CHAPTER 3.

ON PREDICTION AND CHAOS IN STOCHASTIC SYSTEMS

Qiwei Yao and Howell Tong
Institute of Mathematics and Statistics
University of Kent
Canterbury, Kent CT2 7NF, U.K.

1. Introduction

It is well known that deterministic chaos is characterized by the sensitivity to initial conditions. However, it is increasingly recognized that a purely deterministic system rarely exists in reality because stochastic noise is ubiquitous; even the computer generation of time series using a purely deterministic chaotic map cannot be free from rounding errors. Accordingly, it is more pertinent to replace the dynamics by the transition probabilities from states to states. Indeed, Chan and Tong (1994) has shown how a deterministic dynamical system which admits a compact attractor can, in a noisy environment, give rise to an ergodic stochastic system. A convenient framework for this stochastic system is the Markov chain over general state space. As a result, nonlinear autoregressive models emerge quite naturally as a realization of this framework for the study of noisy chaos. Within this considerably enlarged stochastic framework, a new notion of sensitivity to initial conditions has to be developed because the distribution of the states should now be more relevant than the positions of the states (in a single realization). This generalized notion should ideally have points of contact with conventional statistical inference, because it is intuitively clear that if we treat the initial value as an unknown parameter, the more information the state variable at a later time point has about the parameter the more sensitively the associated conditional distribution depends on the initial value.

Point predictions are only the first step in any serious study of the subject. The complete picture can only be provided if the predictive distribution is available. It is therefore of substantial practical importance to estimate the interval predictors and the predictive distribution from the observed series and, if possible, to provide

indicators of their sensitivity to initial conditions. Being different from the linear case, the nonlinear prediction has three interesting features: (i) the *dependence* of the current position in the state space; (ii) the *sensitivity* to the current state; and (iii) the *non-monotonicity* of the accuracy in multi-step prediction (cf. Yao and Tong 1994).

The rest of the paper is set up as follows. In Section 2, we study chaos in a stochastic environment, discussing noisy chaos and noise amplification. Section 3 develops nonlinear prediction with a view to the estimation, based on the observed time series, of the predictive distribution and measures of initial-value sensitivities. We illustrate our methodology with both simulated data and real data in Section 4. Some technical conditions are relegated to the appendix.

2. Stochastic Dynamic System

2.1. Noisy Chaos

It is impossible for us to study nonlinear prediction without touching on *stochastic chaotic systems*, or simply *noisy chaos*. There has been no generally accepted definition of chaos in a stochastic system (or even in a deterministic system for that matter), although the term noisy chaos has appeared in the literature from time to time. It is often implicitly assumed that by a stochastic chaotic system is meant a system with a (deterministically) chaotic skeleton (cf. Tong 1990, and others). Intuitively, this assumption does make sense when the noise is additive and the ratio of the noise to the signal in the system is (very) small. However, it does not apply to non-additive noise systems. Further, it is not always proper even for an additive noise system, because in a stochastic system, the dynamic noise will, by permeating through the system dynamics, interact with the system signal throughout the time evolution. Therefore, a proper description must take account of the effect of the random noise. An extreme case is that if the additive noise tends to be overwhelming, the system would behave like a noise process no matter what the skeleton is.

A discrete-time stochastic dynamical system can be describeed by the equation

$$X_t \;=\; F(X_{t-1},\, e_t), \tag{1}$$

for $t \geq 1$, where X_t denotes a state vector in R^d, F is a real vector-valued function, and $\{e_t\}$ is a noise process which satisfies the equality $E(e_t|X_0, \ldots, X_{t-1}) = 0$. If the noise is additive, (1) can be written (by an abuse of notation) as follows

$$X_t \;=\; F(X_{t-1}) \;+\; e_t, \tag{2}$$

To understand the difficulty in giving a general definition of noisy chaos, let us start with deterministic chaos first. It is almost impossible to give a precise mathematical definition of deterministic chaos which encapsulates all that the term implies in the diverse literature. However, it is widely accepted that the sensitive dependence on

initial conditions is a typical feature of a (deterministic) chaotic system, and which can be characteristically described in terms of the well-known Lyapunov exponents (cf. Eckmann and Ruelle 1985, Chatterjee and Yilmaz 1992, Berliner 1992, and the references therein). We do not attempt to give a rigorous mathematical definition of chaos for a stochastic system. Instead, *as a working definition, we say that a stochastic dynamic system is chaotic if the (conditional) distribution of the state variable of the system is sensitive to its initial condition.*

The above description is very rudimentary, and it needs further careful exposition. Superficially, it looks similar to the deterministic case. However, in a stochastic system, 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 conditions. This suggests that asymptotics are unlikely to yield a practically useful characteristic exponent, unless we assume that the different trajectories have the same realization of random noise. Under this assumption, Lyapunov exponents could be defined in a way very much similar to those in deterministic systems, which were initially proposed by Crutchfield *et al.* (1982) and Kifer (1986). (Also see Nychka *et al.* 1992). Obviously, the assumption of the same realization of random noise in different trajectories has an innate drawback in that in practice such trajectories rarely exist.

One way to manifest the sensitivity of the conditional distribution is to use the Kullback-Leibler information. To simplify our discussion, we suppose the system variables are bounded. Let $g_m(y|x)$ denote the conditional density of X_m given $X_0 = x$. We suppose that $g_m(y|x)$ is smooth enough in x. For two nearby initial points x, $x + \delta \in R^d$, after time $m \geq 1$, the divergence of the conditional distribution of X_m is defined as

$$K_m(x;\delta) = \int \{g_m(y|x+\delta) - g_m(y|x)\} \log\{g_m(y|x+\delta)/g_m(y|x)\} dy. \tag{3}$$

Note that it is well known that the Kullback-Leibler information is invariant under one-one differentiable co-ordinate transformations. Therefore, if $Z_t = \phi(X_t)$ for all t, and $\phi(.)$ is a one-one differentiable transformation on R^d, then

$$K_m(x;\delta) = K_m^*(\phi(x), \phi(x+\delta) - \phi(x)),$$

where $K_m^*(.;.)$ denotes the divergence of conditonal distribution of Z_m, which is defined in the same way as in (3). (Cf. Theorem 2.4.1 of Kullback 1967.)

