Quantifying the influence of initial values on nonlinear prediction

QUANTIFYING THE INFLUENCE OF INITIAL VALUES ON NONLINEAR PREDICTION*

Qiwei Yao
University of Kent, UK
and
Southeast University, China

Howell Tong†
University of Kent, UK

Abstract

Motivated by the m-step-ahead prediction problem in nonlinear time series, a brief sketch of stochastic chaotic systems is provided. The accuracy of the prediction depends on the initial value, which is a typical feature of nonlinear but not necessarily chaotic models. However, if the model is chaotic, a small noise can be amplified very quickly through the time evolution at some initial values, thereby decreasing the reliability of the prediction dramatically. Further, if the model is chaotic, small shifts in some initial values can lead to considerable errors in prediction, which can be monitored by the newly defined Lyapunov-like indices. For the nonparametric predictor constructed by the locally linear regression method, the mean squared error may be decomposed into two parts: the conditional variance and the divergence resulted from a small shift in initial values. In fact, the decomposition also holds for more general predictors. A consistent estimator of the Lyapunov-like index is also constructed by the locally linear regression method. Both simulated and real data have been used as illustrations.

Keywords: absolutely regular; chaos; locally linear regression; Lyapunov exponent; Lyapunov-like index; noise amplification; nonlinear prediction; nonlinear time series; nonparametric regression; stochastic dynamical system.

---

*Research partially supported by the Science and Engineering Research Council (U.K.).

†Address for correspondence: Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF,UK.

1 Introduction

The predominance of the assumption of linearity in time series analysis until quite recently has perpetuated the misconception that the reliability of the prediction is independent of the state. Indeed many standard text books in time series analysis have given "error bounds" for the point forecasts which are *uniform* over the state space. Whilst uniformity may well be true for the case of linear least squares prediction, it is certainly untrue for the case of nonlinear prediction. The non-uniformity is not surprising to anybody who has had a first hand experience with the stock market! Although some recent authors (e.g. Tong, 1990 and the references therein) have recognized this fact for particular cases, the literature lacks a precise quantification for the general situation. Moreover, it is of practical significance to analyse the effect of a perturbation of the initial values on the prediction. As long ago as 1956, Wiener warned of the danger of ignoring "*the very real possibility of the self-amplification of small details in the weather map*" (Wiener, 1956 p.247).

In the chaos literature, the concept of a *Lyapunov exponent* has been developed to characterize the sensitive dependence on the initial value of a *deterministic* system, which may take, for example, the form of a nonlinear autoregressive model without noise. Naturally, the concept of the sensitivity to initial conditions as well as the Lyapunov exponent has to be substantially extended in order to cope with a stochastic model. This paper will not present a definitive description of stochastic chaotic systems. However, a heuristic exploration of chaos in stochastic systems will help us to understand the influence of initial values on the prediction of nonlinear time series. It is intuitively clear that the stochastic dynamic noise will, by permeating through the system dynamics, interact with any perturbation of the initial value. First, a small shift in the initial value can lead to considerable change in the distribution of the state variable conditionally on the fixed initial value in the short or medium term, a phenomenon which we call the sensitive dependence on initial values. Next, the stochastic noise in the system can be amplified very rapidly through the time evolution, a feature which is wholly stochastic in that it disappears for a purely deterministic system. In fact, the noise amplification in a nonlinear (not necessarily chaotic) system is not always monotonic (cf. §6.2.2 of Tong 1990). In the context of nonlinear prediction, the influence of initial values may be decomposed into two parts. First, the (theoretical) accuracy of the prediction, which may be represented by the conditional variance, varies with the different initial values. This is a typical feature of nonlinear but *not* necessarily chaotic models. However, if the model is chaotic, the conditional variance can increase very quickly, but not always monotonically, at some initial values due to the noise amplification

through the time evolution. Secondly, the errors in the prediction due to the perturbation in the initial values can be considerable if the model is chaotic, which will be closely monitored by the Lyapunov exponent type quantity, or more specifically the *Lyapunov-like index* defined in Section 2. The above decomposition will be quantified in Theorem 1 in Section 3. As far as we are aware, the decomposition is new. The non-monotonicity of conditional variance implies the possibility that the error of a $(m+1)$-step prediction can be smaller than that of the m-step prediction from the same initial value. This is another interesting feature of nonlinear (but not necessarily chaotic) models.

The plan of the rest of the paper is as follows. Section 2 provides a brief sketch of chaos in a stochastic system with the emphasis on its application in nonlinear prediction. We introduce the Lyapunov-like index, which characterizes the divergence of conditional expectations of the first components of two trajectories with nearby initial values in the short and medium terms. We also present a quantitative description of how a small noise can be amplified rapidly in a chaotic system, and when the non-monotonicity of conditional variance can happen in the case of small noise. In Section 3, the m-step predictors for nonlinear time series, as well as the estimators of the Lyapunov-like indices, are constructed by using the locally linear regression method (cf. Fan 1992). The asymptotic decomposition of the mean squared error of the predictor is developed for the strictly stationary and absolutely regular time series, which shows that the reliability of the prediction does depend on the initial values. The consistency of the estimator of the Lyapunov-like index is proved. In Section 4, the method is illustrated with two sets of simulated data and also the Canadian lynx data and the Wolf's sunspot numbers. All mathematical proofs are relegated to Section 5.

Finally in this section, we introduce some convention on the notation. We always use vectors in the column form. A^τ denotes the transpose of matrix (or vector) A. $A = \circ(1)$ means that all components of A are $\circ(1)$, and other 'order' notations are similarly interpreted. $||\cdot||$ denotes the Euclidean norm. C denotes a generic constant which may be different at different places.

2 Stochastic chaotic systems

2.1 Chaos in dynamical systems

It is almost impossible to give a precise mathematical definition of deterministic chaos which at the same time encapsulates all that the term implies in the diverse literature. However, we are all agreed that the sensitive dependence on initial conditions is a typical feature of a chaotic system,

and which is characterized by the well-known Lyapunov exponents (cf. Eckmann and Ruelle 1985, Chatterjee and Yilmaz 1992, Berliner 1992, and the references therein). Conventionally, a system is chaotic if it is sensitive to its initial conditions (cf. Eckmann and Ruelle 1985), although there are quite a few different operational definitions available in the literature (cf. Tong 1990, Nychka et. al. 1992, Wolff 1992 and others).

A simple system that may generate chaos is the deterministic discrete-time dynamical equation

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

for $t \geq 1$, where X_t denotes a state vector in R^d, and F is a real vector-valued function. We assume that each argument of F has bounded continuous second partial derivatives. Usually in chaos study, F is assumed to be bounded, though we do not put forward this assumption explicitly here. It is well known that the existence of at least one positive Lyapunov exponent is a necessary (but not sufficient) condition for a deterministic system to be chaotic. To see this more precisely, let $\{X_t(x),\ t \geq 0\}$ denote the trajectory starting at $X_0 = x \in R^d$, and $x,\ x+\delta$ be two nearby initial values. Then, after m iterates,

$$X_m(x+\delta) - X_m(x) = F^{(m)}(x+\delta) - F^{(m)}(x) \approx \frac{d}{dx^\tau}\{F^{(m)}(x)\}\delta,$$

where $F^{(m)}$ denotes the m-fold composition of F. To simplify the discussion, let us consider the one-dimensional case ($d = 1$) for the moment. Then, by the chain rule, the (local) Lyapunov exponent (at initial value x) is defined as

$$\kappa(x) = \lim_{m\to\infty} \frac{1}{m} \log |\frac{d}{dx}\{F^{(m)}(x)\}| = \lim_{m\to\infty} \frac{1}{m} \sum_{i=0}^{m-1} \log |F'(X_i(x))|, \tag{2.2}$$

if the limit exists (cf. Ruelle 1989). Hence, we have the approximation

$$|X_m(x+\delta) - X_m(x)| \quad \approx \quad |\delta|\, e^{m\kappa(x)}.$$

Thus, if $\kappa(x) > 0$ for some x, two trajectories with nearby initial values around x diverge at an exponential rate. Notice that the above relation usually does not hold for small m because of the asymptotic nature of $\kappa(x)$. Therefore, it could not be used generally in short-term prediction. Neither does it make sense for very large m because of the global boundedness of the attractor; the trajectories simply keep twisting when they diverge. Hence, we would take the view that the Lyapunov exponent is a useful qualitative characteristic indicating possible deterministic chaos rather than a parameter with which we develop specific quantitative analysis. For the prediction study in a stochastic dynamical system, we need to introduce some Lyapunov-like measures to describe the divergence over a *short* time in a statistical setup.

