Function of Block Size M for the Hénon Map With Noise

		M				
Map estimate	*d*	*50*	*100*	*500*	*2,000*	*20,000*
Local spline[a]	2	.421 (.015)	.426 (.014)	.417 (.015)	.416 (.015)	
Neural net[b]	2	.416 (.020)	.412 (.020)	.408 (.020)	.408 (.019)	
Exact map[c]	2	.415 (.010)	.414 (.009)	.409 (.009)	.408 (.009)	.408 (.009)
Local spline	5	.417 (.009)	.431 (.009)	.406 (.008)	.404 (.008)	
Neural net	5	.417 (.007)	.411 (.009)	.406 (.009)	.405 (.009)	

NOTE: Each Lyapunov exponent estimate is the average of the exponents obtained from N/M disjoint blocks of the data series. The data series consisted of $N = 2{,}000$ values for $M \leq 2{,}000$, and $N = 20{,}000$ for $M = 20{,}000$; d is the dimension of the model.

[a] Average of 14 estimates with standard deviation.

[b] Average of 16 estimates with standard deviation.

[c] 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 $B = N/M$.) The correct value of λ is approximately .408.

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 $d = 4, 5$; for each d we used the smallest value of m such that $2m > d$. Because no noise is present, $\mu = 0$ and thus $\hat{f}(\mathbf{X}_{t-1})$ is equal to x_t. To achieve computational efficiency, $\hat{\sigma}^2$ was computed using within-sample cross-validation (Bates, Lindstrom, Wahba, and Yandell 1986). Let $\bar{x}_t$ denote the prediction of x_t based on the time series where x_t has been omitted. In this situation,

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

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 1 correspond to the value of m that minimized $\hat{\sigma}^2$ for a given d.

The neural net estimates were computed using $K = 7$. The value 7 was chosen because the saturation ratio complied with Gallant and White's (1989) findings and because this value provided stable estimates in preliminary experiments. As is typical in nonlinear least squares estimation problems, the minimization is not always an automatic procedure.