It is known that for small δ, $K_m(x;\delta)$ has the approximation

$$K_m(x;\delta) = \delta^T I_m(x)\delta + o(||\delta||^2), \tag{4}$$

where

$$I_m(x) = \int \dot{g}_m(y|x)\dot{g}_m^T(y|x)/g_m(y|x)dy, \tag{5}$$

and $\dot{g}_m(y|x)$ denotes $dg_m(y|x)/dx$, $\dot{g}_m^T(y|x)$ denotes its transpose (cf. §2.6 of Kullback 1967). If we treat the initial value x as a parameter vector of the distribution, $I_m(x)$ is the Fisher's information matrix, which represents the information on initial value $X_0 = x$ contained in X_m. Roughly speaking, the more information X_m brings, the more sensitively the distribution depends on the initial condition. The converse is also true. Fan, Yao and Tong (1993) has given another measure of sensitivity in the form of an L_2 norm. In theory, there is no difficulty in adopting any norm to measure the distance between $g_m(y|x+\delta)$ and $g_m(y|x)$. However, in practice, interpretability may influence the specific choice.

Besides the divergence of distributions, it is also interesting to look at the divergence in some summarizing characteristics, for example the (conditional) means. For $x \in R^d$ and $m \geq 1$, let $F_m(x) = E(X_m|X_0 = x)$. Then for $\delta \in R^d$,

$$F_m(x+\delta) - F_m(x) = \dot{F}_m(x)\delta + o(||\delta||), \tag{6}$$

where $\dot{F}_m(x)$ denotes $dF_m(x)/dx^T$. For the system with additive noise, $F_1(x) = F(x)$, and it follows from (2) 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, $\dot{F}_m(x)$ can be expressed as

$$\dot{F}_m(x) = E\{\prod_{k=1}^{m} \dot{F}(X_{k-1}) \mid X_0 = x\}. \tag{7}$$

Roughly speaking, assuming that all the factors on the RHS of (7) are of comparable size, it seems plausible that $\dot{F}_m(x)$ grows (or decays) exponentially with m. Let $\nu_m^2(x)$ denote the largest eigenvalue of $\{\dot{F}_m(x)\}^T\dot{F}_m(x)$. It follows from (6) that

$$||F_m(x+\delta) - F_m(x)|| \leq |\nu_m(x)|\,||\delta|| + o(||\delta||),$$

which indicates that the conditional expectation $F_m(x)$ depends on x sensitively when $|\nu_m(x)|$ is large. 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 above approximation is rather rough. Let Y_t denote the first component of X_t. It follows from (2) 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 (6), we have

$$f_m(x+\delta) - f_m(x) = \delta^T\lambda_m(x) + \circ(\parallel \delta \parallel), \tag{8}$$

where $\lambda_m(x) = df_m(x)/dx$. We call $\lambda_m(.)$ the m-step *Lyapunov-like index*, or simply the m-LI (Yao and Tong 1994). When $d = 1$,

$$\lambda_m(x) = \nu_m(x) = E\{\prod_{k=1}^{m} \frac{d}{dx} f(X_{k-1}) \mid X_0 = x\} = E\{\prod_{k=1}^{m} \lambda_1(X_{k-1}) \mid X_0 = x\}. \quad (9)$$

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

So far we have derived some indices which describe the sensitive dependence of the conditional distributions on the initial conditions. However, this falls short of a mathematical definition of noisy chaos. The essential barrier lies with the fact that the 'clear' cutoff between deterministic chaotic systems and deterministic non-chaotic systems becomes vague after stochastic noise comes into play in the systems. To see this, let us consider the following two models

$$\begin{aligned} \text{Model I}: \quad & Y_t = 0.230 Y_{t-1}(16 - Y_{t-1}) + 0.4\epsilon_t; \\ \text{Model II}: \quad & Y_t = 0.222 Y_{t-1}(16 - Y_{t-1}) + 0.4\epsilon_t, \end{aligned} \quad (10)$$

where $\{\epsilon_t\}$ is a sequence of independent random variables with the standard normal distribution truncated in the interval $[-12,\ 12]$. It is known that the skeleton of Model I is chaotic (cf. Hall and Wolff 1993, for example). However the skeleton of Model II is not chaotic, but is a limit cycle with period 8. The scatter plots of 3000 data points generated from each of the above models are displayed in Figures 1a and 1b, which show that the difference between these two models is more in quantity than in quality. The random noise masks the distinction in the structures of the skeletons.

2.2. Noise Amplification

Another intersting feature of a nonlinear system is noise amplification. We measure the amplification of noise by comparing the conditional variance of the system variables $\{X_t\}$ (given the initial conditions X_0) with the variance of the innovations $\{e_t\}$. We will see that the amplification of noise varies with the initial values, and is not necessarily monotonic in time evolution. In fact, the sensitive dependence on the initial values and the noise amplification are related to each other, and both of them are dictated by some functions of the derivatives of $F(.)$. In this sense, we say that a small noise is expected to be amplified rapidly through the dynamics if the system is chaotic.

Deissler and Farmer (1991) studied the noise amplification in a different way. They considered the distance between the state variables in a purely deterministic system and the counterparts in the system perturbed by additive system noise. This approach seems improper in the statistical context, because the underlying deterministic skeleton is unknown.

To highlight how nonlinear dynamics amplify noise without going through too much technical detail, we restrict our discussion here to the one-dimensional system

(a)

(b)

Figure 1: The scatter plots of Y_{t+1} against Y_t for (a) Model I; (b) Model II.

with additive noise. Suppose that 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 mean value 0 and variance σ_0^2. Suppose that ϵ_t, is distributed on a bounded set which is independent of t. Then for $\sigma_m^2(x) \equiv \mathrm{Var}(Y_m|Y_0 = x)$, it can be proved that as $\sigma_0 \to 0$,

$$\sigma_m^2(x) = \sigma_0^2 \, \mu_m(x) \, (1 + o(1)), \tag{11}$$

where

$$\mu_m(x) \;=\; 1 \;+\; \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \dot{f}[f^{(k)}(x)] \right\}^2 . \tag{12}$$

(Cf. Yao and Tong 1994.) It is easy to see that if the absolute value of $\dot{f}(x)$ 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. The rapid increase of $\sigma_m^2(x)$ with respect to m is a manifestation of noise amplification. On the other hand, (12) implies that

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

Thus, $\mu_{m+1}(x) < \mu_m(x)$ if $\{\dot{f}[f^{(m)}(x)]\}^2 < 1 - 1/\mu_m(x)$. By (11), it is possible that for such x and m, $\sigma_{m+1}^2(x) < \sigma_m^2(x)$. This suggests that from the same initial value, the error of an $(m+1)$-step ahead prediction could be smaller than that of an m-step ahead prediction in some cases (cf. Tong 1990, Yao and Tong 1994).

