Fields Institute Communications
Volume 11, 1997

Estimating Local Lyapunov Exponents

Zhan-Qian Lu
National Center for Atmospheric Research
P.O. Box 3000
Boulder, CO, USA 80307-3000

Richard L. Smith
Statistical Laboratory
Cambridge University, UK
and
Department of Statistics
University of North Carolina
Chapel Hill, NC, USA 27599-3260

Abstract. Lyapunov exponents are among the principal measures of the chaotic behavior of a dynamical system. However, they are difficult to estimate when the only information available is a sample of data from trajectories of the system. *Local* Lyapunov exponents are easier to estimate and, in some contexts, a more useful measure of the predictability of a system than global Lyapunov exponents. In this paper we propose a method of estimating the dynamical system and its local Lyapunov exponents using local linear and local quadratic regression. Approximations to the bias and variance of such estimators are quoted, and we illustrate the results with a number of examples.

1 Introduction

In recent years there have been many attempts to apply dynamical systems theory to understand the long-term chaotic behavior of non-linear time series. The pioneering book by Tong [1990] was followed by a number of collections of papers, such as Tong and Smith [1992], Drazin and King [1992] and Grenfell et al. [1994]; see also Isham [1993] and Jensen [1993] for reviews of this very active area.

One major theme of this research has been the attempt to estimate dynamical invariants such as the (fractal) dimension of an attractor, and entropy and information measures. This is useful for two reasons: first because they are interesting measures in their own right of the long-term behavior of a dynamical system, and second

1991 *Mathematics Subject Classification.* Primary 58F13, 62M20; Secondary 62G07.

This paper is based in part on the first author's PhD thesis at the University of North Carolina, Chapel Hill. The first author was supported by the NCAR Geophysical Statistics Project, sponsored by NSF grant DMS-9312686. The second author was supported in part by NSF grant DMS-9205112.

because they form the basis for tests to distinguish between chaos and randomness. Of all these measures, perhaps the most widely used is the correlation dimension, introduced by Grassberger and Procaccia [1983] and, as much as anything, popular because there is a relatively simple algorithm to estimate it, at least in its simplest form. L.A. Smith [1988], R.L. Smith [1992] and Wolff [1994] are among the authors to have studied the statistical properties of the Grassberger-Procaccia procedure and related estimators and tests. Recent work by Cutler [1993, 1994] has expanded enormously on the meaning and interpretation of the correlation dimension and other fractal dimensions, but they still lack any direct interpretation in terms of the dynamics of the system. On the other hand, *Lyapunov exponents* are much more readily interpretable: the positivity of the largest Lyapunov exponent is often taken as the definition of chaos and it provides a direct measure of the sensitivity of the system to perturbation of initial conditions. However, Lyapunov exponents have traditionally been regarded as much harder to estimate. Early attempts such as those of Wolf et. al [1985] and Eckmann et al. [1986] lacked any clear measure of statistical performance. The approach developed by McCaffrey et al. [1992] and Nychka et al. [1992] was much more statistical, being based on nonparametric regression to estimate the dynamic map followed by direct calculations on the estimated map. Nevertheless it is still hard, in their approach, to obtain analytic results for such quantities as the asymptotic bias and variance of estimators. On the other hand, *local Lyapunov exponents* – Lyapunov exponents over a finite time horizon – can be studied much more explicitly. Wolff [1992] presented one approach to local Lyapunov exponents (henceforth LLEs) in the one-dimensional case.

The main aim of the present paper is to develop the theory of LLEs, in the general multidimensional case, in conjunction with recent results on function estimation by local polynomial regression. This method of nonparametric regression has been developed a great deal in the past few years, with explicit asymptotic results being available for the bias and variance of estimators thus allowing theoretical treatment of such questions as optimal kernel bandwidth, cf. Fan [1993], Ruppert and Wand [1994].

Lu [1996] extended the work of Ruppert and Wand [1994] to consider the estimation of derivatives of multivariate functions in the standard nonparametric regression setting, in which all the observation vectors are independent, and Lu [1994] extended those results to time series with correlated observations. Developing those results, Lu [1994] showed how to obtain approximations to the bias and variance of a local Lyapunov exponent estimator based on local linear or local quadratic regression. These results are reviewed in Sections 2–4 of the present paper. In Section 5 we describe an implementation of the methodology, and in Section 6 we give some numerical and real-data examples.

2 Local Lyapunov Exponents in Stochastic Systems

The general model we consider is of the form

$$x_{t+1} = m(x_t, x_{t-1}, ..., x_{t-p+1}) + \sigma\epsilon_{t+1}, \tag{2.1}$$

where $m : \mathbb{R}^p \to \mathbb{R}$ is some nonlinear function, $\sigma \geq 0$ and $\{\epsilon_t,\ -\infty < t < \infty\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1; moreover, ϵ_{t+1} is assumed independent of $\{x_s,\ s \leq t\}$. In the case $\sigma = 0$, (2.1) is often taken as a generic representation of a deterministic dynamical system in either discrete

or continuous time, as a consequence of Takens' embedding theorem. See, e.g., Section 3.2.6 of Isham [1993] for one of many discussions of this. However, very few real observed time series are expected to be deterministic dynamical systems, and adding the random noise term $\sigma\epsilon_{t+1}$ is the most obvious way of introducing a random component into the model. The determination of the *embedding dimension* p is one critical decision which has itself been the focus of detailed discussion in the literature, such as Cheng and Tong [1992].

The model (2.1) may also be written in *state space form*: define X_t to be the vector $(x_{t+p-1}, x_{t+p-2}, ..., x_t)^T$ and let $\Sigma = \sigma I_p$ (I_p is the $p \times p$ identity matrix), $E_t = (\epsilon_{t+p}, 0, ..., 0)$ and

$$M(X_t) = \left(m(x_{t+p-1}, x_{t+p-2}, ..., x_t), x_{t+p-1}, ..., x_{t+1}\right)^T, \tag{2.2}$$

then

$$X_{t+1} = M(X_t) + \Sigma E_{t+1}. \tag{2.3}$$

We are interested in the question of sensitivity to initial conditions. Consider two realisations $\{X_t,\ t \geq 0\}$ and $\{X'_t,\ t \geq 0\}$ of (2.3), starting from neighboring initial values X_0, X'_0 but subject to the same noise sequence $\{E_t\}$. After ℓ steps we will have, by a first-order Taylor expansion together with the chain rule,

$$\begin{aligned} X_\ell - X'_\ell &\approx \nabla M(X_{\ell-1})\nabla M(X_{\ell-2})...\nabla M(X_0)(X_0 - X'_0) \\ &= J^{(\ell)} \cdot (X_0 - X'_0) \quad \text{say} \end{aligned} \tag{2.4}$$

where $\nabla M(X)$ is the Jacobian matrix of the map M evaluated at X, and $J^{(\ell)}$ is the product of ℓ such matrices. Under very general conditions, the matrix

$$\Lambda = \lim_{\ell \to \infty} \left[\{J^{(\ell)}\}^T J^{(\ell)}\right]^{1/(2\ell)} \tag{2.5}$$

exists; moreover, the logarithms of the eigenvalues of Λ, ordered as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, are invariants of the system (not dependent on X_0 or the specific sequence of $\{E_t\}$). These are the Lyapunov exponents of the system. In particular, the largest Lyapunov exponent λ_1 is a direct measure of the exponential rate at which the trajectories $\{X_t\}$, $\{X'_t\}$ separate, and the property $\lambda_1 > 0$ is often taken as the definition of a chaotic system. The review papers of Isham [1993] and Jensen [1993] may be consulted for more details about the existence and properties of Lyapunov exponents.