A stochastic dynamical system with additive noise can be described by the equation

$$X_t = F(X_{t-1}) + e_t \tag{2.3}$$

for $t \geq 1$, where F is the same as in (2.1), $\{e_t\}$ is a d-dimensional noise process, and $E(e_t|X_0, \ldots, X_{t-1}) = 0$. We do not attempt to give a rigorous mathematical definition of chaos for such a stochastic system. Operationally, we say that a stochastic system is chaotic if (i) for some $m \geq 1$, the conditional distribution of X_m given $X_0 = x$ depends on the initial value x sensitively for some values of x; (ii) for all $t \geq 1$, the noise e_t is amplified very rapidly starting at some initial values of X_0 through the time evolution. We shall describe manifestations of (i) and (ii) later.

Intuitively, we would expect that the conditional distribution of X_m given $X_0 = x$ can, under certain conditions, depend sensitively on x for some small or moderate rather than large m because of the accumulation of noise through the time evolution. It would seem unlikely that after a long time, the stochastic system still has a strong memory of its initial value, especially when the amount of noise is considerable in comparison with the magnitude of the system signal. We do not rule out the possibility that conditions (i) and (ii) might be equivalent. However, it remains an open problem as to how to formulate them properly and prove (or disprove) the general equivalence. On the other hand, it seems not always proper to say that the stochastic system (2.3) is chaotic if its skeleton (2.1) (cf. Tong 1990) is deterministically chaotic, because if the noise tends to be overwhelming in (2.3), the stochastic system would behave like a stochastic noise process no matter what the skeleton is. A challenging question is to quantify the amount of permissible noise $\{e_t\}$ on the dynamic $F(.)$ without smearing the qualitative characteristics of the latter. A relevant result for continuous-time systems can be found in Kresting (1991).

To prepare the grounds for the prediction study in Section 3, we discuss two characteristic features of stochastic chaotic systems from some special angles in the following Sections 2.2 and 2.3. More general discussion on the sensitivity of the conditional distribution to initial values will be reported elsewhere. Some remarks on 'Lyapunov exponents' in stochastic systems will be given in Section 2.4.

2.2 Lyapunov-like indices

Suppose that $\{X_t,\ t \geq 1\}$ are given as in (2.3). For $x \in R^d$ and $m \geq 1$ let

$$F_m(x) = E(X_m|X_0 = x).$$

Suppose that x and $x+\delta$ are two nearby initial points of the process. Then after time m, the divergence of the conditional expectations of X_m is

$$F_m(x+\delta) - F_m(x) = \Lambda_m(x)\delta + \circ(\| \delta \|), \tag{2.4}$$

where $\Lambda_m(x) = dF_m(x)/dx^\tau$ is a $d \times d$ matrix, especially $\Lambda(x) \equiv \Lambda_1(x) = dF(x)/dx^\tau$. It follows from (2.3) that

$$F_m(x) = E\{F(X_{m-1})|X_0 = x\} = E\{F(F(X_{m-2}) + e_{m-1})|X_0 = x\}$$

$$= E\{\, F(\, \ldots\, (\, F(x)\, +\, e_1\,)\, +\, \ldots\,)\, +\, e_{m-1}) \mid X_0 = x\,\}.$$

By the chain rule of matrix differential, Λ_m can be expressed as

$$\Lambda_m(x) = E\{\prod_{k=1}^{m} \Lambda(X_{k-1}) \mid X_0 = x\}. \tag{2.5}$$

Roughly speaking, assuming that all the factors in the RHS of the above expression are of comparable size, it seems plausible that $\Lambda(x)$ grows (or decays) exponentially with m. Let $\nu_m^2(x)$ denote the largest eigenvalue of $\Lambda_m^\tau(x)\Lambda_m(x)$. It follows from (2.4) that

$$\| F_m(x+\delta) - F_m(x) \| \leq |\nu_m(x)| \, \| \delta \| + \circ(\| \delta \|). \tag{2.6}$$

In the special case $d = 1$, $\nu_m(x) = E\{\prod_{k=1}^m \frac{d}{dx} F(X_{k-1}) \mid X_0 = x\}$, and $F_m(x+\delta) - F_m(x) \approx \nu_m(x)\delta$. Thus, for the values of x such that $|\nu_m(x)|$ is large, a small shift δ in the initial value can lead to considerable divergence in the conditional expectations. This means that the conditional expectation $F_m(x)$ depends on x sensitively when $|\nu_m(x)|$ is large. This can be considered a manifestation of the sensitive dependence of the conditional distribution on the initial value. Since, in practice, the observations will almost certainly be subject to measurement or rounding errors, it seems necessary to take account of this divergence in m-step prediction. However, in the context of prediction, the task is usually to predict one component of X_t instead of the whole vector X_t. Hence, the approximation (2.6) is rather rough. We would concentrate on the divergence in the first component of the system. Let Y_t denote the first component of X_t. It follows from (2.3) that

$$Y_t = f(X_{t-1}) + \epsilon_t,$$

where $f(\cdot)$, and ϵ_t denote respectively the first component of $F(\cdot)$ and the first component of e_t. For $m \geq 1$, and $x \in R^d$, let

$$f_m(x) = E(Y_m|X_0 = x).$$

Obviously, $f_1(x) = f(x)$. Then from (2.4), we have

$$f_m(x+\delta) - f_m(x) = \delta^\tau \lambda_m(x) + \circ(\| \delta \|), \tag{2.7}$$

where $\lambda_m(x) = df_m(x)/dx$, which is equal to the transpose of the first row vector of the matrix $\Lambda_m(x)$ in (2.5). We call $\lambda_m(.)$ the m-step *Lyapunov-like index*, or simply the m-LI. When $d = 1$,

$$\lambda_m(x) = \nu_m(x) = E\{\prod_{k=1}^{m} \frac{d}{dx} F(X_{k-1}) \mid X_0 = x\} = E\{\prod_{k=1}^{m} \lambda_1(X_{k-1}) \mid X_0 = x\}. \tag{2.8}$$

We will see in Section 3 that the m-LI plays an important role in the m-step prediction.

2.3 Noise amplification in a stochastic system

Unlike a deterministic system, a stochastic system is typically contaminated by noise. The amplification of noise varies with the initial values, and is not necessarily monotonic in time evolution. This is a typical feature of nonlinear (but not necessarily chaotic) models. Further, a small noise is expected to be amplified rapidly through the dynamics if the system is chaotic. To highlight this interesting phenomenon, we restrict our discussion here to one-dimensional systems.