3. Nonlinear Prediction

By using the ideas developed in Section 2, we study the prediction of nonlinear time series. From (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, (12) indicates that a small noise could be amplified quickly when the system starts at some initial values, which means that the m-step prediction based on these initial values could be unreliable even for small m. Further, when the system is chaotic, a small change in the initial value x would lead to considerable divergence in the states at time m (cf. (8) and (9)). In this case, it is worth our while to take account of the error in prediction because there is always some 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) to estimate both the prediction functions and their derivatives (i. e. the Lyapunov-like indices) simultaneously. However, it does not mean that our results only hold for these particular estimates. In specific practical applications, parametric (nonlinear) models would be more appealing provided that they could be properly justified. Our results can be easily extended to these cases.

3.1. Model

Suppose that $\{Y_t,\ -\infty < t < \infty\}$ is a one-dimensional strictly stationary time series, which has the property that given $\{Y_i,\ i \le t\}$, the conditional distribution of Y_{t+1} depends on $\{Y_i,\ i \le t\}$ only through X_t, where $X_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+1})^T$. 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

$$Y_t \ = \ f(X_{t-1}) \ + \ \epsilon_t, \tag{13}$$

where $\epsilon_t \ = \ Y_t \ - \ f(X_{t-1})$. Define $F(X_{t-1}) \ = \ (f(X_{t-1}), Y_{t-1}, \ldots, Y_{t-d+1})^T$, $e_t \ = \ (\epsilon_t, 0, \ldots, 0)^T$. Then equation (2) holds. In what follows, the time series model is said to be chaotic if the corresponding stochastic dynamic system is chaotic.

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 (13) that for all t, Y_{t+m} can be expressed as

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

with

$$E\{\ \epsilon^{(m)}_{t+m} \mid Y_k,\ k \le t\ \} \ = \ 0, \qquad \text{a.s.} \tag{14}$$

3.2. Point Predictors

It is easy to see from (14) 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 $X_n = (Y_n, \ldots, Y_{n-d+1})^T$. 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 (8) holds. Hence, based on the observations $(Y_{-d}, Y_{-d+1}, ..., Y_n)$, the estimation problem can be described as a weighted least-squares problem, namely finding f_m and λ_m to minimize

$$\sum_{t=1}^{n-m} \Big\{Y_{t+m} - f_m(x) - \lambda_m^T(x)(X_t - x)\Big\}^2 K\Big(\frac{X_t - x}{h}\Big), \tag{15}$$

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

$$\hat{f}_m(x) = \{T_0(x) - S_1^T(x)S_2^{-1}(x)T_1(x)\}/\{S_0(x) - S_1^T(x)S_2^{-1}(x)S_1(x) + h^2\}, \tag{16}$$

$$\hat{\lambda}_m(x) = \{S_2(x) - S_1(x)S_1^T(x)/S_0(x)\}^{-1}\{S_1(x)T_0(x)/S_0(x) - T_1(x)\}, \tag{17}$$

where

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

$$S_2(x) = \frac{1}{n-m}\sum_{t=1}^{n-m}(x - X_t)K\Big(\frac{X_t - x}{h}\Big)(x - X_t)^T, \tag{18}$$

and

$$T_0(x) = \frac{1}{n-m}\sum_{t=1}^{n-m} Y_{t+m}K\Big(\frac{X_t - x}{h}\Big), \quad T_1(x) = \frac{1}{n-m}\sum_{t=1}^{n-m}(x - X_t)Y_{t+m}K\Big(\frac{X_t - x}{h}\Big). \tag{19}$$

For technical reasons, we add h^2 into the denominator in (16), which has little effect for large n.

Theorem 1. Assume that conditions (A1) – (A7) hold for some $m \geq 1$, which are listed in Appendix.

(i) For $x \in \{p(x) > 0\}$ and $\delta \in R^d$

$$\lim_{n\to\infty} E\{[Y_{n+m} - \hat{f}_m(x)]^2 | X_n = x + \delta\} \;=\; \sigma_m^2(x+\delta) + \{\delta^T \lambda_m(x)\}^2 + R_m, \quad \text{a.s.,} \tag{20}$$

where $R_m = \circ(\| \delta \|^2)$ as $\| \delta \| \to 0$, $\lambda_m(x) = df_m(x)/dx^T$ is the m-LI, and $\sigma_m^2(x) = \mathrm{Var}(Y_m | X_0 = x)$.

(ii) For $x \in \{p(x) > 0\}$, as $n \to \infty$, $\hat{\lambda}_m(x)$ converges to $\lambda_m(x)$ in probability.

For the proof, we can easily adapt the arguments in Yao and Tong (1994) which were based on different assumptions on the mixing condition. The first part of 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 the initial value which is related to the m-LI. When $\delta = 0$, i. e. X_n is fully known, equation (20) 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. When the measurement error δ is small but not zero, such as rounding errors in measurement etc., usually the right hand side of (20) is dominated by the conditional variance $\sigma_m^2(x+\delta) = \sigma_m^2(x) + O(\| \delta \|)$. However, for a chaotic system, the m-LI $\lambda_m(x)$ can be very large for some values of x (cf. (7) and (8)), in which case the term $\{\delta^T \lambda_m(x)\}^2$ 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 fact, the asymptotic decomposition (20) dose not depend on the special

choice of $\hat{f}_m$. It holds for any estimator which converges to f_m in mean square. 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 (15).

In (13), the noise term ϵ_t is not necessarily homogeneous as indicated in the second expression in (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 1 (ii) says that $\hat{\lambda}_m$ is a weakly consistent estimator of the m-LI λ_m. In fact, the estimate of λ_m can be improved by using the locally quadratic regression instead of the locally linear regression (cf. Fan et. al. 1993). We use the latter for the simplicity in calculation.