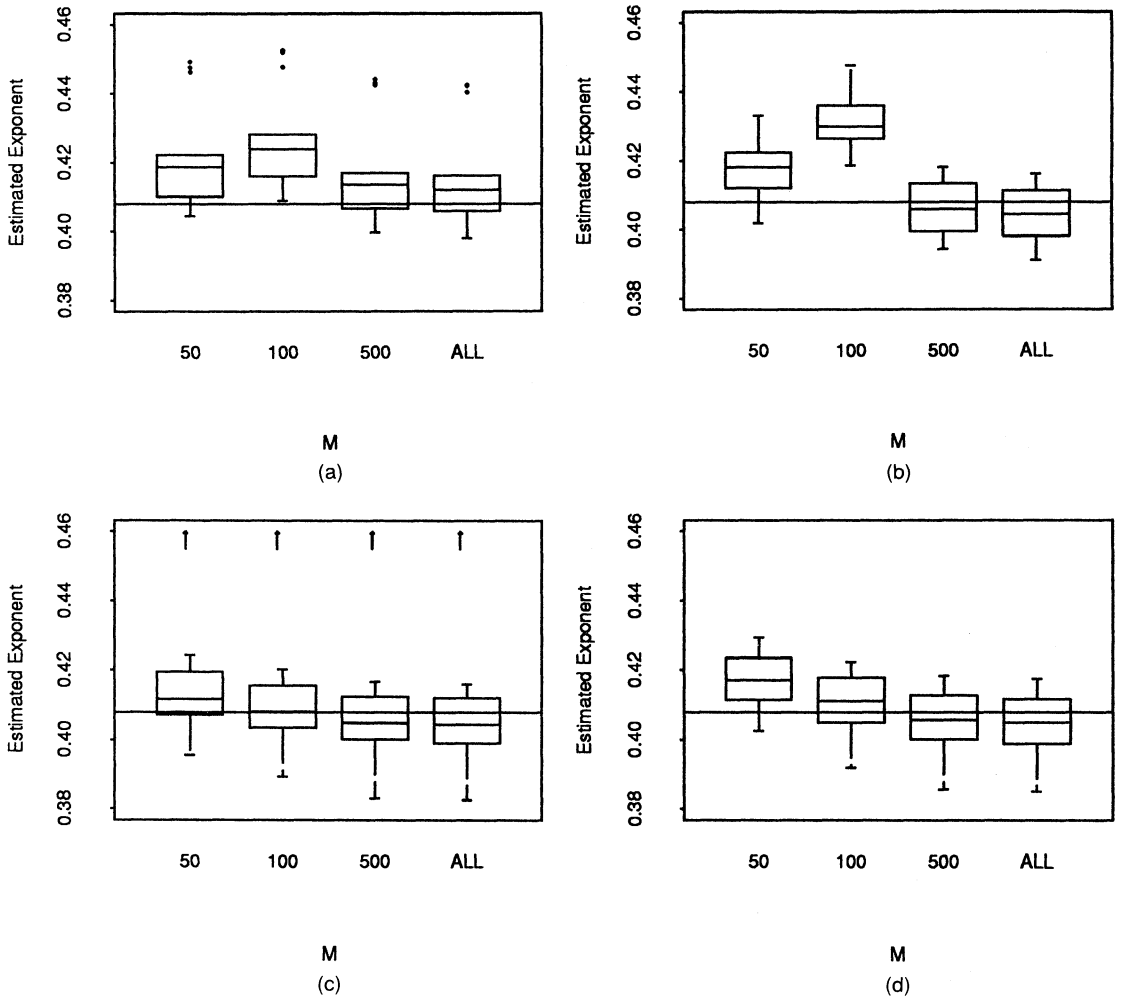

Figure 3. Distribution of Lyapunov Exponent Estimates From the Noisy Hénon System. Boxplots summarize the distribution of estimates as a function of method (local spline, neural net), the embedding dimension (d = 2, 5) and the block length (M). The horizontal line shows the true value of λ_1 for this system. The boxplots are based on 14 observations for the local splines and 16 for the neural nets. The four arrows in the lower left plot indicate four points out of bounds. (a) Spline, d = 2; (b) Spline, d = 5; (c) neural net, d = 2; (d) neural net, d = 5.

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, μ. For the Hénon map the estimate was computed for the 20 combinations of $K = 2, \ldots, 6$ and $\log(\mu) = -6, -10, -14, -18$. For the Rössler system the estimate was computed for the 24 combinations of $K = 2, \ldots, 9$ 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 used only 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 λ, the sequence of $\hat{\lambda}$ 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, provided that one uses the value for d that minimizes the expected value of $\hat{\sigma}^2$. Unlike for the neural net estimates, increasing the embedding dimension degrades the accuracy of $\hat{\lambda}$. Finally, it should be noted that in comparison with these two methods, the projection pursuit and radial basis functions yielded poor estimates of λ for the Rössler 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 Hénon System

The distribution of Lyapunov exponent estimates for a noisy Hénon system are summarized by Table 2. Figure 3 gives a graphical display of the simulation results using boxplots. As a reference, estimates based on the true Jacobians also are reported. For all the cases except local splines with $d = 5$ and $M = 100$, $\hat{\lambda}$ estimates λ_M well. The estimates have a small bias relative to the standard deviation; the variability of these estimates is comparable to the estimates when the Jacobian is known. This particular system has a large block

bias relative to the variability of $\hat{\lambda}_M$. Therefore, although the nonparametric regression methods yield accurate estimates of λ_M for small M they differ from the actual Lyapunov exponent λ. 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 λ.

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 that 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 used successfully 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 Hénon system, where f is a simple polynomial, but it is surprising for the Rössler 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 Rössler 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 objective 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 still would not be able to reproduce the ridge functions 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 terminated before a global minimum was reached.

The simulation results for a noisy system, though very limited, have promise. Both methods yield accurate estimates of λ, provided that the block size is chosen correctly ($M = N = 2{,}000$). Unlike the results for the deterministic Hénon system, 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 to 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 $M \to \infty$ 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 that this is a good strategy for other noisy systems, however.

The instability of the Hénon system with added noise is not surpirising. Initial values that evolve into chaotic solutions for the Hénon system are contained in a bounded subset of $\mathbb{R}^2$; other initial values cause the system to diverge. The added noise forced the Hénon system out of the stable region and into the set of unstable solutions. Most likely many other deterministic chaotic systems will exhibit instability in the presence of noise. The level of noise necessary to destabilize a chaotic system will depend on the stability of the deterministic system and will be very system-dependent.

The numerical results suggest 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 is determining a better data-based estimate 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 λ not based on local splines, $\hat{\sigma}^2$ was calculated *out of sample.* That is, the predictions were made at each point in a large auxiliary data set that is distinct from the data used in estimating the map. Unfortunately, this approach is not feasible if the amount of data is limited, and within-sample estimates need further study.