This definition of the Lyapunov exponents of a stochastic system is the simplest one operationally, but it seems artificial because it introduces the rather unrealistic idea that two trajectories starting from different starting positions may nevertheless be subject to the same random shocks. In particular, this criticism is made by Tong [1995], pp. 405-6. However, a different viewpoint, also given by Tong, leads to practically the same thing. If we define $F_\ell(x) = \mathrm{E}\{X_\ell \mid X_0 = x\}$, the conditional mean of the ℓ-step prediction from a starting point x, then we may consider the effect of a small (vector) perturbation δ in x. We have

$$F_\ell(x + \delta) - F_\ell(x) = \dot{F}_\ell(x)\delta + o(||\delta||),$$

where

$$\dot{F}_\ell(x) = \mathrm{E}\left\{\prod_{j=1}^{\ell} \nabla M(X_{j-1}) \mid X_0 = x\right\}$$

(Tong [1995], p. 408). Thus the matrix $\dot{F}_\ell$ given by Tong is the same as our $J^{(\ell)}$ but with an additional "averaging over trajectories". The contrast between the two definitions is discussed by Yao and Tong [1994], who define "Lyapunov-like indices" which are not exactly the same as Lyapunov exponents but are nevertheless very similar. Our own point of view is that there are by now a number of ways of defining sensitivity to initial conditions, and ultimately one should consider all of them, but the Lyapunov exponents defined by (2.5) are a natural starting point for discussion about how one should estimate these quantities.

Nevertheless, one clear disadvantage of this definition is that, from the point of view of practical evaluation, (2.5) is difficult to handle because there is no good way to decide how large ℓ must be taken to guarantee a good approximation to Λ. In the definition of local Lyapunov exponents, we do not take the limit and work directly with

$$\Lambda^{(\ell)} = \left[\{J^{(\ell)}\}^T J^{(\ell)}\right]^{1/(2\ell)} \tag{2.6}$$

for finite ℓ; if we then denote the logarithms of the eigenvalues of $\Lambda^{(\ell)}$ as

$$\lambda_1^{(\ell)} \geq \lambda_2^{(\ell)} \geq \ldots \geq \lambda_p^{(\ell)}, \tag{2.7}$$

these are called the local Lyapunov exponents of order ℓ. In contrast to the global Lyapunov exponents, they do depend on X_0 and indeed the entire trajectory $\{X_0, X_1, \ldots, X_{\ell-1}\}$. However for some purposes this can be a positive feature, for example in helping to determine to what extent the short-term predictability of the system depends on the current state.

A second positive feature of local as opposed to global Lyapunov exponents is that there is a clear-cut procedure for estimating them from sample data. Indeed, if we can obtain an estimate of ∇M, then we can carry through the steps (2.4), (2.6) and (2.7) directly and so obtain estimates of the LLEs for any finite order ℓ, for any given starting point on the trajectory. The properties of the LLE estimators are then governed by those of the estimator $\widehat{\nabla M}$.

2.1 Application to the Lorenz Attractor

Figure 1 shows some plots generated from the Lorenz [1963] system. This is defined by the differential equations

$$\begin{aligned}
\frac{dx(t)}{dt} &= -\sigma\{x(t) - y(t)\}, \\
\frac{dy(t)}{dt} &= -x(t)y(t) + rx(t) - y(t), \\
\frac{dz(t)}{dt} &= x(t)y(t) - bz(t),
\end{aligned}$$

with $\sigma = 10$, $b = \frac{8}{3}$, $r = 28$. A trajectory was calculated using a one-step forward Euler method with step size 0.001. The resulting system was recorded at time intervals of 0.05, and we consider LLEs corresponding to a time step of 0.1, so that $\ell = 2$ for this example. Figure 1(a) shows the time series of LLE spectra. The means of the three LLEs are 3.98, −2.11 and −15.58, which are a long way from the global Lyapunov exponents of 0.90, 0 and −14.57. There is also substantial variability in the plot, illustrating the sensitivity of the LLE to the starting position.

Figure 1(b–d) show the distribution of the largest LLE in the projected phase space (x, y), (x, z) and (y, z). We use this to classify the phase space into three

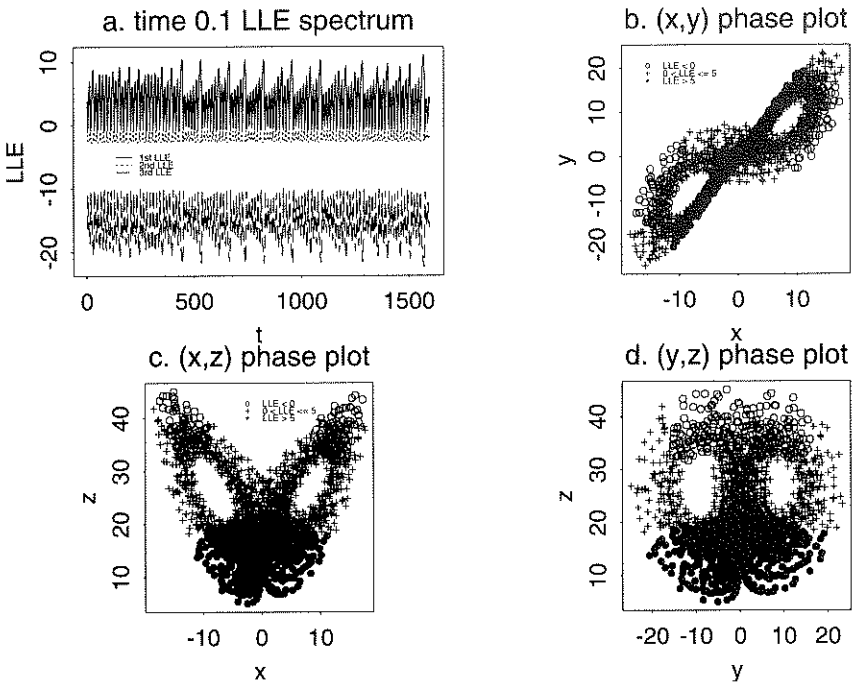

Figure 1 Local Lyapunov exponents for the Lorenz system. Plot (a) shows the two-step LLEs, $\lambda_1^{(2)}$ (top), $\lambda_2^{(2)}$ (middle) and $\lambda_3^{(2)}$ (bottom) each plotted against time t (horizontal axis). The remaining plots show three different projections of the attractor with regions of small $\lambda_1^{(2)}$ (circles), medium $\lambda_1^{(2)}$ (crosses) and high $\lambda_1^{(2)}$ (black dots) marked.

areas which are "predictable" ($\lambda_1^{(2)} < 0$, denoted by circles), "fairly unpredictable" ($0 < \lambda_1^{(2)} \leq 5$, denoted by crosses) and "very unpredictable" ($\lambda_1^{(2)} > 5$, denoted by black dots). The three regions are well separated in phase space. The predictability increases in the z direction as z increases. The most unpredictable region is in the region of low z and in the interchange of the two wings (most noticeable in the (x, y) plot). The practical implication of this result is that, for short-term forecasting, one may do better in certain regions than others. This example is, of course, deterministic, but essentially the same interpretation is made for stochastic systems.

