PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

On Prediction and Chaos in Stochastic Systems

Qiwei Yao and Howell Tong

Phil. Trans. R. Soc. Lond. A 1994 **348**, 357-369
doi: 10.1098/rsta.1994.0096

On prediction and chaos in stochastic systems†

By Qiwei Yao and Howell Tong

*Institute of Mathematics and Statistics, University of Kent,
Canterbury, Kent CT2 7NF, U.K.*

We propose a new measure of sensitivity to initial conditions within a stochastic environment and explore its connection with nonlinear prediction and statistical estimation. We use modern statistical developments to construct and illustrate pointwise predictors and predictive intervals/distributions.

1. Introduction

A purely deterministic system rarely exists in reality because stochastic noise is ubiquitous. Accordingly, it is more pertinent to replace the dynamics by the transition probabilities from states to states. A convenient framework for this stochastic system is the Markov chain over a general state space and nonlinear autoregressive models emerge naturally as a realization of this framework for the study of noisy chaos (cf. Chan & Tong 1994). Within this considerably enlarged stochastic framework, a new notion of sensitivity to initial conditions has to be developed.

Point predictions are only the first step in any serious study of the subject. To complete the picture, we need 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. The nonlinear prediction has three distinctive 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 & Tong 1994).

2. Stochastic dynamic system

(*a*) *Noisy chaos*

Just as in a deterministic system, there has been no generally accepted definition of chaos in a stochastic system, although the term noisy chaos has appeared in the literature. By a stochastic chaotic system is sometimes meant a system with a (deterministically) chaotic skeleton. However, this approach is not always appropriate because the dynamic noise will, by permeating through the system dynamics, interact with the system signal throughout the time evolution. An ex-

† This paper was produced from the authors' disk by using the TEX typesetting system.

Phil. Trans. R. Soc. Lond. A (1994) **348**, 357–369
Printed in Great Britain

treme 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 described by the equation

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

for $t \geqslant 1$, 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, (2.1) can be written (by an abuse of notation) as follows

$$X_t = F(X_{t-1}) + e_t. \tag{2.2}$$

It is widely accepted that the sensitive dependence on initial conditions is a typical feature of a deterministic chaotic system, and this can be characteristically described in terms of the well-known Lyapunov exponents (see Eckmann & Ruelle 1985). 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. 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 with substantial noise still has a strong memory of its initial conditions. This suggests that asymptotics are unlikely to yield a practically useful characteristic exponent.

One way to manifest the sensitivity of the conditional distribution is to use the Kullback–Leibler–type 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 $X_0 = x$, which is assumed smooth enough in x. For nearby initial points x, $x + \delta \in R^d$, after time $m \geqslant 1$, the divergence of the conditional distributions of X_m is defined as

$$K_m(x;\delta) = \int \{g_m(y|x+\delta) - g_m(y|x)\} \log\{g_m(y|x+\delta)/g_m(y|x)\}\, \mathrm{d}y. \tag{2.3}$$

For small δ, $K_m(x;\delta)$ has the approximation

$$K_m(x;\delta) = \delta^T I_m(x)\delta + o(||\delta||^2), \tag{2.4}$$

where

$$I_m(x) = \int \dot{g}_m(y|x)\dot{g}_m^T(y|x)/g_m(y|x)\, \mathrm{d}y, \tag{2.5}$$

$\dot{g}_m(y|x)$ denotes $\mathrm{d}g_m(y|x)/\mathrm{d}x$, and $\dot{g}_m^T(y|x)$ 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 information on x contained in X_m. Roughly speaking, the more information X_m brings, the more sensitively the distribution depends on the initial condition. Fan *et al.* (1993) has given another measure of sensitivity in the form of an L_2 norm.