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 A: CONVERGENCE RATES FOR DERIVATIVE ESTIMATES

This Appendix gives a brief justification of the convergence rates for thin-plate splines. We first consider the case with noise. From Cox (1984), if $\hat{f}$ is an mth order thin-plate smoothing spline estimate of f, then under suitable conditions on the smoothing parameter and the limiting density of the state vectors $\{\mathbf{X}_t\}$, $1 \le t \le N$, we have

$$E\|\hat{f}_n - f\|_p^2 = \mathcal{O}(N^{-\delta}),$$

with $\delta = (m-p)/(2m+d)$. Here expectation is taken with respect to the random component, $\|\cdot\|_p$ denotes the usual norm for the

pth order Sobolev–Hilbert space $W_2^p[\Omega]$, where $\Omega \subset \mathbb{R}^d$ contains the support of the invariant density of the state vectors. (See Adams [1975] for a discussion of Sobolev spaces.) Also, note that by Chebychev's inequality it follows that

$$\|\hat{f}_n - f\|_p^2 = \mathcal{O}_p(N^{-\delta}).$$

Now consider the norm

$$|g|_k = \max_{1 \le j \le k} \max_{\alpha_1 + \cdots + \alpha_d = j} \sup_{\mathbf{X} \in \Omega} \left| \frac{\partial^j}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} g(\mathbf{X}) \right|.$$

From the Sobolev embedding theorem (Adams 1975, theorem 5.4(8)), there is a $K < \infty$ such that $|g|_k \le K\|g\|_{k+s}$ for all $g \in W_2^{k+s}[\Omega]$, provided that $2s > d$. In particular for $k = 1$, $|h|_1 \le K\|h\|_p$ for $p = 2 + \lfloor d/2 \rfloor$, where $\lfloor d/2 \rfloor$ denotes the integer part of $d/2$. Finally, noting that there is a $K < \infty$ such that $\max_{1 \le t \le N} \|\hat{\mathbf{J}}_t - \mathbf{J}_t\| \le K|\hat{f}_n - f|_1$, β_N may be assumed to have the rate $N^{-\delta}$ with $\delta = (m - 2 - \lfloor d/2 \rfloor)/(2m + d)$. Note that for all values of m, $\delta < \frac{1}{2}$.

Convergence rates for β_N in the deterministic case can be inferred from the interpolation results in Madych and Potter (1985). Paraphrasing their Theorem 2, let $G = \{\mathbf{X}_t\}$, $1 \le t \le N$, and $h_N = \max\{\text{distance } (\mathbf{X}, G \cup \partial\Omega): \mathbf{X} \in \Omega\}$, where $\partial\Omega$ denotes the boundary of Ω. If $2m \ge d$, then there is a constant $C < \infty$ such that

$$\|\hat{f}_n - f\|_p \le C h_N^{m-p} \|f\|_m.$$

Now setting $p = 2 + \lfloor d/2 \rfloor$, it follows from the previously mentioned embedding theorem that $\beta_N = \mathcal{O}(h_N^{m-p})$. It remains to characterize the maximum distance between a point in Ω and the N state vectors. If the state vectors form a regular grid on Ω, then $h_N \approx N^{-1/d}$. Under these circumstances for m sufficiently large—that is, ($m > d + 2 + \lfloor d/2 \rfloor$)—one may obtain a convergence rate for β_N that is faster than N^{-1}. But for points associated with the trajectory of a dynamic system, the rate on h_N is not clear. One reasonable conjecture is that $h_N \approx N^{-1/d^*}$, where d^* is the fractal dimension of the attracting set for the system. It also should be noted that the set Ω comprising the attractor for the system may not be regular and may violate assumptions such as those of Theorem 5.4 of Adams (1975).

APPENDIX B: PROOF OF THEOREM 3.1

Let $\alpha = \sup_{k \le 1} \|\mathbf{J}_k\|$ and $B_N = \max_{1 \le k \le N} \|\hat{\mathbf{J}}_k - \mathbf{J}_k\|$. Note that by Assumption 1, $B_N = \mathcal{O}_p(\beta_N)$. Also let $\mathbf{U}_t = \hat{\mathbf{J}}_t - \mathbf{J}_t$; then

$$\hat{\mathbf{T}}_M - \mathbf{T}_M = (\mathbf{J}_M + \mathbf{U}_M)(\mathbf{J}_{M-1} + \mathbf{U}_{M-1}) \cdots (\mathbf{J}_1 + \mathbf{U}_1) - \mathbf{J}_M \mathbf{J}_{M-1} \cdots \mathbf{J}_1. \quad \text{(B.1)}$$