3 Estimation by Local Polynomial Regression

Nychka et al. [1992] review a number of strategies for estimating m, concentrating on thin-plate splines and neural nets. Here we shall work with the alternative method of local polynomial regression which has been popularized in particular by the work of Fan [1993]. We shall concentrate on the case of local quadratic regression: to estimate $m(X)$ and $\nabla m(X)$ at a location $X \in \mathbb{R}^p$, based on data $X_0, ..., X_n$,

choose a scalar a, a p-vector b and a $p \times p$ symmetric matrix L to minimize

$$\sum_{t=0}^{n-1} \left\{ x_{t+1} - a - b^T(X_t - X) - (X_t - X)^T L (X_t - X) \right\}^2 K\left(\frac{X_t - X}{h}\right), \qquad (3.1)$$

where K is a p-dimensional kernel function and h is a scalar *bandwidth parameter*. We then take a as our point estimate of $m(X)$ and b as our estimate of the gradient $\nabla m(X)$. If we restrict $L = 0$ then (3.1) becomes local linear regression, and of course we may include higher-order powers of $X_t - X$ to get local polynomial regression of any order. However, following results in Fan [1993], Ruppert and Wand [1994] and Lu [1996], it appears to be especially appropriate to use local *linear* regression when estimating the function itself, and local *quadratic* regression for estimating first-order derivatives. Since the Lyapunov exponents depend on accurate estimation of the first-order derivatives, we concentrate on local quadratic regression as our estimation technique.

The solution of (3.1) may be expressed in matrix form as follows: let $q = (p+1)(p+2)/2$, $\mathbf{Y} = (x_1, ..., x_n)^T$, W =diag$\{K((X_0 - X)/h), ..., K((X_{n-1} - X)/h)$ and $\mathbf{X}$ the $T \times q$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & (X_0 - X)^T & \text{vech}^T\{(X_0 - X)(X_0 - X)^T\} \\ \vdots & \vdots & \vdots \\ 1 & (X_{n-1} - X)^T & \text{vech}^T\{(X_{n-1} - X)(X_{n-1} - X)^T\} \end{pmatrix}$$

where vech^T denotes the row vector consisting of the columns on and below the diagonal of a symmetric matrix. Let $\beta = (a, b^T, \text{vech}^T L)^T$. Then the problem becomes choosing β to minimize $(\mathbf{Y} - \mathbf{X}\beta)^T W (\mathbf{Y} - \mathbf{X}\beta)$ for which it is well known that the least-squares solution is

$$\widehat{\beta} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W \mathbf{Y}. \qquad (3.2)$$

provided $(\mathbf{X}^T W \mathbf{X})^{-1}$ exists. From (3.2), we estimate $\widehat{m}(X) = \widehat{a}$ and $\widehat{\nabla m}(X) = \widehat{b}$ for any specified value of X. As is usual in kernel density estimation, we fix K to be some spherically symmetric kernel and concentrate our attention on the "optimal" choice of bandwidth h. Previous results by Lu [1994, 1996] have shown how to obtain bias and variance approximations for the estimates of m, ∇m and ultimately the local Lyapunov exponents derived via (2.7), as functions of the bandwidth h. In theory at least, therefore, we may specify h in such a way as to minimize the asymptotic mean squared error. We shall next summarize these results.

4 Bias and Variance Approximations

We shall consider the estimation of the ℓ-step LLE starting from a fixed value of $X \in \mathbb{R}^p$, say $X = X_{(0)}$. We assume that the first-order derivatives of m at $X_{(0)}$ and also the next $\ell - 1$ data points, $X_{(1)}, ..., X_{(\ell-1)}$, are estimated by local quadratic regression as described in Section 3. We then form an estimated matrix $\widehat{J}^{(\ell)}$ as in (2.4), and estimated LLEs $\widehat{\lambda}_1^\ell \geq ... \geq \widehat{\lambda}_1^\ell$ from (2.6) and (2.7), in each case substituting the estimates $\widehat{\nabla M}$ for ∇M. The final estimates are sensitive to the bandwidth h used in defining $\widehat{\nabla M}$. To gain some indication of suitable h values to use, we proceed by calculating approximations to the bias and variance of these estimators.

The development of bias and variance approximations proceeds in three steps. First, we consider the estimation of ∇M or equivalently ∇m (recall (2.2)). Second, we develop such results for the matrix $J^{(\ell)}$ given by (2.4). These results will in turn be applied to the estimation of the LLEs $\lambda_1^{(\ell)}, ..., \lambda_p^{(\ell)}$. We shall only outline the results here, concentrating on those that are directly relevant to our end objectives, and omitting all proofs. For full details the reader is referred to Lu [1994].

First, then, suppose $\nabla m(X)$ is estimated for some fixed $X \in \mathbb{R}^p$ by the procedure described in Section 3. The main assumptions we require are:

(i) m is four times continuously differentiable in a neighborhood of X,
(ii) the stationary density of X_t, denoted f, exists and is continuous, and $f(X) > 0$,
(iii) the $\{\epsilon_t\}$ are i.i.d. and $\mathrm{E}|\epsilon_t|^{2+\delta} < \infty$ for some $\delta > 0$,
(iv) the kernel K is spherically symmetric and $\int u_1^{12} K(u_1, ..., u_p) du_1 ... du_p < \infty$.

Define

$$\mu_k = \int u_1^k K(u_1, ..., u_p) du_1 ... du_p, \quad \nu_k = \int u_1^k K^2(u_1, ..., u_p) du_1 ... du_p$$

and let $b(X)$ denote the p-vector whose ith entry is

$$\mu_4 \frac{\partial^3 m}{\partial x_i^3} + 3\mu_2^2 \sum_{j \neq i} \frac{\partial^3 m}{\partial x_i \partial x_j^2}.$$

Also let A be the $p \times p$ matrix

$$A = \frac{\sigma^2 \nu_2}{\mu_2^2 f(X)} I_p.$$

Under these conditions, if $n \to \infty$, $h \to 0$ in such a way that $nh^{p+2} \to \infty$, we have that

$$n^{1/2} h^{1+p/2} \left\{ \widehat{\nabla m}(X) - \nabla m(X) - \frac{h^2}{6\mu_2} b(X) \right\} \xrightarrow{d} \mathcal{N}(0, A) \tag{4.1}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean μ and covariance matrix Σ. In terms of bias and variance considerations, (4.1) is saying that

$$\text{Bias of } \widehat{\nabla m}(X) \approx \frac{h^2}{6\mu_2} b(X),$$

$$\text{Variance of } \widehat{\nabla m}(X) \approx n^{-1} h^{-p-2} A.$$

Lu [1994] also shows that if the conditions are satisfied at k distinct points X, say $X_{(0)}, ..., X_{(k-1)}$, then the estimates $\widehat{\nabla m}(X_{(0)}), ..., \widehat{\nabla m}(X_{(k-1)})$ are asymptotically independent. In the following discussion, we are particularly interested in the case where $X_{(k)}$, $k = 0, 1, ..., \ell - 1$, form a time series, as outlined in the introduction to this section.