To use (20) 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

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

where $S_k(\cdot)$, $k = 0, 1, 2$, are as given in (18), 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. (19)). Now, we get an estimator for σ_m^2,

$$\hat{\sigma}_m^2(x) = \hat{\zeta}_m(x) - [\hat{f}_m(x)]^2, \tag{21}$$

where $\hat{f}_m$ is given in (16). However, 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. Yao and Tong 1994). Note that the positivity of $\hat{\sigma}_m^2(\cdot)$ cannot always be guaranteed even if 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 (11) that the variation of $\sigma_m(x)$ is dominated by the variation of the functions $\mu_m(x)$. Equation (12) suggests the following estimator for μ_m,

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

where $\hat{f}_k$ and $\hat{\lambda}_1$ are given in (16) and (17). Simulations show that this estimator is quite good in small-noise experiments (cf. Yao and Tong 1994).

3.3. Interval Predictors

In a stochastic system, an interval predictor is much more relevant than a point predictor, especially in the case of a relatively large noise. A natural way to construct

a predictive interval is to estimate the conditional percentiles of Y_m given X_0. Specifically, for $\alpha \in [0,1]$, the 100α-th conditional percentiles of Y_m given $X_0 = x \in R^d$ is defined as

$$\xi_{\alpha,m}(x) \;=\; \arg\min_{|a|<\infty} E\{\, R_\alpha(Y_m - a) \mid X_0 = x \,\},$$

where the loss function

$$R_\alpha(y) = \begin{cases} (1-\alpha)|y| & y \le 0, \\ \alpha|y| & y > 0. \end{cases} \tag{22}$$

It is well known that the relation $\alpha = P\{Y_m \le \xi_{\alpha,m}(x) | X_0 = x\}$ holds. Therefore, given $\{Y_t,\ t \le n\}$, Y_{n+m} will be in the interval $[\xi_{\alpha/2,m}(X_n),\ \xi_{1-\alpha/2,m}(X_n)]$ with probability $1-\alpha$. In fact, the conditional distribution of Y_{n+m} given X_n is determined by the values of $\xi_{\alpha,m}(X_n)$ for $0 \le \alpha \le 1$.

Similar to Section 3.2, we use the locally linear regression to estimate $\xi_{\alpha,m}(.)$ as well as its derivative $\dot{\xi}_{\alpha,m}(.) \in R^d$. More precisely, we use the estimators $\hat{\xi}_{\alpha,m}(x) = \hat{a}$ and $\hat{\dot{\xi}}_{\alpha,m}(x) = \hat{b}$, by setting $(\hat{a},\ \hat{b})$ as the minimizer of the function (with respect to a and b respectively)

$$\sum_{t=1}^{n-m} R_\alpha\{Y_{t+m} - a - b^T(X_t - x)\} K\left(\frac{X_t - x}{h}\right), \tag{23}$$

where $K(.)$ is a probability density function on R^d, and $h = h(n)$ is a bandwidth. Unlike (15), (23) does not have an explicit solution for $(\hat{a},\ \hat{b})$. Moreover, since $R_\alpha(y)$ is not differentiable at $y = 0$, either a smooth approximation of $R_\alpha(.)$ or a more complicated software development seems necessary in order to compute the estimates numerically (cf. Bloomfield and Steiger 1983).

An alternative approach is to change the loss function (22) to a quadratic function

$$Q_\omega(y) = \begin{cases} (1-\omega)y^2 & y \le 0, \\ \omega y^2 & y > 0, \end{cases}$$

for $\omega \in [0,1]$, the 100 ω-th conditional expectile of Y_m is defined as

$$\tau_{\omega,m}(x) \;=\; \arg\min_{|a|<\infty}\ E\{\, Q_\omega(Y_m - a) \mid X_0 = x \,\},$$

(cf. Neway and Powell 1987). Obviously, in the case $\omega = \frac{1}{2}$, this definition reduces to the conditional mean $E(Y|X = x)$. Since $Q_\omega(.)$ has a continuous first derivative, $\tau_{\omega,m}(x)$ satisfies the equation

$$E\{\, L_\omega(Y_m - \tau_{\omega,m}(x)) \mid X_0 = x \,\} = 0,$$

where

$$L_\omega(x) = \begin{cases} (1-\omega)y & y \le 0, \\ \omega y & y > 0. \end{cases} \tag{24}$$

Consequently, we have

$$\omega = \frac{E\{\ |Y_m - \tau_{\omega,m}(x)|;\ \ Y_m \leq \tau_{\omega,m}(x) \mid X_0 = x\ \}}{E\{\ |Y_m - \tau_{\omega,m}(x)| \mid X_0 = x\ \}}, \tag{25}$$

where $E\{X;\ A\}$ denotes $E\{XI_A\}$, and I_A is the indicator function of the set A. Notice that for $\xi_{\alpha,m}(.)$, we can rewrite the relation

$$\alpha = \frac{E\{\ 1;\ \ Y_m \leq \xi_{\alpha,m}(x) \mid X_0 = x\ \}}{E\{\ 1\ \mid X_0 = x\ \}}. \tag{26}$$

Comparing the above two expressions, we can see that given $X_0 = x$, the percentile $\xi_{\alpha,m}(x)$ specifies the position below which $100\alpha\%$ of the (probability) mass of Y_m lies; while the expectile $\tau_{\omega,m}(x)$ determines, again given $X_0 = x$, the point such that $100\omega\%$ of the mean absolute distance between it and Y_m comes from the mass below it. Based on this interpretation, $\tau_{\omega,m}(x)$ can also be used to construct a predictive interval: given $\{Y_t,\ t \leq n\}$, predict Y_m to lie in the interval $[\tau_{\omega/2,m}(X_n),\ \tau_{1-\omega/2,m}(X_n)]$ with $100(1-\omega)\%$ 'coverage'.