Suppose the process begins at the initial value $Y_0 = x \in R$, and for $t \geq 1$,

$$Y_t = f(Y_{t-1}) + \epsilon_t,$$

where $\{\epsilon_t, t \geq 1\}$ is a noise process with

$$E(\epsilon_t | Y_k,\ k < t) = 0, \qquad \sigma^2 \equiv \mathrm{Var}(\epsilon_t) = \mathrm{Var}(\epsilon_t | Y_k,\ k < t). \tag{2.9}$$

Therefore $E(\epsilon_t \epsilon_s | Y_k,\ k < t) = 0$ for all $t > s$. We also assume that for all $t \geq 1$, $|\epsilon_t| < \zeta$ a.s., where $\zeta > 0$ is a small constant. By Taylor's expansion, it is easy to see that for $m \geq 1$,

$$Y_m = f^{(m)}(x) + \epsilon_m + \lambda[f^{(m-1)}(x)]\epsilon_{m-1} + \ldots + \left\{\prod_{k=1}^{m-1} \lambda[f^{(k)}(x)]\right\} \epsilon_1 + O_p(\zeta^2), \tag{2.10}$$

where $\lambda(x) = df(x)/dx$, and $f^{(k)}$ denotes the k-fold composition of f. Let $\sigma_m^2(x) = \mathrm{Var}(Y_m | Y_0 = x)$. Then $\sigma_1^2(x) \equiv \sigma^2$ and for $m > 1$,

$$\sigma_m^2(x) = \mu_m(x)\sigma^2 + O(\zeta^3), \tag{2.11}$$

where

$$\mu_m(x) \ = \ 1 \ + \ \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \lambda[f^{(k)}(x)] \right\}^2. \tag{2.12}$$

If the absolute value of the one-step Lyapunov-like index $\lambda_1(x) = \lambda(x)$ is greater than 1 for a large range of values of x, $\mu_m(x)$ can be very large for moderate (and even small) m (cf. Fig. 3 in Section 4). The rapid increase of $\sigma_m^2(x)$ with respect to m is a manifestation of noise amplification. It is easy to see from (2.8) and (2.7) that in this one-dimensional case, the

conditional expectation also depends on the initial value sensitively when $|\lambda(x)|$ is greater than 1 for a large range of values of x. On the other hand, (2.12) implies that

$$\mu_{m+1}(x) = 1 + \mu_m(x)\{\lambda[f^{(m)}(x)]\}^2.$$

Thus, $\mu_{m+1}(x) < \mu_m(x)$ if $\{\lambda[f^{(m)}(x)]\}^2 < 1 - 1/\mu_m(x)$. By (2.11), it is possible that for such x and m, $\sigma^2_{m+1}(x) < \sigma^2_m(x)$. This suggests that from the same initial value, the error of a $(m+1)$-step ahead prediction could be smaller than that of the m-step ahead prediction in some cases.

In the case that $f(.)$ is linear, $\lambda(.)$ is a constant, and the remainder in the RHS of (2.10) is zero. Since the noise is homogeneous as assumed in (2.9), $\sigma^2_m(x)$ does not depend on x, and is monotonically increasing as m increases.

Deissler and Farmer (1989) have discussed noise amplification of a deterministic system.

2.4 Some remarks

The purpose of this paper is to indicate the dependence of nonlinear prediction on its initial values, which will be further explored later on in this paper. To this end, it is inevitable that we should touch on the notion of a stochastic chaotic system. For a stochastic system as defined in (2.3), it seems to us that the problem as to how to define the Lyapunov exponent is still not completely resolved, although quite a few attempts have been made. For example, Crutchfield et al. (1982) has taken the probability average in the conventional definition of the Lyapunov exponent (initially designed for a deterministic system), which however seems to lose its intuitive appeal. Recently, Nychka et.al. (1992) has adopted a measure which describes the divergence of the trajectories under the assumption that the different trajectories have the *same* realization of random noise. We ourselves have difficulties in perceiving the existence of such trajectories in practice.

Perhaps an alternative direction worthy of exploration is as follows. Suppose the system is one-dimensional ($d = 1$). Define

$$\kappa(x) = \lim_{m\to\infty} \frac{1}{m} \log |\frac{d}{dx} F_m(x)| = \lim_{m\to\infty} \frac{1}{m} \log |E\{ \prod_{i=0}^{m-1} \frac{d}{dx} F(X_i) \mid X_0 = x\}|, \tag{2.13}$$

which can also be expressed in terms of the m-LIs, namely $\kappa(x) = \lim_{m\to\infty} \frac{1}{m} \log |\lambda_m(x)|$. If $\kappa(x)$ exists, we have $|F_m(x+\delta) - F_m(x)| \approx e^{m\kappa(x)}|\delta|$. Positive $\kappa(x)$ entails a possible exponential divergence of the conditional expectations of the trajectories with nearby initial values. When the noise fades away ($e_t \equiv 0$), $\kappa(x)$ reduces to the well-known Lyapunov exponent of a deterministic system (cf. (2.2)). However, $\kappa(x)$ defined in (2.13) only makes sense when the noise

is relatively small, or more precisely the ratio of signal to noise is large, because if the noise becomes overwhelming, the system tends to have a 'short memory' of its history. In this case, $F_m(x) = E\{X_m|X_0 = x\}$ could be nearly a constant when m is sufficiently large. Further, for a stochastic chaotic system, as we mentioned in Subsection 2.1, it makes more practical sense to focus the sensitivity of the conditional distributions (on the initial values) on the short and moderate (rather than the long) terms. This suggests that asymptotics are unlikely to yield a practically useful characteristic exponent. Yet another possible direction is to explore how the ergodic theory of random transformations (e.g. Kifer 1986) could be brought to bear in the present context.

3 m-step prediction

By using the ideas developed in Section 2, we study the prediction of nonlinear time series. From (2.11), we expect that the 'error bounds' of the prediction will vary with the initial value. This is a typical feature of nonlinear (but not necessarily chaotic) model. If the model is stochastically chaotic, (2.12) indicates that a small noise can be amplified quickly when the system starts at some initial values, which means that the m-step prediction based these initial values could be unreliable even for small m. Further, when the m-LI $\lambda_m(x)$ is large, a small change in the initial value x would lead to considerable divergence in the states at time m (cf. (2.7)). In this case, it is worthwhile to take account of the error in prediction due to the measurement error in the initial value.

Since we do not assume any specific form of the model, we choose as our technical tool the nonparametric kernel regression method based on locally linear fit (or simply, locally linear regression, cf. Fan 1992, Ruppert and Wand 1992) to estimate both the prediction functions and their derivatives (i. e. the Lyapunov-like indices) simultaneously. As far as we know, this is the first time that the locally linear regression method has been adopted in a time series context.

3.1 Model

Suppose that $\{Y_t,\ -\infty < t < \infty\}$ is a one-dimensional strictly stationary time series, and the following equality holds for $m \geq 1$

$$E\{\, Y_{t+m} \mid Y_k,\ k \leq t \,\} \ = \ E\{\, Y_{t+m} \mid X_t \,\}, \quad -\infty < t < \infty, \tag{3.1}$$

where $X_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+1})^\tau$, and $d \geq 1$ is a known integer. Obviously, condition (3.1) is weaker than the assumption that $\{Y_t\}$ is a Markov process in d steps. Given the observations

$\{Y_t,\ -d+1 < t \le n\}$, we shall predict the random variables Y_{n+m} for $m = 1, 2, \ldots$. In fact, the time series model can be considered a special case of a stochastic dynamical system. To see this, let $f(x) = E(Y_1|X_0 = x)$. Then Y_t can be expressed as

$$Y_t \ = \ f(X_{t-1}) \ + \ \epsilon_t, \tag{3.2}$$

where $\epsilon_t = Y_t - f(X_{t-1})$. Define $F(X_{t-1}) = (f(X_{t-1}), Y_{t-1}, \ldots, Y_{t-d+1})^\tau$, $e_t = (\epsilon_t, 0, \ldots, 0)^\tau$. Then equation (2.3) holds. In what follows, the time series model is said to be chaotic if the corresponding stochastic dynamic system is chaotic. When $f(.)$ is a linear function, we have the usual linear AR model.