It immediately follows from (4.1) and the definition of $M(X)$ that we can write

$$n^{1/2} h^{1+p/2} \{ \widehat{\nabla M}(X_{(k)}) - \nabla M(X_{(k)}) - h^2 B_k \} \xrightarrow{d} W_k \tag{4.2}$$

where

$$B_k = \frac{1}{6\mu_2} \begin{pmatrix} b(X_{(k)})^T \\ \mathbf{0} \end{pmatrix}, \quad W_k = \frac{\sigma \sqrt{\nu_2}}{\mu_2 \sqrt{f(X_{(k)})}} \begin{pmatrix} Z_k^T \\ \mathbf{0} \end{pmatrix}.$$

Here Z_k is a p-vector of i.i.d. standard normal random variables and $\mathbf{0}$ is a $(p-1)\times p$ matrix of zeros. Moreover, $\{W_k\}$ corresponding to distinct fixed values of $\{X_{(k)}\}$ are independent.

Now we consider the second step, in which these results are extended to the estimation of $J^{(\ell)} = G_{\ell-1}...G_0$ say, where $G_k = \nabla M(X_{(k)})$.

Taking a first-order expansion using the chain rule, we have

$$\widehat{J}^{(\ell)} - J^{(\ell)} \approx \sum_{k=0}^{\ell-1} G_{\ell-1}...G_{k+1}(\widehat{G}_k - G_k)G_{k-1}...G_0.$$

Define

$$\begin{aligned} B^{(\ell)} &= \sum_{k=0}^{\ell-1} G_\ell...G_{k+1}B_kG_{k-1}...G_0, \\ W^{(\ell)} &= \sum_{k=0}^{\ell-1} G_\ell...G_{k+1}W_kG_{k-1}...G_0. \end{aligned} \tag{4.3}$$

Then it can be shown that

$$n^{1/2}h^{1+p/2}\{\widehat{J}^{(\ell)} - J^{(\ell)} - h^2B^{(\ell)} - O(h^4)\} \xrightarrow{d} W^{(\ell)}. \tag{4.4}$$

Noting that only the first rows of the matrices B_k and W_k are non-zero, the expressions for $B^{(\ell)}$ and $W^{(\ell)}$ in (4.3) can be simplified, as follows. Let $g_k^{(\ell)}$ denote the first column of the matrix $G_\ell...G_{k+1}$ and let $H_k = G_{k-1}...G_0$. Then

$$B^{(\ell)} = \frac{1}{6\mu_2}\sum_{k=0}^{\ell-1} g_k^{(\ell)}b(X_{(k)})^TH_k, \qquad W^{(\ell)} = \frac{\sigma\sqrt{\nu_2}}{\mu_2}\sum_{k=0}^{\ell-1}\frac{g_k^{(\ell)}Z_k^TH_k}{\sqrt{f(X_{(k)})}}. \tag{4.5}$$

The third step is to extend this to estimating the eigenvalues of $\{J^{(\ell)}\}^TJ^{(\ell)}$. In fact it is easier to work with the *singular values* $\delta_1^{(\ell)} \geq \delta_2^{(\ell)} \geq \delta_p^{(\ell)}$ of $J^{(\ell)}$; the singular values of $J^{(\ell)}$ are the square roots of the eigenvalues of $\{J^{(\ell)}\}^TJ^{(\ell)}$ so the LLEs will be given by

$$\lambda_j^{(\ell)} = \frac{1}{\ell}\log\delta_j^{(\ell)}, \quad j = 1,...,p.$$

Our estimation procedure will be to let $\{\widehat{\delta}_j^{(\ell)},\ j = 1,...,p\}$ denote the singular values of $\widehat{J}^{(\ell)}$ and then to define

$$\widehat{\lambda}_j^{(\ell)} = \frac{1}{\ell}\log\widehat{\delta}_j^{(\ell)}, \quad j = 1,...,p. \tag{4.6}$$

We now need to consider the asymptotic distribution of $\{\widehat{\delta}_j^{(\ell)},\ j = 1,...,p\}$. Lu [1994] gives the following result. Suppose X_n, B_n and W are random $p\times p$ matrices, A a fixed $p\times p$ matrix and c_n a scalar for each n. Suppose as $n\to\infty$, we have that $c_n\to\infty$, $B_n \xrightarrow{p} 0$ and $c_n(X_n - A - B_n) \xrightarrow{d} W$. For any matrix M let $\delta(M) = (\delta_1(M),...,\delta_p(M))^T$ denote the singular values of M arranged in decreasing order. Suppose the singular values of A are all distinct and positive. We can write $A = U\Delta V^T$ where U and V are orthogonal matrices and Δ is a diagonal matrix with diagonal entries $\delta_1(A),...,\delta_p(A)$. For any square matrix M let $\mathrm{dgv}(M)$ denote the column vector consisting of the diagonal entries of M. Then

$$c_n\left\{\delta(X_n) - \delta(A) - \mathrm{dgv}(U^TB_nV) - o_p(B_n)\right\} \xrightarrow{d} \mathrm{dgv}(U^TWV). \tag{4.7}$$

A recent paper by Eaton and Tyler [1994] has reviewed results of this nature including the case in which the singular values of A are not all distinct; this is however much harder to deal with, and we do not consider it here.

We can now see that, provided the singular values of $J^{(\ell)}$ are distinct, a combination of (4.4), (4.5) and (4.7), together with a simple delta function argument to handle the transformation (4.6), gives the result we are looking for. The final result is in the following:

Theorem 4.1 *Suppose $J^{(\ell)}$ has p distinct non-zero singular values and write its singular value decomposition in the form $J^{(\ell)} = U^{(\ell)}\Delta^{(\ell)}(V^{(\ell)})^T$ where $U_1^{(\ell)}, ..., U_p^{(\ell)}$ are the columns of $U^{(\ell)}$ and $V_1^{(\ell)}, ..., V_p^{(\ell)}$ are the columns of $V^{(\ell)}$. Then*

$$n^{1/2}h^{1+p/2}\left\{ \begin{pmatrix} \widehat{\lambda}_1^{(\ell)} \\ \vdots \\ \widehat{\lambda}_p^{(\ell)} \end{pmatrix} - \begin{pmatrix} \lambda_1^{(\ell)} \\ \vdots \\ \lambda_p^{(\ell)} \end{pmatrix} - \frac{h^2}{6\mu_2\ell} \begin{pmatrix} (\delta_1^{(\ell)})^{-1} \sum_{k=0}^{\ell-1} (U_1^{(\ell)})^T g_k^{(\ell)} b(X_{(k)})^T H_k V_1^{(\ell)} \\ \vdots \\ (\delta_p^{(\ell)})^{-1} \sum_{k=0}^{\ell-1} (U_p^{(\ell)})^T g_k^{(\ell)} b(X_{(k)})^T H_k V_p^{(\ell)} \end{pmatrix} - o(h^2) \right\}$$
$$\xrightarrow{d} \mathcal{N}(0, \Sigma^{(\ell)}). \tag{4.8}$$