To estimate $\tau_{\omega,m}(.)$, we minimize the function

$$\sum_{t=1}^{n-m} Q_\omega\{Y_{t+m} - a - b^T(X_t - x)\}\, K\left(\frac{X_t - x}{h}\right),$$

and define the estimators $\hat{\tau}_{\omega,m}(x) = \hat{a}$, $\hat{\dot{\tau}}_{\omega,m}(x) = \hat{b}$. It is easy to see that $\{\hat{\tau}_{\omega,m}(x),\ \hat{\dot{\tau}}_{\omega,m}(x)\}$ satisfies the following equation

$$\begin{cases} \sum_{t=1}^{n-m} L_\omega\{Y_{t+m} - \hat{\tau}_{\omega,m}(x) - (X_t - x)^T\hat{\dot{\tau}}_{\omega,m}(x)\}K\left(\frac{X_t - x}{h}\right) = 0, \\ \sum_{t=1}^{n-m}(X_t - x)L_\omega\{Y_{t+m} - \hat{\tau}_{\omega,m}(x) - (X_t - x)^T\hat{\dot{\tau}}_{\omega,m}(x)\}K\left(\frac{X_t - x}{h}\right) = 0. \end{cases}$$

Here, $L_\omega(.)$ is the piecewise linear function defined by (24). Based on this relation, it is easy to construct a fast iterative algorithm to computing $\{\hat{\tau}_{\omega,m}(x),\ \hat{\dot{\tau}}_{\omega,m}(x)\}$ (cf. Yao and Tong 1992). Although a predictive interval based on conditional expectiles is convenient to compute, it does not have the conventional probability interpretation in general. However, if we think that the conditional expectation is a good point predictor, $[\tau_{\omega/2,m}(X_n),\ \tau_{1-\omega/2,m}(X_n)]$ could be considered a reasonable interval predictor extended from the conditional expectation. Yao and Tong (1992) has pointed out that, in a special case, the above asymmetric least squares approach can be used to estimate conditional percentiles directly.

Theorem 2. Assume that conditions (A3) – (A8) listed in Appendix hold for some $m \geq 1$.

(i) For $x \in \{p(x) > 0\}$,

$$\sqrt{nh^d}\{\hat{\xi}_{\alpha,m}(x) - \xi_{\alpha,m}(x) - h^2\mu_1\} \xrightarrow{d} N(0, \sigma_1^2),$$

$$\sqrt{nh^{d+2}}\{\hat{\dot{\xi}}_{\alpha,m}(x) - \dot{\xi}_{\alpha,m}(x) - h\mu_2\} \xrightarrow{d} N(0, \Sigma_2),$$

where

$$\mu_1 = \frac{1}{2}\sigma_0^2 \mathrm{tr}\{\ddot{\xi}_{\alpha,m}(x)\} + o(1), \qquad \mu_2 = \frac{1}{2\sigma_0^2}\int uu^T \ddot{\xi}_{\alpha,m}(x) u K(u)\, du + o(1),$$

$$\sigma_1^2 = \frac{\alpha(1-\alpha)\int K^2(u)du}{p(x)[g_m(\xi_{\alpha,m}(x)|x)]^2}, \qquad \Sigma_2 = \frac{\alpha(1-\alpha)\int uu^T K^2(u)du}{p(x)\sigma_0^2[g_m(\xi_{\alpha,m}(x)|x)]^2}.$$

(ii) For $x \in \{p(x) > 0\}$,

$$\sqrt{nh^d}\{\hat{\tau}_{\omega,m}(x) - \tau_{\omega,m}(x) - h^2\mu_3\} \xrightarrow{d} N(0, \sigma_3^2),$$

$$\sqrt{nh^{d+2}}\{\hat{\dot{\tau}}_{\omega,m}(x) - \dot{\tau}_{\omega,m}(x) - h\mu_4\} \xrightarrow{d} N(0, \Sigma_4),$$

where

$$\mu_3 = \frac{1}{2}\sigma_0^2 \mathrm{tr}\{\ddot{\tau}_{\omega,m}(x)\} + o(1), \qquad \mu_4 = \frac{1}{2\sigma_0^2}\int uu^T \ddot{\tau}_{\omega,m}(x) u K(u)\, du + o(1),$$

$$\sigma_3^2 = \frac{\int K^2(u)du \mathrm{Var}\{\dot{Q}_\omega(Y_m - \tau_{\omega,m}(x))|X_0 = x\}}{p(x)\gamma^2},$$

$$\Sigma_4 = \frac{\int uu^T K^2(u)du \mathrm{Var}\{\dot{Q}_\omega(Y_m - \tau_{\omega,m}(x))|X_0 = x\}}{p(x)\sigma_0^2\gamma^2},$$

and $\gamma = 2\omega P\{Y_m \le \tau_{\omega,m}(x)|X_0 = x\} + 2(1-\omega)P\{Y_m > \tau_{\omega,m}(x)|X_0 = x\}$.

Using the convexity lemma (cf. Pollard 1991), Yao and Tong (1992) proved Theorem 2 (ii) in the special case $d = 1$. The multidimensional case is technically more involved, but contains no fundamentally new ideas for the current version. Theorem 2 (i) can be proved in a similar way (also see Fan, Hu and Truong 1992).

Theorem 2 gives the asymptotic normality of the the estimators for the conditional percentiles, expectiles and their derivatives. Notice that $\hat{\tau}_{1/2,m}(x) = \hat{f}_m(x)$ and $\hat{\dot{\tau}}_{1/2,m}(x) = \hat{\lambda}_m(x)$. Therefore, Theorem 2 (ii) also includes the asymptotic normality of the point estimators as a special case. As shown in the theorem, the 'asymptotic bias' is of the order of h^2 for the estimators $\hat{\xi}_{\alpha,m}$ and $\hat{\tau}_{\omega,m}$, and order of h for the estimators of their derivatives; they come from the error in the local approximation of the underlying curve by a linear function. A locally quadratic fit will improve the estimation for the derivatives (cf. Fan et. al. 1993). However, it creates further complications in practical implementation.

We use the following two kinds of intervals to predict Y_{n+m} from $\{Y_t,\, t \le n\}$

$$[\hat{\xi}_{\alpha/2,m}(X_n), \quad \hat{\xi}_{1-\alpha/2,m}(X_n)], \tag{27}$$

$$[\hat{\tau}_{\omega/2,m}(X_n), \quad \hat{\tau}_{1-\omega/2,m}(X_n)]. \tag{28}$$

The interval (27) is based on the conditional percentiles, which has the conventional probability interpretation. The interval (28) does not have any probability meaning. Relations (25) and (26) show that the interval (28) is constructed in the same way as (27) except that we use the mean distance instead of the probability mass.