To study the m-step prediction, we define

$$f_m(x) \ = \ E(\ Y_m \mid X_0 = x\ ),$$

for $x \in R^d$ and $m \ge 1$. It follows from (3.1) that for all t, Y_{t+m} can be expressed as

$$Y_{t+m} \ = \ f_m(X_t) \ + \ \epsilon_{t+m}^{(m)},$$

with

$$E\{\ \epsilon_{t+m}^{(m)} \mid Y_k,\ k \le t\ \} \ = \ 0, \qquad \text{a.s.} \tag{3.3}$$

3.2 Predictor

It is easy to see from (3.3) that the (theoretical) least squares predictor of Y_{n+m} based on $\{Y_t, t \le n\}$ is $f_m(X_n)$, which only depends on the latest vector $X_n = (Y_n, \ldots, Y_{n-d+1})^\tau$. In what follows, an estimator of the function $f_m(x)$ is constructed by using the locally linear regression method, which also produces an estimator of the m-LI $\lambda_m(x) \equiv df_m(x)/dx$. The idea of the locally linear regression is very simple: for a small shift $\delta \in R^d$, the equality (2.7) holds. Hence, the estimation problem can be described as a weighted least-squares problem, namely finding f_m and λ_m to minimize

$$\sum_{t=1}^{n-m} \{Y_{t+m} - f_m(x) - \lambda_m^\tau(x)(X_t - x)\}^2 K\left(\frac{x - X_t}{h}\right), \tag{3.4}$$

where $K(\cdot)$ is a probability density function on R^d, and $h = h(n)$ is a bandwidth. Simple calculation yields

$$\hat{f}_m(x) = \{T_0(x) - S_1^\tau(x)S_2^{-1}(x)T_1(x)\}/\{S_0(x) - S_1^\tau(x)S_2^{-1}(x)S_1(x) + h^2\}, \tag{3.5}$$

$$\hat{\lambda}_m(x) = \{S_2(x) - S_1(x)S_1^\tau(x)/S_0(x)\}^{-1}\{S_1(x)T_0(x)/S_0(x) - T_1(x)\}, \tag{3.6}$$

where

$$S_0(x) = \frac{1}{n-m}\sum_{t=1}^{n-m} K\left(\frac{x-X_t}{h}\right), \quad S_1(x) = \frac{1}{n-m}\sum_{t=1}^{n-m}(x-X_t)K\left(\frac{x-X_t}{h}\right),$$

$$S_2(x) = \frac{1}{n-m}\sum_{t=1}^{n-m}(x-X_t)K\left(\frac{x-X_t}{h}\right)(x-X_t)^\tau, \tag{3.7}$$

and

$$T_0(x) = \frac{1}{n-m}\sum_{t=1}^{n-m} Y_{t+m}K\left(\frac{x-X_t}{h}\right), \quad T_1(x) = \frac{1}{n-m}\sum_{t=1}^{n-m}(x-X_t)Y_{t+m}K\left(\frac{x-X_t}{h}\right). \tag{3.8}$$

For technical reasons, we add h^2 into the denominator in (3.5), which has little effect for large n.

3.3 Asymptotic properties

To discuss the asymptotic properties of $\hat{f}_m(x)$ and $\hat{\lambda}_m(x)$, we need the following assumptions.

(A1) All second partial derivatives of $f_m(x)$ are bounded and continuous.

(A2) X_t has the probability density function p, and $|p(x) - p(y)| \leq C \parallel x - y \parallel$ for any $x, y \in R^d$.

(A3) The conditional variance

$$\sigma_m^2(x) = \mathrm{Var}(Y_m | X_0 = x) \tag{3.9}$$

is bounded and continuous.

(A4) $E|Y_t|^{2\alpha} < \infty$ for some $\alpha > 2$.

(A5) Let $\beta_n = E\left[\sup_{A \in \Im_n^\infty} |P(A|\Im_{-\infty}^0) - P(A)|\right]$, where $\Im_k^n$ is the σ-field generated by $\{Y_t : t = k, \ldots, n\}$ $(n \geq k)$. Then $\beta_n = O(n^{-(2+\eta)/\eta})$ for some constant η in $(0, \alpha - 2)$, where α is given in (A4). Furthermore, there exists a positive integer $n_1 = n_1(n)$ such that for $n_2 \equiv [n/(2n_1)] > 0$,

$$\limsup_{n\to\infty}(1 + 6\sqrt{e}\beta_{n_1}^{1/(n_2+1)})^{n_2} < \infty. \tag{3.10}$$

(A6) $K(\cdot)$ is a bounded density function on R^d, and

$$\int xK(x)dx = 0, \quad \int xx^\tau K(x)dx = \sigma_0^2 I_d, \quad \int \parallel x \parallel^8 K(x)dx < \infty,$$

where I_d denotes the $d \times d$ identity matrix.

(A7) The bandwidth $h = n^{-\theta}$ with θ in $(0, (2+d)^{-1}/2)$.

(A8) For n_1 given in (A5) and θ given in (A7), $\limsup_{n\to\infty} n_1/n^{d\theta} < \infty$.

Assumption (A5) implies that the process $\{Y_t\}$ is absolutely regular. The condition $\beta_n = O(n^{-(2+\eta)/\eta})$ is for technical convenience, which is not the weakest possible. Condition (3.10), together with (A7) and (A8), allows us to apply a probability inequality of Roussas and Ioannides (1988) on triangular arrays of weakly dependent random variables. The other conditions are self-explanatory.

Theorem 1. Assume that (A1) – (A8) hold for some $m \geq 1$. Then, for $x \in \{p(x) > 0\}$ and $\delta \in R^d$

$$\lim_{n\to\infty} E\{[Y_{n+m} - \hat{f}_m(x)]^2 | X_n = x + \delta\} \;=\; \sigma_m^2(x+\delta) + \{\delta^\tau \lambda_m(x)\}^2 + R_m, \quad \text{a.s.,} \tag{3.11}$$

where $R_m = \circ(\| \delta \|^2)$ as $\| \delta \| \to 0$, $\lambda_m(x) = df_m(x)/dx$ is the m-LI, and $\sigma_m^2(x)$ is the conditional variance given in (3.9).

Theorem 2. Assume that (A1) – (A8) hold for some $m \geq 1$. For $x \in \{p(x) > 0\}$, as $n \to \infty$, $\hat{\lambda}_m(x)$ converges to $\lambda_m(x)$ in probability.

The proofs of the theorems will be given in Section 5.

Theorem 1 shows that the mean squared error of the predictor $\hat{f}_m$ at the initial value x, which has a small shift from the true but unobservable value $X_n = x + \delta$, can be decomposed into two parts: (i) the conditional variance; (ii) the error due to the small shift at initial value which is related to the m-LI. When $\delta = 0$, i. e. X_n is fully known, (3.11) becomes

$$\lim_{n\to\infty} E\{\, [Y_{n+m} - \hat{f}_m(x)]^2 \,|\, X_n = x \,\} \;=\; \sigma_m^2(x) \quad \text{a.s.,}$$

which shows that the accuracy of the prediction in a nonlinear (but not necessarily chaotic) model does depend on the initial value x, which is strikingly different from the case of a linear prediction. (In a linear model with homogeneous noise as indicated in (2.9), $\sigma_m^2(.)$ is a constant.) But if the model is chaotic, $\sigma_m^2(x)$ can be very large for some moderate or even small m (cf. (2.11) and (2.12)). When the measurement error δ is small but not zero, such as rounding errors in measurement etc., usually the right hand side of (3.11) is dominated by the conditional variance $\sigma_m^2(x+\delta) = \sigma_m^2(x) + O(\| \delta \|)$, because the second term is of the order of $||\delta||^2$. For example, in the case that the model is linear (and stationary), the m-LI $\lambda_m(x)$ is a constant vector with norm less than one and the term $\{\delta^\tau \lambda_m(x)\}^2$ can therefore be ignored. However, for a chaotic system, the m-LI $\lambda_m(x)$ can be very large for some values of x (cf. (2.7) and (2.8)), in which case

the term $\{\delta^\tau \lambda_m(x)\}^2$ can no longer be ignored. In this sense, we say that the m-step prediction is sensitive to the initial values when the model is chaotic.

In (3.2), the noise term ϵ_t is not necessarily homogeneous as indicated in the second expression in (2.9). However if it is, $\sigma_1^2(x) \equiv \sigma_1^2$ is a constant. In this case, the variation of the asymptotic mean squared prediction error is dictated by $\lambda_1(x)$.

Theorem 2 shows that $\hat{\lambda}_m$ is a weakly consistent estimator of the m-LI λ_m. In fact, it can be proved that strong consistency also holds, but the proof will be considerably more complicated. Yao and Tong (1992b) have established the asymptotic normality for more general estimators which include $\hat{f}_m(x)$ and $\hat{\lambda}_m(x)$ as special cases.