It is also interesting to look at the divergence in some summarizing characteristics, for example the (conditional) means. Among other things, Yao & Tong (1994) have considered the following case. Let Y_t denote the first component of

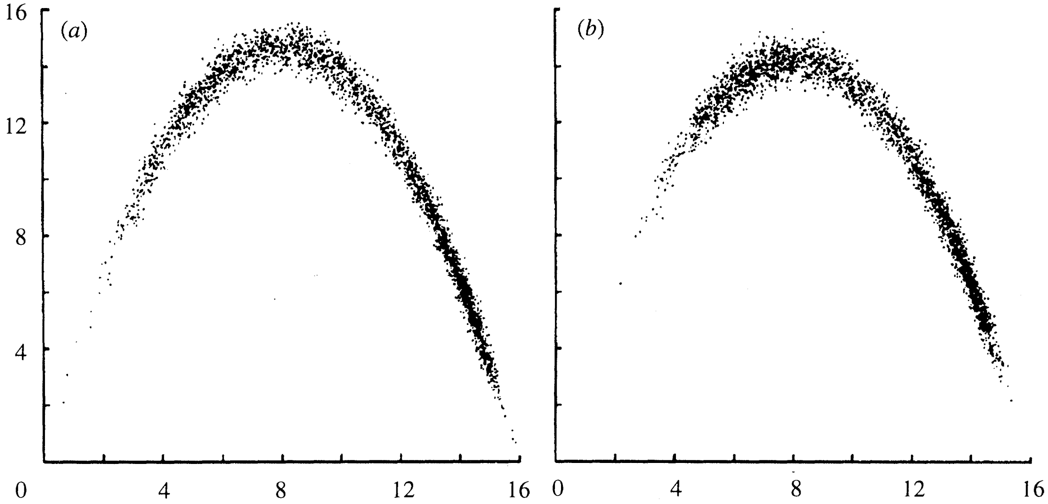

Figure 1. The scatter plots of Y_{t+1} against Y_t of model (*a*) $Y_t = 0.230Y_{t-1}(16 - Y_{t-1}) + 0.4\epsilon_t$; (*b*) $Y_t = 0.222Y_{t-1}(16 - Y_{t-1}) + 0.4\epsilon_t$, where $\{\epsilon_t\}$ is a sequence of IID standard normal random variables. Note that the skeleton of model (*a*) is chaotic while that of model (*b*) is a limit cycle with period 8.

X_t. It follows from (2.2) 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 $m \geqslant 1$, and $x \in R^d$, let $f_m(x) = E(Y_m|X_0 = x)$. Obviously, $f_1(x) = f(x)$. Then we have

$$f_m(x+\delta) - f_m(x) = \delta^{\mathrm{T}}\lambda_m(x) + \mathrm{o}(\| \delta \|), \tag{2.6}$$

where $\lambda_m(x) = \mathrm{d}f_m(x)/\mathrm{d}x$. We call $\lambda_m(.)$ the m-step *Lyapunov-like index*, or simply the m-LI. When $d = 1$,

$$\lambda_m(x) = E\left\{\prod_{k=1}^{m} \frac{\mathrm{d}}{\mathrm{d}x} f(X_{k-1}) \mid X_0 = x\right\} = E\left\{\prod_{k=1}^{m} \lambda_1(X_{k-1}) \mid X_0 = x\right\}. \tag{2.7}$$

We will see in § 3 its role in the pointwise prediction.

We remark that the 'clear' cut-off between deterministic chaotic systems and deterministic non-chaotic systems is masked by the presence of stochastic noise. Figure 1 illustrates the situation.

(*b*) *Noise amplification*

We measure the amplification of noise by comparing the conditional variance of X_t (given the initial conditions X_0) with the variance of e_t. Deissler & Farmer (1991) studied the noise amplification in a different way. They considered the distance between the state variables in a (known) purely deterministic system and their counterparts in the system perturbed by additive system noise. This approach seems inappropriate in the statistical context, because the underlying deterministic skeleton is now typically unknown.

As an illustration, let us consider a one-dimensional system with additive noise:

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

where $\{\epsilon_t, t \geqslant 1\}$ is a noise process with mean value 0 and variance σ_0^2 and $Y_0 = x \in R$. 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 may be proved that as $\sigma_0 \to 0$,

$$\sigma_m^2(x) = \sigma_0^2 \, \mu_m(x) \, (1 + o(1)), \tag{2.8}$$

where

$$\mu_m(x) = 1 + \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \dot{f}[f^{(k)}(x)] \right\}^2 \tag{2.9}$$

(Yao & Tong 1994). Thus, if $|\dot{f}(x)| > 1$ 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, (2.12) 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 (2.11), 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 & Tong 1994).