From Theorem 1, we have learned that in nonlinear prediction, we should take account of two kinds of error: the error caused by the stochastic noise, which varies over the state space, and the error caused by a small shift in the initial value. In the context of interval prediction, the first kind of error can be tentatively represented by the width of the interval. To monitor the second kind of error, i.e. the sensitivity of the predictive intervals, we can use the estimates of the derivatives of the conditional percentiles or expectiles. The estimates presented in the next subsection also offer measures for the sensitivity.

3.4. Estimates of $I_m(x)$

To estimate the Fisher information $I_m(x)$ as given in (5), we use the method proposed by Fan, Yao and Tong (1993). For simplicity, let us discuss the first order case, i.e. $d = 1$, noting that the generalisation to the higher order cases is straightforward. Notice that

$$I_m(x) = 4 \int \left\{ \frac{d\sqrt{g_m(y|x)}}{dx} \right\}^2 dy.$$

We first construct the estimators for $\sqrt{g_m(y|x)}$ and its derivative first. Let $q_m(x, y)$ denote $\sqrt{g_m(y|x)}$.

For given bandwidths h_1 and h_2, let

$$C_m(X_i, Y_i) = \#\{(X_t, Y_t), 1 \le t \le n \ : \ ||X_t - X_i|| \le h_1 \text{ and } |Y_{t+m} - Y_{i+m}| \le h_2\},$$

$$C_m(X_i) = \#\{X_t, 1 \le t \le n, \ : \ ||X_t - X_i|| \le h_1\},$$

for $1 \le i \le n$. Then

$$Z_t \equiv \sqrt{C_m(X_t, Y_t)/\{C_m(X_t)\, h_2\}}$$

is a natural estimate of $q_m(x, y)$ at $(x, y) = (X_t, Y_t)$. Fitting it into the context of locally linear regression, we estimate $q_m(x, y)$ and its derivative with respect to x, denoted by $\dot{q}(x, y)$, by using $\hat{q}_m(x, y) = \hat{a}$ and $\hat{\dot{q}}_m(x, y) = \hat{b}$, where $(\hat{a}, \hat{b})$ are the minimizer of the function

$$\sum_{t=1}^{n-m} \{Z_t - a - b^T(X_t - x)\}^2 K\left(\frac{X_t - x}{h_1}, \frac{Y_t - y}{h_2}\right),$$

K being a probability density function on R^{d+1}. Consequently, we estimate $I_m(x)$ by

$$\hat{I}_m(x) = 4 \int \{\hat{\dot{q}}_m(x, y)\}^2 \, dy.$$

For further discussion of this estimation, we refer to Fan, Yao and Tong (1993).

4. Examples

We have shown, via asymptotics, that the performance of nonlinear prediction is influenced by the initial values. In this section, we use two simulated models and two real data sets to illustrate the finite-sample behaviour. We use Gaussian kernel in our estimation.

4.1. Logistic Map

We begin with the simple one-dimensional model (10). In fact, its skeleton is a transformed logistic map with the coefficient $3.68(=16 \times 0.23)$. We adopt the transformation in order to enlarge the dynamic range of the model. A sample of 1200 is generated from model (10). Note that $\sigma_1^2(x) \equiv 0.16$ (also cf. Fig. 1(a)); 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. 2, 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 5.6 and 10.4 (see Fig. 2(b)). We use the first 1000 observations to estimate the unknown functions. 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 estimated conditional variance $\hat{\sigma}_m^2(x)$ (cf. (21)) are plotted in Fig. 3 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. 3(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 5.6 or 10.4, which is in agreement with the observation from Fig. 2(b). Similar remarks apply to the two-step and the four-step predictions.

To see how a small shift in the initial values affects the prediction, we round the 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. 4 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 (17).

Fig. 5 presents the predictive intervals constructed by the estimated conditional percentiles $\hat{\xi}_{\alpha,m}(.)$, together with 200 real values. To estimate the conditional percentiles, we use the multidimensional *downhill simplex method* (cf. §10.5 of Press et. al. 1992). Fig. 5 shows that the width of the interval varies with respect to the initial value. For example, in the case of $m = 3$, the width attains its maximum around $x = 8$, and its minimum about $x = 5.6$ and 10.4 (cf. Fig. 5(b)). Notice that the presented intervals are supposed to contain the predicted values with probability 0.9;

(a)

(b)

(c)

Figure 2: The scatter plots of Y_{t+m} against Y_t for Logistic map: (a) $m = 2$; (b) $m = 3$; (c) $m = 4$

(a)

(b)

(c)

Figure 3: The plots of the 200 m-step predicted values of Logistic map, and the corresponding absolute prediction errors against their initial values, as well as the estimated conditional variance $\hat{\sigma}_m^2(x)$: (a) $m = 2$ ($h = 0.25$); (b) $m = 3$ ($h = 0.2$); (c) $m = 4$ ($h = 0.18$). Diamonds — predicted values; impulses — absolute prediction errors; solid curve — $\hat{\sigma}_m^2(x)$.

(a)

(b)

Figure 4: The plots of the 200 m-step predicted values of Logistic map, and the corresponding absolute prediction errors against their rounded initial values, and the estimated function $|\hat{\lambda}_m(x)|$; (a) $m = 1$ ($h = 0.32$); (b) $m = 2$. Diamonds — absolute prediction errors; solid curve — $|\hat{\lambda}_m(x)|$.

(a)

(b)

(c)

Figure 5: The predictive interval $[\hat{\xi}_{0.05,m}(x),\ \hat{\xi}_{0.95,m}(x)]$, and 200 real values for Logistic map. (a) $m = 2$ ($h = 0.5$); (b) $m = 3$ ($h = 0.42$); (c) $m = 4$ ($h = 0.37$). Solid curve — $\hat{\xi}_{0.95,m}(x)$; dotted curve — $\hat{\xi}_{0.05,m}(x)$; diamonds — real values.