In fact, the asymptotic decomposition (3.11) dose not depend on the special choice of $\hat{f}_m$. It also holds when $\hat{f}_m$ is the conventional Nadaraya-Watson estimator (or any other estimator which converges to f_m in mean square). Note that the Nadaraya-Watson kernel estimator for f_m can be interpreted as the weighted least-squares solution of (3.4) with the restriction $\lambda_m \equiv 0$. Fan (1992) has shown that the Nadaraya-Watson estimator has a larger bias (of the order h^2) than the locally linear regression estimator, especially in the case when $\| \lambda_m(x) \|$ is large, which is a typical feature of the chaotic system. However, in most practical cases, the difference between the two methods in estimating f_m is not so obvious. Our preference for the locally linear regression stems mainly from the fact that it offers a natural and convenient estimator for λ_m by virtue of the weighted least-squares formulation around (3.4).

In the above approach, we only consider the effect of measurement errors on the prediction through the initial values. Perhaps it could be argued that we should also consider the effect on the estimation of $f_m(.)$ and $\lambda_m(.)$. This will lead to an interesting exploration of the *robustness* of locally linear regression against measurement errors. Since the resulted effect on prediction is not directly due to the nonlinearity of the model, it will not be considered in this paper.

We will not discuss in any detail how to choose the bandwidth h. A frequently used bandwidth selection technique is the cross-validation method (cf. Stone 1977). A more refined method for local linear smoothers has been developed by Fan and Gijbels (1993).

3.4 Conditional variances

To use (3.11) in practice, we need to estimate the conditional variance $\sigma_m^2(x)$. In principle, we can use the locally linear regression method to estimate the second conditional moment $E(Y_m^2|X_0 = x)$ by

$$\hat{\zeta}_m(x) = \{V_0(x) - S_1^\tau(x)S_2^{-1}(x)V_1(x)\}/\{S_0(x) - S_1^\tau(x)S_2^{-1}(x)S_1(x)\},$$

where $S_k(\cdot)$, $k = 0, 1, 2$, are as given in (3.7), and $V_k(\cdot)$, $k = 0, 1$, are defined in the same way as $T_k(\cdot)$ with Y_{t+m}^2 replacing Y_{t+m} (cf. (3.8)). Now, we get an estimator for σ_m^2,

$$\hat{\sigma}_m^2(x) = \hat{\zeta}_m(x) - [\hat{f}_m(x)]^2, \tag{3.12}$$

where $\hat{f}_m$ is given in (3.5). It can be proved that the estimator $\hat{\zeta}_m$ is consistent under some conditions (cf. Lemma 4 of Yao and Tong 1992a). Therefore, the estimator in (3.12) is also consistent. Any smooth regression method would suggest using different bandwidths for the first and second conditional moments. In practice, for the sake of convenience, we tend to adopt the same bandwidth whilst bearing in mind the possibility of misleading results sometimes (cf. Fig. 3 in Section 4). Note that the positivity of $\hat{\sigma}_m^2(\cdot)$ cannot always be guaranteed even though the same bandwidth is used in estimating the first and second conditional moments.

The discussion in Section 2.2 offers us a tentative way to estimate a 'profile' of $\sigma_m^2(x)$ when the noise terms are small. In the case $d = 1$, it is easy to see from (2.11) that the variation of $\sigma_m(x)$ is dominated by the variation of the functions $\mu_m(x)$. Equation (2.12) suggests the following estimator for μ_m,

$$\hat{\mu}_m(x) = 1 + \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \hat{\lambda}_1[\hat{f}_k(x)] \right\}^2, \tag{3.13}$$

where $\hat{f}_k$ and $\hat{\lambda}_1$ are given in (3.5) and (3.6). Simulations show that this estimator is quite good in small-noise experiments (cf. Fig. 3 in Section 4) and suggest the possibility of avoiding estimating $\sigma_m^2(x)$ directly by estimating $\mu_m(x)$ instead.

4 Examples

We have shown, via asymptotics, that the performance of m-step-ahead prediction is influenced by the initial values. However, its finite-sample behaviour is unknown. In this section, we use two simulated examples and two real data sets to illustrate its finite-sample behaviour. In all of them, the Gaussian kernel is used. In each case, we use the cross validation method to choose a primary value of the bandwidth, and then adjust it upwards by a small amount in the light of Hall and Johnstone (1992). For the sake of simplicity, we use the same bandwidth for estimating both regression function and its derivative. In order to show the performance of $\hat{f}_m(x)$ with $m = 1, 2, 3, 4, \ldots$ etc., we add small amounts of (stochastic) dynamic noise to chaotic deterministic systems in Examples 1 and 2. It is obvious that an 'interval predictor' is much more relevant than a point predictor for a stochastic system, especially in the case of a relatively large noise. Yao and Tong (1992b) have developed nonparametric regression estimation for conditional

percentiles and expectiles, which can be used to construct prediction intervals. We shall report some of the work in this direction elsewhere.

Example 1. We begin with the simple one-dimensional model

$$Y_t = 0.246 Y_{t-1}(16 - Y_{t-1}) + \epsilon_t \qquad t \geq 1, \tag{4.1}$$

where ϵ_t, $t \geq 1$, are independent random variables with the same distribution as the random variable 0.05η, and η is equal to the sum of 48 independent random variables each uniformly distributed on $[-0.5, 0.5]$. According to the central limit theorem, we can treat ϵ_t as being nearly a normal random variable with mean 0 and variance 0.1^2. However, it has a bounded support $[-1.2,\ 1.2]$. Note that bounded support of ϵ_t is necessary for the stationarity of the time series (cf. Chan and Tong 1994). In fact, the skeleton of (4.1) is a transformed logistic map with the coefficient $3.936\,(= 16 \times 0.246)$. We have adopted the transformation in order to enlarge the dynamic range of the model. A sample of 1200 is generated from model (4.1). Note that $\sigma_1^2 \equiv 0.01$; therefore the one-step prediction is uniformly good for different initial values. Hence, the case is not reported here. The scatter plots of Y_{t+m}, for $m = 2, 3, 4$, against Y_t are displayed in Fig. 1, which show obvious change of the variability of Y_{t+m} with respect to the different values of Y_t. For example, in the case $m = 3$, the variability of Y_{t+m} is at its largest when Y_t is around 8, and at its smallest when Y_t is about 5 and 11 (see Fig. 1(b)). We use the first 1000 observations to estimate $f_m(\cdot), \lambda_m(\cdot)$, $\sigma_m^2(\cdot)$ and so on (i.e. $n = 1000$). The last 200 observations are used to demonstrate the quality of prediction. The predicted values for those 200 observations together with their absolute prediction errors and the estimated conditional variance $\hat{\sigma}_m^2(x)$ (cf. (3.12)) are plotted in Fig. 2 for the cases of two, three, and four steps ahead. Since rounding errors in the calculation are below 10^{-6}, the accuracy is dominated by the conditional variance. For example, Fig. 2(b) shows that the three-step-ahead prediction is at its worst when the initial value is around 8, and at its best when the initial value is near 5 or 11, which is in agreement with the observation from Fig. 1(b). Similar remark applies to two-step and four-step predictions. Fig. 3 displays the estimated curves $\hat{\sigma}_m^2(x)$ accompanied by $\mu_m(x)$ (cf. (2.12)) and its estimator $\hat{\mu}_m(x)$ (cf. (3.13)), for $m = 2, 3, 4$. It can be proved that $\sigma_2^2(x)$, like $\mu_2(x)$, attains its maximum at $x = 8$. However, $\hat{\sigma}_2(x)$ is misleading around $x = 8$ (Fig. 3 (a)). Similar remark applies to the cases of $m = 3$ and $m = 4$. Fig. 3 also shows that the estimators $\hat{\mu}_m(x), m = 2, 3, 4$, as given in (3.13), are obviously better than the corresponding $\hat{\sigma}_m^2(x)$. We would suggest relying on $\hat{\mu}_m(x)$ rather than $\hat{\sigma}_m^2(x)$ at least in the small noise case.

Fig. 4(a) shows that $\mu_3(x)$ is less than $\mu_2(x)$ when x is near 5 or 7. A similar situation

can be seen for the estimated conditional variances in Fig. 4(b). Fig. 4(c) shows that at a few initial values near 5 and 11, the absolute errors of the three-step-ahead prediction are smaller than those of the two-step-ahead prediction.