Here the covariance matrix $\Sigma^{(\ell)}$ has entries $\sigma_{ij}^{(\ell)}$ where

$$\sigma_{ij}^{(\ell)} = \frac{\sigma^2\nu_2}{\ell^2\mu_2^2\delta_i^{(\ell)}\delta_j^{(\ell)}} \sum_{k=0}^{\ell-1} \frac{1}{f(X_{(k)})} \left\{ (U_i^{(\ell)})^T g_k^{(\ell)} \right\} \left\{ (U_j^{(\ell)})^T g_k^{(\ell)} \right\} (V_i^{(\ell)})^T H_k^T H_k V_j^{(\ell)}. \tag{4.9}$$

The practical interpretation of this result is that the bias of $\widehat{\lambda}_j^{(\ell)}$ is approximately

$$\frac{h^2}{6\mu_2\ell}(\delta_j^{(\ell)})^{-1} \sum_{k=0}^{\ell-1} (U_j^{(\ell)})^T g_k^{(\ell)} b(X_{(k)})^T H_k V_j^{(\ell)} \tag{4.10}$$

while its variance is approximately $n^{-1}h^{-2-p}\sigma_{jj}^{(\ell)}$.

This leads to an expression for the mean squared error (MSE=bias^2+variance) and we can choose h to minimize this, at least from a theoretical point of view, and we can also calculate the resulting MSE.

Lu [1994, 1996] also gave expressions for the bias and variance of $\widehat{\nabla M}$, and consequently the LLEs, in the case where estimation is via local linear regression instead of local quadratic regression. For estimation of M itself, rather than ∇M, we use local linear regression. In this case explicit expressions for the bias and variance of local linear regression have been given by Ruppert and Wand [1994].

5 Implementation

The kernel estimator was computed by solving the weighted least squares equation (3.2) using standard software for the solution of least squares problems. As part of the solution, this gave us the minimized weighted sum of squares

$$S = (\mathbf{Y} - \mathbf{X}\widehat{\beta})^T W (\mathbf{Y} - \mathbf{X}\widehat{\beta}). \tag{5.1}$$

For local linear regression, the method is the same, but the "vech" columns in the definition of $\mathbf{X}$ and the corresponding components of β are omitted. The kernel used here was $K(x) = C_p(1 - |x|^2)$ where $|\cdot|$ denotes L_2 norm ($|(x_1, ..., x_p)|^2 = \sum_1^p x_j^2$) and C_p is chosen so that the kernel integrates to 1 (this criterion leads to $C_p = \pi^{-\frac{p}{2}}\Gamma(2 + \frac{p}{2})$).

The bias term, which depends on higher-order derivatives of m, is not directly computable except as a theoretical calculation in cases where m is known (as in the cosine map example, Section 6 below). One possible strategy is to extend the method of local polynomial regression to estimate higher-order derivatives and to use them to derive sample-based estimators of the bias. This is known to be somewhat problematic but we shall use this method in two examples in Section 6.

The variance term, however, is more directly computable from (4.9). The main difficulty with this formula is the need for a separate estimate of f, the density of data points in the assumed stationary distribution. It is possible to avoid estimating this by an alternative strategy, which we now outline.

Consider $\widehat{\beta}$ defined by (3.2). The data vector $\mathbf{Y}$ has covariance matrix $\sigma^2 I$, so the covariance matrix of $\widehat{\beta}$ is

$$\sigma^2(\mathbf{X}^T W\ \mathbf{X})^{-1}\ \mathbf{X}^T W^2\ \mathbf{X}(\mathbf{X}^T W\ \mathbf{X})^{-1}. \tag{5.2}$$

We need to estimate σ^2. However after defining S by (5.1), we can easily see that $\mathrm{E}\{S\} = K\sigma^2$, where

$$K = \mathrm{tr}\left\{W - W\ \mathbf{X}(\mathbf{X}^T W\ \mathbf{X})^{-1}\ \mathbf{X}^T W\right\}. \tag{5.3}$$

Consequently S/K is an unbiased estimator of σ^2, and by combining this with (5.2) we obtain an unbiased sample estimate of the covariance matrix of $\widehat{\beta}$, for any particular evaluation of (3.2).

Finally, by extracting the components of β which correspond to elements of ∇M, and passing these through the same operations as led to (4.9), we obtain sample-based estimators of the variances of the estimated LLEs, which avoids the direct evaluation of f.

6 Numerical Examples

Our first example is totally artificial, but is intended to illustrate one method for selecting the bandwidth h in a situation where we can calculate the "correct" answer. Then we briefly discuss two well-known data sets.

The artificial example is the cosine map, which is closely related to the better-known Hénon map. In its deterministic form, it consists of iterating the second-order nonlinear difference equation $x_{t+1} = \cos(2.8x_t) + 0.3x_{t-1}$ (Nychka et al. [1992]), and discarding initial transients. Figure 2(a) shows a plot of x_{t+1} against x_t, while Figure 2(b) shows the same plot with additive noise (as in (2.1)) with $\sigma = 0.2$. Our analysis is based on the data in Figure 2(b).

We first consider estimation of the function $m(\cdot)$, and then go on to consider the LLEs. Moreover, we do separate analyses of the full data set which consists of 10,000 data points, and a much shorter data set consisting only of the first 400 data points. This is intended to bring out the difference between large and relatively small sample sizes. Thus there are four analyses in all, and one of the features which emerges is the need for quite different bandwidths in the different cases.