(a)

(b)

(c)

Figure 6: The predictive interval $[\hat{\tau}_{0.05,m}(x),\ \hat{\tau}_{0.95,m}(x)]$, and 200 real values for Logistic map. (a) $m = 2$ ($h = 0.25$); (b) $m = 3$ ($h = 0.2$); (c) $m = 4$ ($h = 0.18$). Solid curve — $\hat{\tau}_{0.95,m}(x)$; dotted curve — $\hat{\tau}_{0.05,m}(x)$; diamonds — real values.

(a)

(b)

Figure 7: The estimated Fisher information $\hat{I}_m(x)$, and the derivatives of conditional percentiles and expectiles for Logistic map. (a) $m = 1$ ($h_1 = 0.61,\ h_2 = 0.24$ for $\hat{I}_1(x)$); (b) $m = 2$ ($h_1 = 0.57,\ h_2 = 0.22$ for $\hat{I}_2(x)$). Solid curve — $\hat{I}_m(x)$; dashed curve — $\{(\dot{\hat{\xi}}_{0.05,m}(x))^2 + (\dot{\hat{\xi}}_{0.95,m}(x))^2\}^{1/2}$; dotted curve — $\{(\dot{\hat{\tau}}_{0.05,m}(x))^2 + (\dot{\hat{\tau}}_{0.95,m}(x))^2\}^{1/2}$.

we notice that about 10% of the 200 samples lie outside the intervals. The predictive intervals constructed by the estimated conditional expectile $\hat{\tau}_{\omega,m}(.)$ are displayed in Fig. 6.

To monitor the sensitivity of the predictive interval to the initial value, we plot the three sensitive measures in Fig. 7. It shows that the profiles of the Fisher information $I_m(x)$, $\{(\hat{\dot{\xi}}_{0.05,m}(x))^2 + (\hat{\dot{\xi}}_{0.95,m}(x))^2\}^{1/2}$ and $\{(\hat{\dot{\tau}}_{0.05,m}(x))^2 + (\hat{\dot{\tau}}_{0.95,m}(x))^2\}^{1/2}$ are generally quite similar. Comparing Fig. 7(b) with Fig. 5(a), or Fig. 6(a), we can see that where the sensitive measures are large, small shift will lead to a relatively large change in the intervals (e.g. x near 5 and 11), and vice versa. (Also see Fig. 4.)

4.2. Hénon Map

We clothe a Hénon map with dynamic noise to obtain

$$Y_t = 6.8 - 0.19Y_{t-1}^2 + 0.28Y_{t-2} + 0.2\epsilon_t \quad t \geq 1, \tag{29}$$

where ϵ_t, $t \geq 1$, are as described in equation (10). A sample of 1200 observations is generated from this model. 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_{t-1}, Y_{t-2}). Fig. 8 reports the predicted values together with the corresponding true values. The estimated values of the conditional variance at these points are shown in Fig. 9, which indicate the accuracy of the prediction. (Note the occasional negative estimates as discussed in Section 3.1.) 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. 8 (a) with Fig. 9 (a)). It can also be seen in Fig. 8 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. 10, we have rounded the first half of the checking sample in the same way as in Fig. 4. (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 $\parallel \hat{\lambda}_m \parallel$ instead of $\hat{\lambda}_m$.

Table 1 reports the two-step and three-step ahead predictive intervals for the first 10 checking sample. Here we use the interval predictor $[\hat{\xi}_{0.05,m},\ \hat{\xi}_{0.95,m}]$, which contains the predicted variable with probability 0.90 theoretically. The bandwidth is chosen as 0.53 for two-step prediction and 0.48 for three-step prediction. It has so happened that in both cases of $m = 2$, and 3, 1 out of 10 true values lies outside the estimated interval. The results also show that the width of the interval varies considerably with respect to the initial values. For example, for the case $m = 2$, the shortest interval in the table is [4.27, 4.76] with the width 0.49, and the largest interval is [-7.76, -5.67] with the width 2.09. We also report the values of $\{||\hat{\dot{\xi}}_{0.05,m}||^2 + ||\hat{\dot{\xi}}_{0.95,m}||^2\}^{1/2}$ as our sensitive measures, although we could not see their effects in the reported results because the rounding errors in the calculation are below 10^{-6}. We can also

(a)

(b)

Figure 8: The plots of the 200 m-step predicted values for Hénon Map, 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.

(a)

(b)

Figure 9: The plots of the 200 estimates values of $\hat{\sigma}_m^2$ against the first component of their initial values for Hénon Map: (a) $m = 2$; (b) $m = 3$.

(a)

(b)

Figure 10: 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 for Hénon Map: (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: Interval prediction for Henon map

Initial values		Two-step ahead prediction			Three-step ahead prediction		
Y_t	Y_{t-1}	Y_{t+2}	PI	SM	Y_{t+3}	PI	SM
-6.49	8.12	4.99	[3.80, 5.00]	1.00	2.24	[2.36, 4.27]	2.56
1.19	-6.49	2.24	[2.27, 3.51]	2.01	6.84	[5.63, 7.16]	3.00
4.99	1.19	6.84	[6.45, 7.26]	2.44	-1.42	[-2.57, -0.42]	6.80
2.24	4.99	-1.42	[-3.39, -1.44]	4.09	8.27	[6.27, 8.28]	2.55
6.84	2.24	8.27	[7.06, 8.38]	1.68	-7.12	[-7.08, -3.96]	4.82
-1.42	6.84	-7.12	[-7.76, -5.67]	2.57	-1.03	[-2.65, 2.94]	6.07
8.27	-1.42	-1.03	[-0.15, 1.52]	5.44	4.36	[2.29, 4.96]	2.44
-7.12	8.27	4.36	[4.27, 4.76]	2.04	2.74	[1.96, 3.17]	1.94
-1.03	-7.12	2.74	[1.48, 3.28]	3.90	6.35	[5.98, 8.03]	4.31
4.36	-1.03	6.35	[5.94, 6.81]	2.54	0.19	[-1.44, 0.83]	6.15

PI: predictive interval; SM: sensitive measure.

use conditional expectiles to construct predictive intervals. The results are similar and are therefore omitted here.

4.3. Lynx Data

We present the results of pointwise prediction for $m = 1$ and 2 for the Canadian lynx data for 1821-1934 (listed in Tong 1990) in Table 2. 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.