To see how a small shift in the initial values affects the prediction, we round an initial value x to the nearest value from amongst $[x]$, $[x]+0.5$, and $[x]+1$, where $[x]$ denotes the integer part of x. Hence, $|\delta| \leq 0.5$. Fig. 5 shows that for $m = 1, 2$, the absolute prediction error increases as $|\hat{\lambda}_m(x)|$ increases, which is consistent with the asymptotic conclusion presented in Theorem 1. There, $\hat{\lambda}_m$ is estimated by using (3.6). Notice that when x is near 8, the prediction errors are less than 0.5.

(Fig. 1 – Fig. 5 are about here.)

Example 2. We clothe a Hénon map with dynamic noise to obtain

$$Y_t = 6.8 - 0.19Y_{t-1}^2 + 0.28Y_{t-2} + \epsilon_t \quad t \geq 1 \tag{4.2}$$

where ϵ_t, $t \geq 1$, are independent random variables with the same distribution as random variable 0.02η, and η is the same as in Example 1. Hence, ϵ_t can be approximately considered to be normal with mean 0 and variance 0.2^2. But it has a bounded support $[-2.4, 2.4]$. A sample of 1200 observations is generated from this model. Similar to Example 1, the first 1000 observations are used for estimation, and the remaining 200 observations for checking the prediction. Although there are two components for each initial value (or rather initial vector), we only plot the data against its first component, namely Y_{t-1} of (Y_{-1}, Y_{t-2}). Fig. 6 reports the predicted values together with the corresponding true values. The estimated values of the conditional variance at these points are shown in Fig. 7, which indicate the accuracy of the prediction. (Note the occasional negative estimates as discussed in Section 3.4.) For example, when the first component of the initial value is near -6.8 or 6.5, the two-step prediction is good (compare Fig. 6 (a) with Fig. 7 (a)). It can also be seen in Fig. 6 that when the first component of the initial value is near 0, the curve has two branches depending on the signs of the second component. The prediction is evidently better when the second component is negative.

In Fig. 8, we have rounded the first half of the checking sample in the same way as in Fig. 5. (Using the complete checking sample would clutter the figure with too many points.) Note that $\hat{\lambda}_m$ is a two-dimensional vector now and we plot $\| \hat{\lambda}_m \|$ instead of $\hat{\lambda}_m$.

(Fig. 6 – Fig. 8 are about here.)

Example 3. On applying the analysis to the Canadian Lynx data for 1821-1934 (listed in Tong 1990), the results for $m = 1$ and 2 are reported in Table 1. Here, we choose $d = 4$. We use the data for 1821-1924 (i.e. $n = 104$) to estimate $f_m(\cdot), \lambda_m(\cdot)$ etc., and the last 10 data to check the predicted values. The bandwidth is chosen as 0.55 for one-step prediction and 0.50 for two-step prediction. The column under $\hat{\sigma}_2^2$ is not complete due to the omission of a negative estimate. Roughly speaking, the prediction is reasonably good though there is evidence of under-prediction. For the case of one-step ahead, the prediction errors are less than 0.1 when $\| \hat{\lambda}_1(x) \|$ is less than 1. They tend to be larger when $\| \hat{\lambda}_1(x) \|$ is 'large'. Occasionally (e.g. in 1934) the error is small even though $\| \hat{\lambda}_1(x) \|$ is 'large'. For the two-step prediction, $\hat{\sigma}_2^2$ and $\| \hat{\lambda}_2 \|$ provide some indication of the prediction reliability. Typically, in 1927 the values of both $\hat{\sigma}_2^2$ and $\| \hat{\lambda}_2 \|$ are large, and the error of the prediction is also large.

(Table 1 is about here.)

Example 4. In many respects, the Wolf's annual sunspot numbers are known to be quite a challenging set of data (see e.g. Tong 1990). We use the data for 1700-1978 (i.e. $n = 279$) to estimate the predictor function and its related functions, and the data for 1979-1991 (i.e. 13 points) to check the prediction reliability as monitored by $\| \hat{\lambda}_1 \|$. In the fitting, we adopt $d = 4$ and $h = 6.43$. The results are summarized in Table 2. The overall impression is that $\| \hat{\lambda}_1 \|$ tends to be small (around 1 say) for the 'trough-years' and large for the 'peak-years'. With the exception of 1985 and 1987, the prediction reliability is fairly closely monitored by reference to $\| \hat{\lambda}_1 \|$.

(Table 2 is about here.)

5 Proofs

We use the same notations as in Section 3. Further, we always assume that conditions (A1)-(A8) hold for a fixed $m \geq 1$. The proofs of Theorem 1 and Theorem 2 will be presented in a series of lemmas. We only give the proofs for the special case $d = 1$. The case $d > 1$ requires more details but does not involve fundamentally new ideas.

Lemma 1. As $n \to \infty$,

$$E\, S_k(x) = p(x) W_k h^{d+k} + O(h^{d+k+1}) \qquad \text{for} \quad k = 0, 1, 2;$$

$$E\, T_k(x) = f_m(x) p(x) W_k h^{d+k} + O(h^{d+k+1}) \qquad \text{for} \quad k = 0, 1,$$

where $W_0 = 1$, $W_1 = 0 \in R^d$, $W_2 = \sigma_0^2 I_d$, and σ_0^2 is as given in (A5).

Lemma 1 follows from standard techniques in kernel density estimation.

Lemma 2. As $n \to \infty$, for any $\epsilon > 0$

$$P\{\| S_k(x) - ES_k(x) \| / h^{d+k} > \epsilon\} = \circ(h^4) \quad \text{for} \quad k = 0, 1, 2; \tag{5.1}$$

$$\| T_k(x) - ET_k(x) \| / h^{d+k} \xrightarrow{P} 0 \quad \text{for} \quad k = 0, 1. \tag{5.2}$$

Proof. When $d = 1$, we have a uniform expression for $k = 0, 1, 2$,

$$S_k(x) \;=\; \frac{1}{n-m} \sum_{t=1}^{n-m} (x - X_t)^k K\left(\frac{x - X_t}{h}\right).$$

For some large $M_n > 0$, which will be specified later, define

$$S_k^+(x) = \frac{1}{n-m} \sum_{t=1}^{n-m} (x - X_t)^k K\left(\frac{x - X_t}{h}\right) I_{\{|x - X_j| \geq M_n\}},$$

and $S_k^-(x) = S_k(x) - S_k^+(x)$. It follows from assumptions (A2), (A6) that

$$\begin{aligned} E|S_k^+(x)| &\leq \int_{\{|x-y| \geq M_n\}} |x - y|^k K\left(\frac{x - y}{h}\right) p(y)\, dy \\ &= \; p(x)\, h^{k+1} \int_{\{|z| \geq M_n/h\}} |z|^k K(z)\, dz \; (1 + \circ(1)) \\ &\leq C p(x) h^{k+1} \left\{ \int_{\{|z| \geq M_n/h\}} K(z) dz \right\}^{\frac{1}{2}} (1 + \circ(1)), \end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality. Consequently by Chebyshev's inequality and (A6), for any $\epsilon > 0$

$$P\{|S_k^+(x) - ES_k^+(x)|/h^{k+1} > \epsilon\} \;\leq\; 2\epsilon^{-1} h^{-(k+1)} E|S_k^+(x)| \;\leq\; C(h/M_n)^4 (1 + \circ(1)). \tag{5.3}$$

On the other hand, it follows from Theorem 3.1 of Roussas and Ioannides (1988) that

$$P\{|S_k^-(x) - ES_k^-(x)|/h^{k+1} > \epsilon\} \;\leq\; C_2 \exp\{-C_1 \epsilon^2 h^{2(k+1)} (n-m)/M_n^2\} \tag{5.4}$$

for $0 < \epsilon \leq C_3 M_n h^{-(k+1)}/n_1$, where $C_i, \quad i = 1, 2, 3$, are some positive constants. Take $M_n = n^\zeta$ for some $\zeta \in (0, 1/2 - 3\theta)$. Then the right hand side of (5.3) tends to 0 faster than h^4. Assumptions (A7) and (A8) guarantee that there exists some positive constant ϵ for which the inequality (5.4) holds for all $n \geq 1$. Furthermore, the right hand side of (5.4) is also equal to $\circ(h^4)$. The relation (5.1) follows from (5.3) and (5.4) immediately.