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

$$\hat{\mathbf{T}}_M - \mathbf{T}_M = \sum_{\omega \in \mathcal{J}, |\omega| < M} A_M(\omega) A_{M-1}(\omega) \cdots A_1(\omega). \quad \text{(B.2)}$$

Hence, using the assumed bounds on $\|\mathbf{J}_k\|$ and $\|\mathbf{U}_k\|$,

$$\begin{aligned} \|\hat{\mathbf{T}}_M - \mathbf{T}_M\| &\le \sum_{\omega \in \mathcal{J}, |\omega| < M} \alpha^{|\omega|} B_N^{(M - |\omega|)} \\ &= \sum_{k=0}^{M-1} \binom{M}{k} \alpha^k B_N^{M-k} = (\alpha + B_N)^M - \alpha^M \\ &= \alpha^M[(1 + (B_N/\alpha))^M - 1] \\ &\le \alpha^M(e^{MB_N/\alpha} - 1) \\ &= \mathcal{O}_p(M\alpha^{M-1}\beta_N). \end{aligned} \quad \text{(B.3)}$$

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

$$|\hat{\gamma}_M - \gamma_M| \le \|\hat{\mathbf{T}}_M - \mathbf{T}_M\|^{1/M} = \mathcal{O}_p((\alpha^{M-1} M \beta_N)^{1/M}).$$

Assumption 3b provides that α is almost surely finite, which implies that $\alpha^{(M-1)/M} = \mathcal{O}_p(1)$. Thus

$$|\hat{\gamma}_M - \gamma_M| = \mathcal{O}_p(\beta_N^{1/M}).$$

Under Assumption 2, the conditions of Kifer's Theorem III.1.1 (Kifer 1986, p. 90) are met and λ_M converges almost surely to λ. Furthermore, by Assumption 3a, $lim\ 1/M \log\|\mathbf{T}_M\| > -\infty$, which implies that $\lambda > -\infty$. Therefore, outside of an exceptional set whose probability converges to 0 as $M, N \to \infty$, $\hat{\gamma}_M$ and γ_M are bounded away from 0. By the mean value theorem

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

where γ^* lies between $\hat{\gamma}_M$ and γ_M; with this bound, (3.1) now follows.

Now (3.2) will be considered. From the hypotheses on β_N and (B.3), $\|\hat{\mathbf{T}}_M - \mathbf{T}_M\| = o_p(\rho^{M-1})$. It follows from the Furstenberg–Kesten theorem that $\|\mathbf{T}_M\|\rho^{-M} \to +\infty$ with probability 1. From this fact it follows that $\|\hat{\mathbf{T}}_M - \mathbf{T}_M\|/\|\mathbf{T}_M\| \xrightarrow{p} 0$ in probability and, by the triangle inequality, $\|\hat{\mathbf{T}}_M\|/\|\mathbf{T}_M\| \xrightarrow{p} 1$. Using the relation $|\log(b) - \log(a)| \le |b - a|/\min(b, a)$,

$$\begin{aligned} M|\hat{\lambda}_M - \lambda_M| &= |\log\|\hat{\mathbf{T}}_M\| - \log\|\mathbf{T}_M\|| \\ &\le \left| \frac{\|\hat{\mathbf{T}}_M\| - \|\mathbf{T}_M\|}{\|\mathbf{T}_M\| \mathbb{X}_M} \right|, \end{aligned} \quad \text{(B.4)}$$

where $\mathbb{X}_M = \min(\|\hat{\mathbf{T}}_M\|/\|\mathbf{T}_M\|, 1)$. From the previous remarks it follows that $\mathbb{X}_M \xrightarrow{p} 1$, and thus

$$|\hat{\lambda}_M - \lambda_M| \le \frac{\|\hat{\mathbf{T}}_M - \mathbf{T}_M\|}{M\|\mathbf{T}_M\|}(1 + o_p(1)) \quad \text{as} \quad M \to \infty.$$

The result now follows immediately from this bound and (B.3).

[*Received July 1990. Revised September 1991.*]

REFERENCES

Abarbanel, H. D. I., Brown, R., and Kadtke, J. B. (1989), "Prediction and System Identification in Chaotic Nonlinear Systems: Time Series With Broadband Spectra," *Physics Letters A,* 138, 401–408.