We also perform the interval prediction using conditional percentiles to this data set. The results are reported in Table 3. The bandwidth is chosen as 0.57 for one-step prediction and 0.51 for two-step prediction. We use the predictor $[\hat{\xi}_{0.05,m},\ \hat{\xi}_{0.95,m}]$. In the case $m = 1$, two predictive intervals (out of the ten) do not cover the true values. In the case $m = 2$, although all the intervals contain the corresponding true values, the widths of the intervals are considerably larger than those in the case $m = 1$ (except for the year 1925).

Table 2: Point 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 3: Interval prediction of the Canadian lynx data (on natural log scale)

Year	True value	Predictive interval	
		$m = 1$	$m = 2$
1925	8.18	[7.88, 8.67]	[7.84, 8.36]
1926	7.98	[7.35, 8.27]	[6.89, 8.47]
1927	7.34	[6.48, 7.88]	[5.92, 7.58]
1928	6.27	[5.68, 8.09]	[4.77, 8.47]
1929	6.18	[4.97, 6.35]	[4.76, 7.29]
1930	6.50	[5.75, 6.43]	[5.31, 6.53]
1931	6.91	[5.99, 6.97]	[6.28, 7.41]
1932	7.37	[7.04, 7.63]	[6.65, 7.87]
1933	7.88	[7.07, 7.83]	[7.31, 8.07]
1934	8.13	[7.55, 8.40]	[7.22, 8.32]

Table 4: Point prediction of the sunspot numbers

Year	True value	error ($\hat{f}_1$)	$\| \hat{\lambda}_1 \|$	error(MARS)
1979	155.4	- 8.88	2.64	-21.67
1980	154.7	-47.26	6.32	-9.24
1981	140.5	- 5.83	2.98	-10.55
1982	115.9	-32.0	12.59	-12.58
1983	66.6	2.80	1.10	15.76
1984	45.9	1.01	0.96	-2.59
1985	17.9	17.94	1.16	2.75
1986	13.4	-2.57	0.64	-7.66
1987	29.2	-19.73	0.92	-2.82
1988	100.2	-53.67	3.92	-24.27
1989	157.6	35.56	8.27	-9.32
1990	142.6	34.51	9.84	4.38
1991	145.7	-11.63	3.08	-27.90
1992	94.3	-10.40	11.35	12.98

4.4. Sunspot Data

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 1979-1992 (i.e. 14 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 4. 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 $\| \hat{\lambda}_1 \|$.

We also fit this data set by using the MARS method with linear base functions (cf. Friedman 1991, Lewis and Stevens 1991). Using the sunspot data in the period 1700-1978, the selected model is

$$\begin{aligned} Y_t \;=\; & 181.67 - 0.23Y_{t-3} + 0.86(Y_{t-1} - 190.2)^- - 1.93(Y_{t-2} - 79.7)^- \\ & -0.0023(Y_{t-2} - 79.7)^-(Y_{t-9} - 190.2)^- - 0.0081(Y_{t-1} - 190.2)^-(Y_{t-2} - 92.6)^- \\ & -0.0043Y_{t-1}Y_{t-3}(Y_{t-2} - 21.3)^-, \end{aligned}$$

where $x^- = x$ if $x < 0$, and 0 otherwise. The prediction errors of this model for the sunspot data in the period 1979-1992 are reported in the last column of Table 4. Again, we have better prediction in the 'trough-years' than the 'peak-years'.

5. Appendix: The Regularity Conditions

To discuss the asymptotic properties of the estimators, we need the following

assumptions.

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

(A2) The conditional variance $\sigma_m^2(x) = \mathrm{Var}(Y_m|X_0 = x)$ is bounded and continuous.

(A3) The joint density of distinct elements of (X_1, Y_1, X_k, Y_k) is bound by a constant independent of k.

(A4) X_t has the probability density function p, and $|p(x) - p(y)| \leq C \parallel x - y \parallel$ for any $x, y \in R^d$.

(A5) The precess $\{Y_t\}$ is ρ-mixing, i.e., $\rho_j = \sup_{U \in \Im_{-\infty}^0, V \in \Im_j^\infty} \mathrm{Corr}(U, V) \to 0$, as $j \to \infty$, where $\Im_i^j$ is the σ-field generated by $\{Y_k,\ k = i, \ldots, j\}$. Further, we assume that $\sum_{k=1}^{\infty} \rho_k < \infty$.

(A6) $K(\cdot)$ is a continuous density function with a bounded support in R^d, and $\int xK(x)dx = 0$, $\int xx^T K(x)dx = \sigma_0^2 I_d$, where I_d denotes the $d \times d$ identity matrix.

(A7) The bandwidth $h \to 0$, $nh^{2+d} \to \infty$, and $(\log n)/(nh^d) \to 0$.

(A8) For any compact subset $B \in R^d$, there exists a constant c such that for any x, $y \in B$, $|\int z^2 g_m(z|x)dz - \int z^2 g_m(z|y)dz| \leq ||x - y||$, where $g_m(y|x)$ denotes the conditional density of Y_m given X_0.

6. References

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

Bloomfield, P. and Steiger, W.L. (1983). *Least Absolute Deviations.* Birkhäuser, Boston.

Chan, K.S. and Tong, H. (1994). A note on noisy chaos. *J. R. Statis. Soc.* **B**, **56**, 301-311.

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**, 45-82.

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., Hu, T.C. and Truong, Y.K. (1992). Robust nonparametric function estimation. Technical Report 035-92, Math.Science Research Inst., Berkeley.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. Technical Report, University of North-Carolina.

Fan, J., Yao, Q., and Tong, H. (1993). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Technical Report, Univ. of Kent.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, **19**, 1-50.

Hall, P. and Wolff, R.C.L. (1993). Properties of invariant distributions and Lyapunov exponents for chaotic logistic maps. Technical Report, Australian National University.

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

Kullback, S. (1967). *Information Theory and Statistics.* Dover Publi., New York.

Lewis, P.A.W. and Stevens, J.G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines. *J. Ameri. Statis. Assoc.*, **86**, 864-877.

Neway, W.K. and Powell, J.K. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-847.

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

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186-198.

Press, W.H., Flannery, B.P., Tenkolsky, S.A., and Vetterling, W.T. (1992). *Numerical Recipes.* Cambridge Univ.Press, Cambridge.

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

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

Yao, Q. and Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *J. R. Statis. Soc.* **B**, **56**, 701-725.