Note that for $k = 0, 1$,

$$T_k(x) = \frac{1}{n-m} \sum_{t=1}^{n-m} f_m(X_t)(x - X_t)^k K\left(\frac{x - X_t}{h}\right)$$

$$+\frac{1}{n-m}\sum_{t=1}^{n-m}\epsilon_{t+m}^{(m)}(x-X_t)^k K\left(\frac{x-X_t}{h}\right). \tag{5.5}$$

It follows from (3.3) and (A4) and (A6) that

$$\begin{aligned} & E\left\{\frac{1}{n-m}\sum_{t=1}^{n-m}\epsilon_{t+m}^{(m)}(x-X_t)^k K\left(\frac{x-X_t}{h}\right)\right\}^2 \\ = \; & \frac{1}{(n-m)^2}\sum_{t=1}^{n-m}\sum_{s=(t-m+1)\wedge 1}^{t} E\left\{\epsilon_{t+m}^{(m)}\epsilon_{s+m}^{(m)}(x-X_t)^k(x-X_s)^k\right. \\ & \left.\times K\left(\frac{x-X_t}{h}\right)K\left(\frac{x-X_s}{h}\right)\right\} \;=\; O(h^{2(2k+1)}/n). \end{aligned} \tag{5.6}$$

Therefore

$$\frac{1}{n-m}\sum_{t=1}^{n-m}\epsilon_{t+m}^{(m)}(x-X_t)^k K\left(\frac{x-X_t}{h}\right) = \circ_p(h^{k+1}).$$

Similar to the proof of (5.1), it can be proved that the same limit holds for the first term on the right hand side of (5.5). Consequently, (5.2) holds.

Lemma 3. As $n\to\infty$, for $x\in\{p(x)>0\}$, $E\{\hat{f}_m(x)-f_m(x)\}^2\to 0$.

Proof. We give the proof in the case $d=1$. Let

$$\hat{p}(x) = \{S_0(x) - S_1^2(x)/S_2(x) + h^2\}/h.$$

It follows from Lemmas 1 and 2 that for any $\epsilon>0$

$$P\{|\hat{p}(x)-p(x)|>\epsilon\} = \circ(h^4). \tag{5.7}$$

From (3.5), we have the following inequality

$$\begin{aligned} & E\{\hat{f}_m(x)-f_m(x)\}^2 \;\le\; 4\,E\{[T_0(x)-S_0(x)f_m(x)]h^{-1}/\hat{p}(x)\}^2 \\ + \; & 4\,E\{S_1(x)T_1(x)S_2^{-1}(x)h^{-1}/\hat{p}(x)\}^2 \;+\; 4f_m^2(x)\,E\{S_1^2(x)S_2^{-1}(x)/\hat{p}(x)\}^2 \\ = \; & 4\,(R_1 \;+\; R_2 \;+\; R_3), \qquad \text{say.} \end{aligned} \tag{5.8}$$

By Cauchy-Schwarz inequality,

$$S_1^2(x)/S_2(x)\le S_0(x); \quad S_1(x)T_1(x)/S_2(x)\le\{S_0(x)T_*(x)\}^{\frac{1}{2}}, \tag{5.9}$$

where $T_*(x)=\frac{1}{n-m}\sum_{t=1}^{n-m}Y_{t+m}^2 K\left(\frac{x-X_t}{h}\right)$. Hence, for $x\in\{p(x)>0\}$,

$$\begin{aligned} & R_1 \le (p(x)/2)^{-2}h^{-2}E\{\,[T_0(x)-S_0(x)f_m(x)]^2\,; \quad |\hat{p}(x)-p(x)|\le p(x)/2\,\} \\ + \; & h^{-4}E\{\,[T_0(x)-S_0(x)f_m(x)]^2\,; \quad |\hat{p}(x)-p(x)|>p(x)/2\,\}. \end{aligned} \tag{5.10}$$

Under (A1)-(A4) and (A6), it is easy to prove that $E\{T_0(x)-S_0(x)f_m(x)\}^4/h^4<\infty$. Lemmas 1 and 2 entail that $\{T_0(x)-S_0(x)f_m(x)\}/h \xrightarrow{P} 0$. By the mean convergence theorem, the first term

on the right hand side of (5.10) converges to 0. It follows from Cauchy-Schwarz inequality and (5.7) that the second term on the right hand side of (5.10) also tends to 0. Therefore $R_1 \to 0$. With the inequalities in (5.9), it can be proved in the similar way that R_2 and R_3 also tend to 0. The lemma follows from the inequality (5.8).

Proof of Theorem 1. It is easy to see from Lemma 3 that

$$E\{[\hat{f}_m(x) - f_m(x)]^2 | X_n\} \to 0, \quad \text{a.s. } (P).$$

Hence by Cauchy-Schwarz inequality, assumption (A3) and the definition of f_m, we have

$$\begin{aligned} & E\{[Y_{n+m} - \hat{f}_m(x)]^2 | X_m = x + \delta\} \\ = \; & E\{[Y_{n+m} - f_m(x+\delta) + f_m(x+\delta) - f_m(x) + f_m(x) - \hat{f}_m(x)]^2 | X_m = x + \delta\} \\ = \; & \sigma_m^2(x+\delta) + \{f_m(x+\delta) - f_m(x)\}^2 + \circ(1). \qquad \text{a.s. } (P). \end{aligned}$$

The theorem follows from the simple Taylor's expansion of $\{f_m(x+\delta) - f_m(x)\}^2$.

Proof of Theorem 2. In the case $d = 1$, (3.6) can be expressed as

$$\hat{\lambda}_m(x) = \{S_1(x)T_0(x) - S_0(x)T_1(x)\}/\{S_2(x)S_0(x) - S_1^2(x)\}. \tag{5.11}$$

It follows from (5.5) and (5.6) that

$$\begin{aligned} S_1(x)T_0(x) - S_0(x)T_1(x) \; &= \; \frac{1}{(n-m)^2} \sum_{t \neq s} \{f_m(X_t) - f_m(X_s)\}(x - X_s) \\ \times K\left(\frac{x - X_t}{h}\right) K\left(\frac{x - X_s}{h}\right) + \circ_p(h^4) \; &= \; h^4 \{ \, R \, + \, \circ_p(1) \, \}, \qquad \text{say.} \end{aligned} \tag{5.12}$$

For $s, t, \geq 1$, define

$$H(X_t, X_s) = \{f_m(X_t) - f_m(X_s)\}(X_t - X_s) K\left(\frac{x - X_t}{h}\right) K\left(\frac{x - X_s}{h}\right),$$

$$H(X_t) = \int H(X_t, y) p(y)\, dy, \quad \text{and} \quad H_0 = E\{H(X_t)\}.$$

It follows from assumptions (A3) and (A4) that

$$E|H(X_t, X_s)|^\alpha < \infty, \qquad \text{and} \qquad \int |H(y,z)|^\alpha p(y) p(z)\, dy\, dz < \infty,$$

where α is as given in (A4). By (A5) and Lemma 2 of Yoshihara (1976),

$$\frac{1}{(n-m)^2} \sum_{t \neq s} \{H(X_t, X_s) - H(X_t) - H(X_s) + H_0\} = \circ(n^{-1}). \tag{5.13}$$

Similar to the proof of Lemma 2, it can be shown that

$$\frac{1}{n-m}\sum_{t=1}^{n-m}\{H(X_t)-H_0\}=\circ_p(h^4).$$

Together with (5.13), we have

$$\begin{aligned}R &= \frac{h^{-4}}{2(n-m)^2}\sum_{t\neq s}\{H(X_t,X_s)-H(X_t)-H(X_s)+H_0\}\\ &+ \frac{h^{-4}}{(n-m)^2}(n-m-1)\sum_{t=1}^{n-m}\{H(X_t)-H_0\} \;-\; \frac{1}{2}\,\frac{n-m-1}{n-m}h^{-4}H_0\\ &= -\frac{1}{2}h^{-4}H_0 \;+\; \circ_p(1).\end{aligned}$$

Some integration operations yield

$$H_0=-h^4p(x)\int\frac{f_m(x-hz)-f_m(x-hy)}{h(y-z)}(y-z)^2K(y)K(z)\,dy\,dz\,(1+\circ(1)).$$

Assumptions (A1) and (A6) imply that for all sufficiently small h,

$$\int\frac{|f_m(x-hz)-f_m(x-hz)|}{|h(y-z)|}(y-z)^2K(y)K(z)\,dy\,dz \;<\; C,$$

where C is a finite constant independent of h. Therefore,

$$H_0 \;\sim\; -h^4p(x)\lambda_m(x)\int(y-z)^2K(y)K(z)\,dy\,dz \;=\; -2h^4\sigma_0^2p(x)\lambda_m(x).$$

Consequently, $R \xrightarrow{P} \sigma_0^2p(x)\lambda_m(x)$. On the other hand, it follows from Lemmas 1 and 2 that $\{S_2(x)S_0(x)-S_1^2(x)\}/h^4 \xrightarrow{P} \sigma_0^2p(x)$. The theorem follows from (5.11) and (5.12) immediately. This completes the proof of Theorem 2.