Adams, R. A. (1975), *Sobolev Spaces,* New York: Academic Press.

Albano, A. M., Abraham, N. B., deGuzman, G. C., Tarroja, M. F. H., Bandy, D. K., Goggia, R. S., Rapp, P. E., Zimmerman, I. D., Greenbaum, N. N., and Bashore, T. R. (1986), "Lasers and Brains: Complex Systems With Low-Dimensional Attractors," in *Dimensions and Entropies in Chaotic Systems,* ed. G. Mayer-Kress, New York: Springer-Verlag, pp. 231–240.

Arnold, L., and Wihstutz, V. (1986), "Lyapunov Exponents: A Survey," in *Lyapunov Exponents* (Lecture Notes in Mathematics 1186), ed. L. Arnold and V. Wihstutz, New York: Springer-Verlag, pp. 1–26.

Athreya, K. B., and Pantula, S. B. (1986), "Mixing Properties of Harris Chains and Autoregressive Processes," *Journal of Applied Probability,* 23, 880–892.

Babloyantz, A., and Destexhe, A. (1986), "Low-Dimensional Chaos in an Instance of Epilepsy," *Proceedings of the National Academy of Science,* 53, 3513–3517.

Bates, D., Lindstrom, M., Wahba, G., and Yandell, B. (1986), "GCVPACK-Routines for Generalized Cross-Validation," Technical Report 775, University of Wisconsin–Madison, Dept. of Statistics.

Bougerol, P., and Lacroix, J. (1986) *Products of Random Matrices with Application to Schrodinger Operators,* Boston: Birkhäuser.

Brandstater, A., Swift, J., Swinney, H. L., Wolf, A., Farmer, J. D., Jen, E., and Crutchfield, P. J., (1983), "Low-Dimensional Chaos in a Hydrodynamic System," *Physical Review Letters,* 51, 1442–1445.

Brandstater, A., and Swinney, H. L. (1987), "Strange Attractors in Weakly Turbulent Couette–Taylor Flow," *Physical Review A,* 35, 2207–2220.

Briggs, K. (1991), "An Improved Method for Estimating Lyapunov Exponents of Chaotic Systems," *Physics Letters A,* 151, 27–32.

Brock, W. A., and Sayers, C. L. (1988), "Is the Business Cycle Characterized by Deterministic Chaos?," *Journal of Monetary Economics,* 22, 71–90.

Brown, R., Bryant, P., and Abarbanel, H. D. I. (1991), "Computing the Lyapunov Spectrum of a Dynamical System From an Observed Time Series," *Physical Review A,* 43, 2787–2805.

Buchler, J. R., and Eichorn, H. (eds.) (1987), *Chaotic Phenomena in Astrophysics* (Annals of the New York Academy of Sciences, Vol. 497), New York: New York Academy of Sciences.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics,* 17, 453–510.

Casdagli, M. (1989), "Nonlinear Prediction of Chaotic Time Series," *Physica,* 35D, 335–356.

Cohen, J. E., Kesten, H., and Newman, C. M. (eds.) (1986), *Random Matrices and Their Applications* (Contemporary Mathematics Vol. 50), Providence RI: American Mathematical Society.

Cox, D. D. (1984), "Multivariate Smoothing Splines," *SIAM Journal of Numerical Analysis,* 21, 789–813.

Eckmann, J.-P., and Ruelle, D. (1985), "Ergodic Theory of Chaos and Strange Attractors," *Reviews of Modern Physics,* 57, 617–656.

Eckmann, J.-P., Kamphorst, S. O., Ruelle, D., and Ciliberto, S. (1986), "Liapunov Exponents From Time Series," *Physical Review A,* 34, 4971–4979.

Eubank, E. (1987), *Spline Smoothing and Nonparametric Regression,* New York: Marcel Dekker.

Farmer, J. D., and Sidorowich, J. J. (1987), "Predicting Chaotic Time Series," *Physical Review Letters,* 59, 845–848.

——— (1988a), "Exploiting Chaos to Predict the Future and Reduce Noise," in *Evolution, Learning, and Cognition,* ed. Y. C. Lee, Singapore: World Scientific Press, p. 277.

——— (1988b), "Predicting Chaotic Dynamics," in *Dynamic Patterns in Complex Systems,* eds. J. A. S. Kelso, A. J. Mandell, and M. F. Schlesinger, Singapore: World Scientific, pp. 265–292.

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

Furstenberg, H., and Kesten, H. (1960), "Products of Random Matrices," *The Annals of Mathematical Statistics,* 31, 457–469.