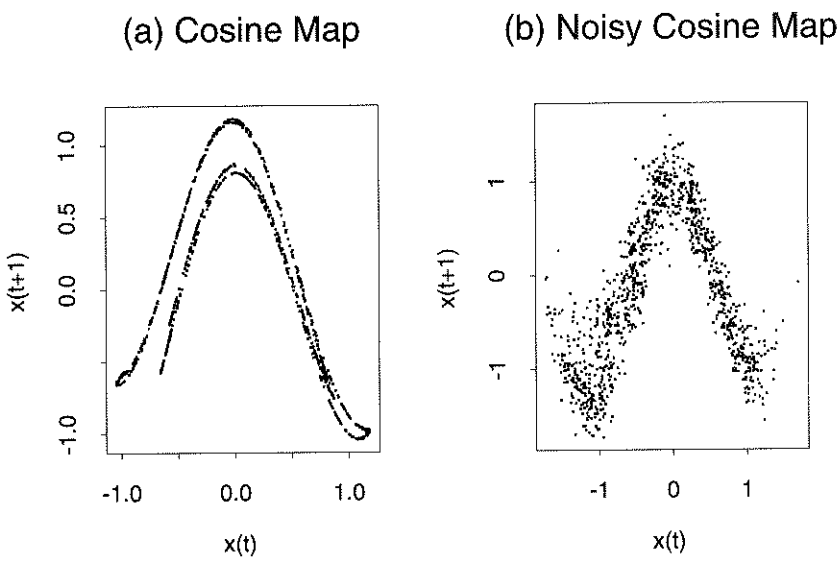

Figure 2 Data plots with and without noise

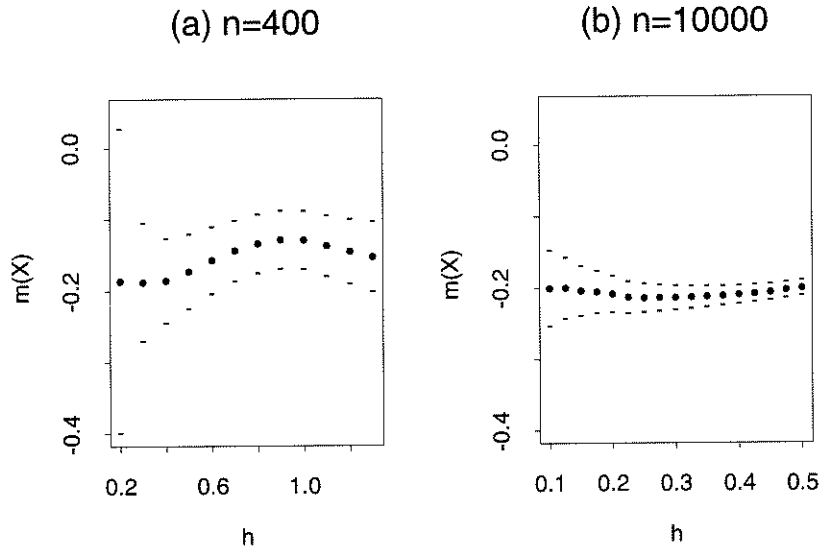

Figure 3 Estimated function values and 95% CIs

The methods are "local" in the sense that they are designed to fit the function and its LLEs in the neighborhood of a specific point, and for the present analyses we work from a single starting point, arbitrarily selected to be $x_0 = 0.74327$, $x_1 = -0.57342$. Figure 3(a) shows a series of estimates of $m(x_1, x_0)$, computed by local linear regression, for several different values of the bandwidth h, for the data set with sample size $n = 400$. Also shown are the upper and lower boundaries of the approximately 95% confidence intervals, using standard errors computed by the

method described in Section 5. Figure 3(b) shows the same thing, for a different set of values of h, computed for $n = 10,000$. The larger sample size in this case allows us to use smaller values of h while still reducing the standard errors.

In each case, the standard errors increase sharply at the left hand end of the displayed range of h values, suggesting that variance is dominating bias. For h too large, however, bias dominates variance. Although it is a highly subjective judgement to pick out an "optimal" value of h from such plots, they can nevertheless be helpful in indicating a suitable range of values and the sensitivity of the estimate to the exact choice of h.

For the present example, being based on a known m, we can compute the bias term exactly, using formulae in Ruppert and Wand [1994], and so work out the exact h to minimize the MSE. For $n = 250$, the best h is 0.364, for which the root mean squared error (RMSE) is $0.038 = 0.28 \times n^{-\frac{1}{3}}$. For $n = 10,000$ the best h is 0.221, and RMSE$= 0.014 = 0.30 \times n^{-\frac{1}{3}}$. Note that although we use an analytic formula for the bias here, the corresponding formula for the variance still depends on the unknown stationary density f, and since this is not computable analytically, we continue to use the method described in Section 5 for the variances. Nevertheless the results between the two sample sizes are very consistent as reflected by the estimated constant of proportionality in the theoretical $n^{-\frac{1}{3}}$ relationship for the RMSE.

Now let us consider estimation of LLEs, for which we take $\ell = 5$. We now use local quadratic regression, and plot both the LLEs and their confidence intervals for a range of h, in Figure 4 for $n = 400$ and Figure 5 for $n = 10,000$. In Figure 4 it is rather hard to pick out an optimal value of h, since all the confidence intervals are rather wide relative to the range of estimates of the LLEs, but we might guess that the LLEs are about 0.2 and −1.3 with standard errors of about 0.08 and 0.25 respectively. In fact, the true values for $\lambda_1^{(\ell)}$ and $\lambda_2^{(\ell)}$ are 0.356 and −1.56, and the best bandwidth for $\lambda_1^{(\ell)}$ is 0.53 with RMSE .105$= 0.47 \times n^{-\frac{1}{4}}$.

Figure 5 shows a very consistent sequence of estimates of the LLEs with sample size $n = 10,000$, though even here it is hard to pick out an optimal value of h. Arbitrarily we select $h = 0.4$ for which the estimate of $\lambda_1^{(\ell)}$ is 0.31 and standard error 0.025. The best h in this case is 0.35 with RMSE$=0.046= 0.46 \times n^{-\frac{1}{4}}$. As might be expected, quoting the standard error without taking bias into account does somewhat underestimate the true RMSE.

We do not consider estimation of the second LLE in the same detail since this is less important for any "chaos" interpretation and in any case the errors of estimation are much larger.

The qualitative messages from this example are, first, that the optimal bandwidth depends on the problem being considered, being larger for smaller sample sizes, and larger for LLE estimation than for estimation of the m function itself. Second, visual inspection of the plots gives reasonable guidance to the choice of bandwidth and the error of the resulting estimates, though the error estimate is always something of an underestimate because it is based only on variance and not at all on bias.

Now we turn to two well-known real data examples, both given by Tong [1990]. Figure 6 is based on $\log_{10}$ counts of Canadian lynxes, 1821–1934; previous analyses are due to Tong [1990], Cheng and Tong [1992], amongst others. The figure shows a time series plot, phase plot, two-year LLE spectrum plot and associated confidence