Acknowledgements. We thank an associate editor and a referee for their constructive and exceptionally thorough reports, which have inspired us to think deeper on the notion of noisy chaos, and also improved the presentation of the paper considerably. We also thank J. Fan and D. Ruppert for sending us their papers before publication.

References

Berliner, L.M. (1992). Statistics, probability and chaos. *Statistical Science*, **7**, 69-122.

Chan, K.S. and Tong, H. (1994). A note on noisy chaos. To appear in *J. R. Statist. Soc.* **B**.

Chatterjee, S. and Yilmaz, M.R. (1992). Chaos, fractals and statistics. *Statistical Science*, **7**, 49-121.

Crutchfield, J.P., Farmer, J.D. and Huberman, B.A. (1982). Fluctuations and simple chaotic dynamics. *Phys. Rep.*, **92**, 45-81.

Deissler, R.J. and Farmer, J.D. (1989). Deterministic noise amplifiers. Tech. Rep., LA-UR-89-4236, Los Alamos Laboratory, USA.

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

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Ameri. Statis. Assoc.*, **87**, 998-1004.

Fan, J. and Gijbels, I. (1993). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. Technial Report, Univ. of North Carolina.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection. *J. R. Statist. Soc.* **B**, **54**, 475-530.

Kersting, G. (1991). How much noise is sufficient to disturb a dynamical system? Technical Report, University of Frankfurt.

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

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

Roussas, G.G. and Ioannides, D. (1988). Probability bounds for sums in triangular arrays of random variables under mixing conditions. *Statistical Theory and Data Analysis II*, ed. K. Matusita, North Holland, 293-308.

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

Ruppert, D. and Wand, M.P. (1992). Multivariate locally weighted least squares regression. Tech. Report.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595-620.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.

Wiener, N. (1956). *Non-Linear Prediction and Dynamics*. Proc. Third Berkeley Symposium on Math. Stat. and Prob. III, 247-252.

Wolff, C.L. (1992). Local Lyapunov exponents: looking closely at chaos. *J. R. Statis. Soc.* **B**, **54**, 353-272.

Yao, Q. and Tong, H. (1992a). On subset selection in non-parametric stochastic regression. To appear in *Statistic Sinica*.

Yao, Q. and Tong, H. (1992b). Asymmetric least squares regression estimation: a nonparametric approach. Technical Report ukc/ims/s92/12a, Univ. of Kent.

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular processes. *Z. Wahr. v. Gebiete*, **35**, 237-252.

Figure Captions

Fig. 1 The scatter plots of Y_{t+m} against Y_t for (a) $m = 2$; (b) $m = 3$; (c) $m = 4$.

Fig. 2 The plots of the 200 m-step predicted values and the corresponding absolute prediction errors against their initial values, as well as the estimated conditional variance $\hat{\sigma}_m^2(x)$: (a) $m = 2$ ($h = 0.13$); (b) $m = 3$ ($h = 0.09$); (c) $m = 4$ ($h = 0.07$). Diamonds — predicted values; impulses — absolute prediction errors; solid curve — $\hat{\sigma}_m^2(x)$.

Fig. 3 The estimated conditional variance $\hat{\sigma}_m^2(x)$, the function $\mu_m(x)$ and its estimator $\hat{\mu}_m(x)$: (a) $m = 2$; (b) $m = 3$; (c) $m = 4$. Solid curve — $\hat{\sigma}_m^2(x)$; dashed curve — $\mu_m(x)/c_m$ ($c_2 = 30, c_3 = 50, c_4 = 70$); dotted curve — $\hat{\mu}_m(x)/c_m$.

Fig. 4 (a) The function $\mu_m(x)$. Solid curve — $\mu_2(x)$; dashed curve — $\mu_3(x)$. (b) The estimated conditional variance $\hat{\sigma}_m^2(x)$. Solid curve — $\hat{\sigma}_2^2(x)$; dashed curve — $\hat{\sigma}_3^2(x)$. (c) The absolute errors of m-step-ahead predictions against initial values. Solid impulses (upwards) — absolute errors of two-step-ahead prediction; dashed impulses (downwards) — absolute errors of three-step-ahead prediction.

Fig. 5 The plots of absolute prediction errors against their (rounded) initial values, and the estimated function $|\hat{\lambda}_m(x)|$; (a) $m = 1$ ($h = 0.19$); (b) $m = 2$. Diamonds — absolute errors; dashed curve — $|\hat{\lambda}_m(x)|$.

Fig. 6 The plots of the 200 m-step predicted values and the corresponding true values against the first component of their initial values: (a) $m = 2$ ($h = 0.47$); (b) $m = 3$ ($h = 0.45$). Diamonds — predicted values; crosses — true values.

Fig. 7 The plots of the 200 estimates values of $\hat{\sigma}_m^2$ against the first component of their initial values: (a) $m = 2$; (b) $m = 3$.

Fig. 8 The plots of the 100 absolute prediction errors and the corresponding estimated values $\| \hat{\lambda}_m \|$ against the first component of their first (rounded) initial values: (a) $m = 1$ ($h = 0.5$); (b) $m = 2$. Diamonds — predicted errors; crosses — $\| \hat{\lambda}_m \|$. (Note that some of the initial values, after rounding, may be coincident. This leads to fewer crosses than diamonds in some columns.)

Table 1

Prediction of the Canadian Lynx data (on natural log scale)

Year	True value	error ($\hat{f}_1$)	$\| \hat{\lambda}_1 \|$	error ($\hat{f}_2$)	$\hat{\sigma}_2^2$	$\| \hat{\lambda}_2 \|$
1925	8.18	-0.05	0.58	-0.13	0.08	0.77
1926	7.98	-0.23	2.67	-0.39	0.69	1.04
1927	7.34	-0.16	2.49	-0.60	1.99	4.21
1928	6.27	0.22	3.12	0.13	1.60	2.30
1029	6.18	-0.43	1.94	-0.45	0.61	3.42
1930	6.50	-0.28	2.34	-0.60	–	3.38
1931	6.91	-0.19	1.23	-0.46	0.37	2.35
1932	7.37	0.02	0.70	-0.21	1.17	1.43
1933	7.88	-0.26	1.21	-0.22	0.08	0.59
1934	8.13	-0.07	2.28	-0.22	0.51	2.02

Table 2

Prediction of the Sunspot numbers

Year	True value	error ($\hat{f}_1$)	$\| \hat{\lambda}_1 \|$
1979	155.4	- 8.88	2.64
1980	154.7	-47.26	6.32
1981	140.5	- 5.83	2.98
1982	115.9	-32.0	12.59
1983	66.6	2.80	1.10
1984	45.9	1.01	0.96
1985	17.9	17.94	1.16
1986	13.4	-2.57	0.64
1987	29.2	-19.73	0.92
1988	100.2	-53.67	3.92
1989	157.6	35.56	8.27
1990	142.6	34.51	9.84
1991	145.7	-11.63	3.08