Gallant, A. R., and White, H. (1992), "On Learning the Derivatives of an Unknown Mapping With Multilayer Feedforward Networks," *Neural Networks,* 5, 129–138.

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

Guckenheimer, J. (1982), "Noise in Chaotic Systems," *Nature,* 298, 358–361.

Guckenheimer, J., and Buzyna, G. (1983), "Dimension Measurements for Geostrophic Turbulence," *Physical Review Letters,* 51, 1438–1441.

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and its Application,* New York: Academic Press.

Heyde, C. C. (1985), "An Asymptotic Representation for Products of Random Matrices," *Stochastic Processes and their Applications,* 20, 307–314.

Heyde, C. C., and Cohen, J. E. (1985), "Confidence Intervals for Demographic Projections Based on Products of Random Matrices," *Theoretical Population Biology,* 27, 120–153.

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

Kot, M., Schaffer, W. M., Truty, G. L., Graser, D. J., and Olsen, L. F. (1988), "Changing Criteria for Imposing Order," *Ecological Modelling,* 43, 75–110.

Kurths, J., and Herzel, H. (1987), "An Attractor in a Solar Time Series," *Physica,* 25D, 165–172.

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

Madych, W. R., and Potler, E. H. (1985), "An Estimate for Multivariate Interpolation," *Journal of Approximation Theory,* 43, 132–139.

May, R. M. (1987), "Nonlinearities and Complex Behavior in Simple Ecological and Epidemiological Models," in *Perspectives in Biological Dynamics and Theoretical Medicine, Annals of the New York Academy of Sciences* (Vol. 504), eds. S. H. Koslow, A. J. Mandell, and M. F. Schlesinger, New York: New York Academy of Sciences, pp. 1–15.

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

Moon, F. C. (1987), *Chaotic Vibrations: An Introduction for Applied Scientists and Engineers,* New York: John Wiley.

Mpitsos, G. J., Creech, H. C., Cohan, C. S., and Mendelson, M. (1988), "Variability and Chaos: Neurointegrative Principles in Self-Organization of Motor Patterns," in *Dynamic Patterns in Complex Systems,* eds. J. A. S. Kelso, A. J. Mandell, and M. F. Schlesinger, Singapore: World Scientific, pp. 162–190.

Mpitsos, G. J., Burton, R. M., Jr., Creech, H. C., and Soinila, S. O. (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 *Proceedings of the 20th Symposium on the Interface,* eds. E. Wegman, D. Grant, and J. Miller, Alexandria, VA: American Statistical Association, pp. 731–736.

Nychka, D. W., Ellner, S., McCaffrey, D., and Gallant, A. R. (1992), "Finding Chaos in Noisy Systems," *Journal of the Royal Statistical Society,* Ser. B, 54, 399–426.

Olsen, L. F., and Degn, H. (1985), "Chaos in Biological Systems," *Quarterly Review of Biophysics,* 18, 165–225.

Ramsay, J. B., and Yuan, H.-J. (1989), "Bias and Error Bars in Dimension Calculations and Their Evaluation in Some Simple Models," *Physics Letters A,* 134, 287–297.

Ruelle, D. (1990), "Deterministic Chaos: The Science and the Fiction," *Proceedings of the Royal Society A,* 427, 241–248.

Schuster, H. G. (1988), *Deterministic Chaos: An Introduction,* (2nd rev. ed.), Weinheim: VCH Verlagsgesellschaft.

Sugihara, G., and May, R. M. (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 *Dynamical Systems and Turbulence, Warwick 1980,* (Lecture Notes in Mathematics, 898), eds. D. Rand and L. S. Young, New York: Springer-Verlag, pp. 366–381.

Vastano, J. E., and Kostelich, E. J. (1986), "Comparison of Algorithms for Determining Lyapunov Exponents From Experimental Data," in *Dimensions and Entropies in Chaotic Systems,* ed. G. Mayer-Kress, New York: Springer-Verlag, pp. 100–107.

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

Wegman, E. J. (1988), "On Randomness, Determinism and Computability," *Journal of Statistical Planning and Inference,* 10, 279–294.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985), "Determining Lyapunov Exponents From a Time Series," *Physica,* 16D, 285–315.

Wolfinger, R. D. (1989), "Rates on Convergence and Asymptotic Normality in Semi-Nonparametric Regression," unpublished Ph.D. dissertation, North Carolina State University, Dept. of Statistics.