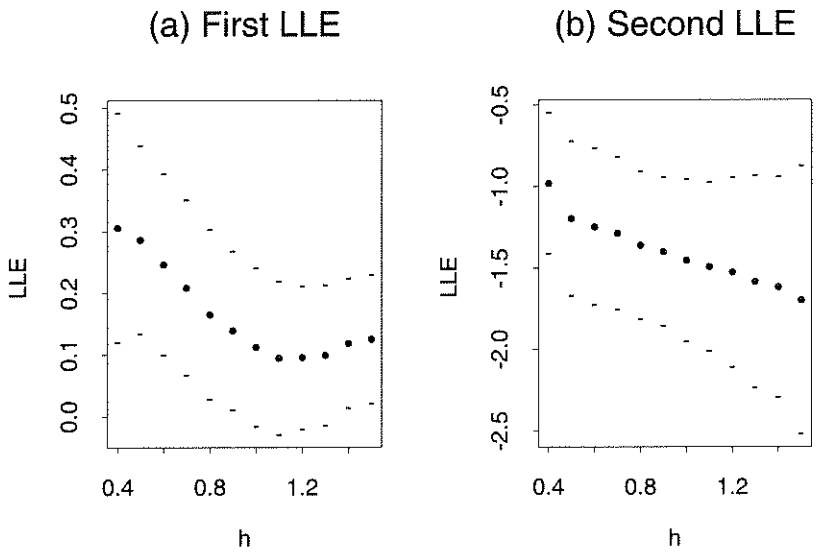

Figure 4 LLE calculations for n=400

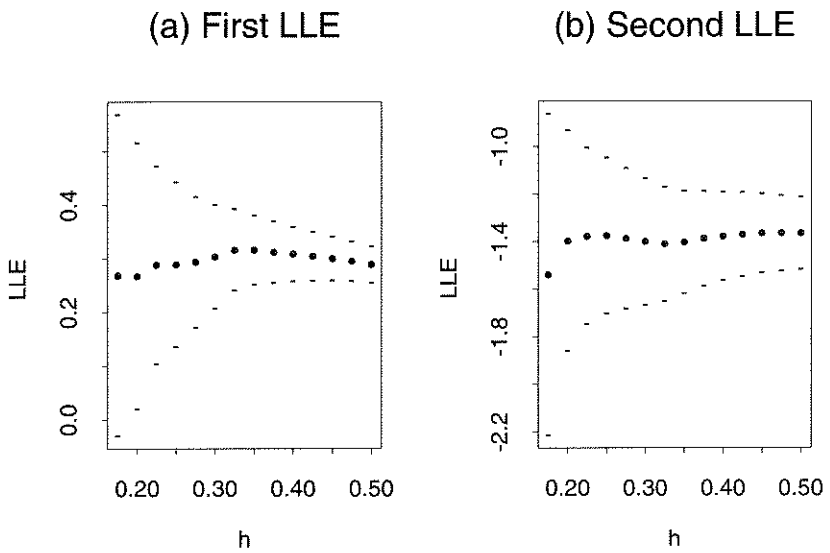

Figure 5 LLE calculations for n=10000

intervals. The LLE computations were based on embedding dimension $p = 2$. Also superimposed on the time series and phase plots are an indicator of the value of the largest LLE (large values denoted by a $+$, small values by o). This shows a clear subdivision of the phase space into less predictable and more predictable regions, much as we saw earlier for the theoretical Lorenz attractor. Figure 6(c) gives an indication of how both LLEs vary along the length of the series. Here the "bias-corrected point estimate" is computed by using a local cubic regression to estimate

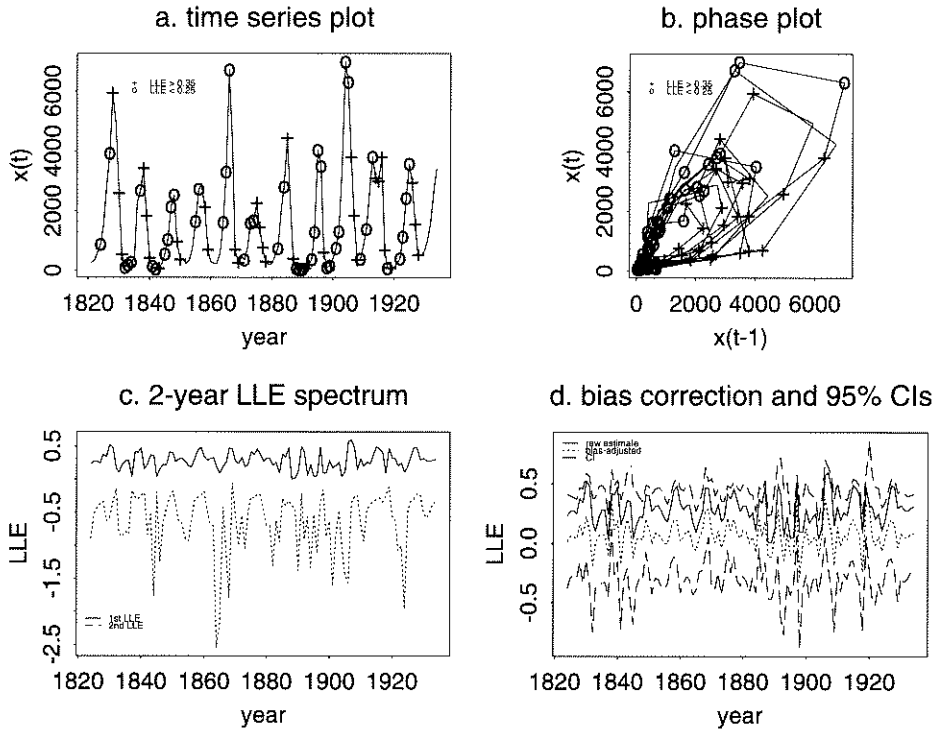

Figure 6 Annual lynx numbers. o denotes a point for which the 2-year largest LLE for $\log_{10}$(lynx) is < 0.25; + denotes a point for which it is > 0.35. The estimation parameters are $p = 2$, $h = 1.4$, $h_1 = 2$.

third-order derivatives of m, and plugging into the theoretical bias formula, which is then subtracted from the prediction interval. The upper and lower boundaries of the confidence intervals are then two standard errors either side of this. A bandwidth $h = 1.4$ was used for the LLE estimation and $h_1 = 2$ as an input to local cubic regression in computing the bias correction. The large bias and standard error are not surprising considering the short length of the series. The bias-adjusted LLE estimates are generally close to 0 which suggests that there is some predictability in the series. It appears that the rising part of each cycle is more predictable than the falling part. The asymmetry of the cycles is evidence of irreversibility, itself a well-known feature of nonlinear time series.

Figure 7 is based on similar calculations for the annual sunspot numbers, 1700–1990. In this case the bias correction has again been computed, but is much less important. The estimates of $\lambda_1^{(\ell)}$ are mostly positive, suggesting unpredictability, with greater level of unpredictability on the rising part than on the falling part of each roughly 11-year cycle. Note that this is the opposite pattern from the lynx example. The period around 1955–1961 is unusual and this has been partly compensated for by using a larger bandwidth ($h = 110$) for this portion than for the rest of the series ($h = 60$). The bandwidth for the bias correction estimate was $h_1 = 120$. The period around 1957 is much more unstable than the rest of the series and is omitted from plots c and d in view of the very large standard errors.

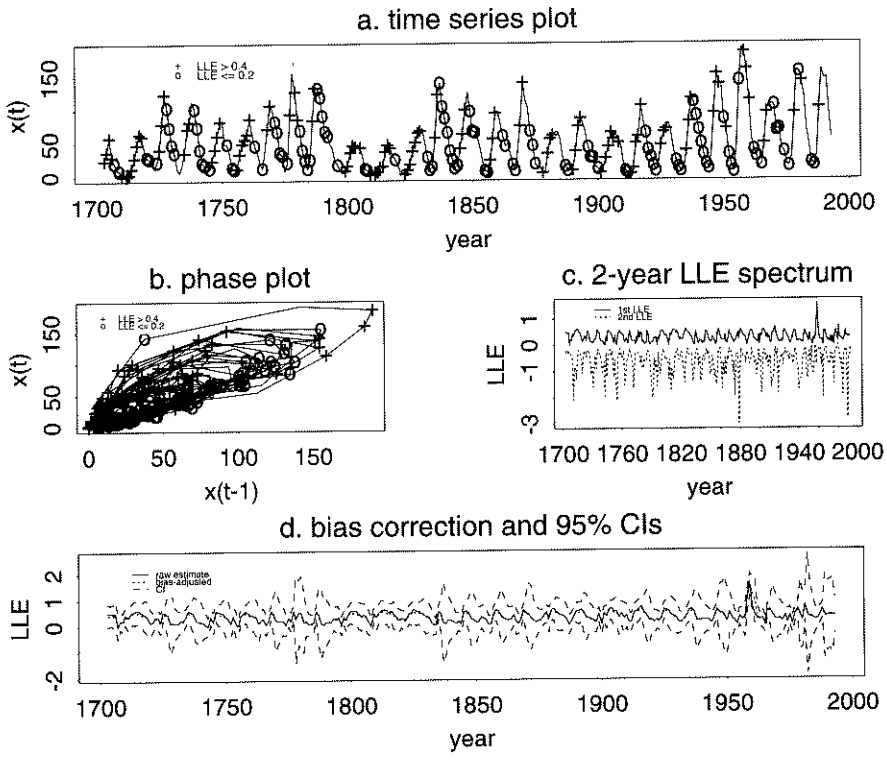

Figure 7 Annual sunspot numbers. o denotes a point for which the 2-year largest LLE is ≤ 0.2; + denotes a point for which it is > 0.4. The estimation parameters are $p = 2$, $h = 80$, $h_1 = 120$.

Once again there is strong evidence of irreversibility of the time series and this is reinforced by the LLE estimates.

7 Summary and Conclusions

Local Lyapunov exponents are an important indicator of the dynamical behavior of a nonlinear system. For large ℓ they approximate the global Lyapunov exponents, and in particular the sign of the largest LLE is an indicator of chaos. For small ℓ, however, they can also be valuable, indicating more or less predictable regions of the phase space.

In this paper, we have presented one method for their estimation, via local polynomial regression, and given theoretical formulae for the bias and variance of the resulting estimators. These formulae help to determine theoretically optimal values for the bandwidth parameter h, and also the overall accuracy of the estimation procedure.

Practical experience with these methods is still very limited, but we have discussed one approach towards bandwidth selection. Of course there are a number of others which could be investigated. Preliminary analyses of two well-known time series, the lynx and sunspots data, have shown the ability of the LLE spectrum to distinguish different portions of the phase space which appear to correspond to different levels of predictability.

However there remain many issues where further work is needed. Two which may be of particular importance are the selection of the order of the model, p, and the behavior of the estimation procedure in higher dimensions, say for $p \geq 5$.

References

Cheng, B. and Tong, H. [1992] *Consistent nonparametric order determination and chaos*, J. R. Statist. Soc. B **54**, 427–449.

Cutler, C. D. [1993] *A review of the theory and estimation of fractal dimension*, Dimension Estimation and Models, edited by H. Tong, World Scientific, Singapore, pp. 1–107.

Cutler, C. D. [1994] *A theory of correlation dimension for stationary time series*, Phil. Trans. R. Soc. (Lond.) **A348**, 343–355.

Drazin, P. G. and King, G. P. [1992] *Interpretation of Time Series from Nonlinear Systems*, North-Holland, Amsterdam.

Eaton, M. L. and Tyler, D. E. [1994] *The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis*, J. Multivariate Analysis **50**, 238–264.

Eckmann, J.-P., Oliffson Kamphorst, S., Ruelle, D. and Ciliberto, S. [1986] *Lyapunov exponents from time series*, Phys. Rev. A **34**, 4971–4979.

Fan, J. [1993] *Local linear regression smoothers and their minimax efficiency*, Ann. Statist. **21**, 196–216.

Grassberger, P. and Procaccia, I. [1983] *Measuring the strangeness of strange attractors*, Physica D **9**, 189–208.

Grenfell, B. T., May, R. M. and Tong, H. (eds.) [1994] *Discussion meeting on chaos and forecasting*, Phil. Trans. R. Soc. (Lond.) **A348**, 325–530.

Isham, V. [1993] *Statistical aspects of chaos: a review*, Networks and Chaos — Statistical and Probabilistic Aspects (ed. O. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall), Chapman and Hall, London, pp. 124–200.

Jensen, J. L. [1993] *Chaotic dynamical systems with a view towards statistics: a review*, Networks and Chaos — Statistical and Probabilistic Aspects (ed. O. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall), Chapman and Hall, London, pp. 201–250.

Lorenz, E. N. [1963] *Deterministic nonperiodic flow*, J. Atmos. Sci. **20**, 130–141.

Lu, Z.-Q. [1994] *Statistical estimation of local Lyapunov exponents: Towards characterizing predictability in nonlinear systems*, Preprint, University of North Carolina at Chapel Hill.

Lu, Z.-Q. [1996] *Multivariate weighted polynomial fit and partial derivative estimation*, Journal of Multivariate Analysis (to appear).

McCaffrey, D., Ellner, S., Gallant, A. R. and Nychka, D. [1992] *Estimating the Lyapunov exponent of a chaotic system with nonparametric regression*, J. Amer. Statist. Assoc. **87**, 682–695.

Nychka, D., Ellner, S., Gallant, A. R. and McCaffrey, D. [1992] *Finding chaos in noisy systems*, J. R. Statist. Soc. B **54**, 399-426.

Ruppert, D. and Wand, M. [1994] *Multivariate locally weighted least squares regression*, Ann. Statist. **22**, 1346–1370.

Smith, L. A. [1988] *Intrinsic limits on dimension calculations*, Physics Letters A **133**, 283–288.

Smith, R. L. [1992] *Estimating dimension in noisy chaotic time series*, J. R. Statist. Soc. B **54**, 329–352.

Tong, H. [1990] *Non-linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.

Tong, H. [1995] *A personal overview of non-linear time series analysis from a chaos perspective*, Scand. J. Statist. **22**, 399–421.

Tong, H. and Smith, R. L. (eds.) [1992] *Royal Statistical Society meeting on chaos*, J.R. Statist. Soc. B **54**, 301–474.

Wolf, A., Swift, J. B., Swinney, H. L. and Vastano, J. A. [1985] *Determining Lyapunov exponents from a time series*, Physica D **16**, 285–317.

Wolff, R. C. L. [1992] *Local Lyapunov exponents: Looking closely at chaos*, J. R. Statist. Soc. B **54**, 353–371.

Wolff, R. C. L. [1994] *Independence in time series: another look at the BDS test*, Phil. Trans. R. Soc. (Lond.) **A348**, 383–395.

Yao, Q. and Tong, H. [1994] *Quantifying the influence of initial values on nonlinear prediction*, J.R. Statist. Soc. B **56**, 701–725.