3. Nonlinear prediction

Suppose that $\{Y_t, -\infty < t < \infty\}$ is a one-dimensional strictly stationary time series such that given $\{Y_i, i \leqslant t\}$, the conditional distribution of Y_{t+1} depends on $\{Y_i, i \leqslant t\}$ only through X_t, where $X_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+1})^{\mathrm{T}}$. Given the observations $\{Y_t, -d+1 < t \leqslant n\}$, 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. For, let $f(x) = E(Y_1 | X_0 = x)$. Then Y_t can be expressed as

$$Y_t = f(X_{t-1}) + \epsilon_t, \tag{3.1}$$

where $\epsilon_t = Y_t - f(X_{t-1})$. Define $F(X_{t-1}) = (f(X_{t-1}), Y_{t-1}, \ldots, Y_{t-d+1})^{\mathrm{T}}, e_t = (\epsilon_t, 0, \ldots, 0)^{\mathrm{T}}$. Then equation (2.2) holds. Henceforth, the time series model is said to be chaotic if the corresponding stochastic dynamic system is chaotic.

(a) *Point predictors*

To study the m-step prediction, we define $f_m(x) = E(Y_m | X_0 = x)$, for $x \in R^d$ and $m \geqslant 1$. It is easy to see that the (theoretical) least squares predictor of Y_{n+m} based on $\{Y_t, t \leqslant n\}$ is $f_m(X_n)$. In practice, we use $\hat{f}_m(X_n)$ as the predictor, where $\hat{f}_m(.)$ is a *reasonable* estimator for $f_m(.)$. In fact, it can be proved that if $E\{[f_m(x) - \hat{f}_m(x)]^2 | X_n\} \to 0$ a.s.,

$$\lim_{n\to\infty} E\{[Y_{n+m} - \hat{f}_m(x)]^2 | X_n = x + \delta\} = \sigma_m^2(x+\delta) + \{\delta^{\mathrm{T}} \lambda_m(x)\}^2 + R_m, \quad \text{a.s.}, \tag{3.2}$$

where $R_m = o(\| \delta \|^2)$ as $\| \delta \| \to 0$, $\lambda_m(x) = \mathrm{d}f_m(x)/\mathrm{d}x^{\mathrm{T}}$ is the m-LI, and $\sigma_m^2(x) = \mathrm{Var}(Y_m|X_0 = x)$ (Yao & Tong 1994).

It can be seen from (3.2) 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: (a) the conditional variance; (b) 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, (3.2) 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 depends on x. When the measurement error δ is small but not zero, such as rounding errors in measurement, etc., usually the right hand side of (3.2) 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. (2.6)). In this sense, we say that the m-step prediction is sensitive to the initial values when the model is chaotic.

In (3.1), the noise term ϵ_t is not necessarily homogeneous as indicated in the second expression in (2.7). 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 $\lambda_1(x)$.

Yao & Tong (1994) have discussed the locally linear kernel estimators for $f_m(.)$, $\lambda_m(.)$, and $\sigma_m^2(.)$.

(b) *Interval predictors*

Of course, 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 100αth conditional percentile of Y_m given $X_0 = x \in R^d$ is defined as

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

where the loss function

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

It is well known that the relation $\alpha = P\{Y_m \leqslant \xi_{\alpha,m}(x)|X_0 = x\}$ holds. Therefore, given $\{Y_t, t \leqslant n\}$, 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 $0 \leqslant \alpha \leqslant 1$.

Similar to § 3 b, 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 (with respect to a and b respectively) of the function.

$$\sum_{t=1}^{n-m} R_\alpha\{Y_{t+m} - a - b^T(X_t - x)\}K\left(\frac{X_t - x}{h}\right), \tag{3.3}$$

where $K(.)$ is a probability density function on R^d, and $h = h(n)$ is a bandwidth. Unfortunately, (3.3) 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 more complicated software development seems necessary in order to compute the estimates numerically (cf. Bloomfield & Steiger 1983).

An alternative approach is to change the loss function (3.3) to a quadratic function

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

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

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

(cf. Neway & Powell 1987). Note that this reduces to $E(Y|X = x)$ when $\omega = \frac{1}{2}$.

Now, $\tau_{\omega,m}(x)$ can also be used to construct a predictive interval: given $\{Y_t, t \leqslant n\}$, 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 in the usual way 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 construct a fast iterative algorithm to compute $\{\hat{\tau}_{\omega,m}(x), \hat{\dot{\tau}}_{\omega,m}(x)\}$ (cf. Yao & 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, $[\tau_{\omega/2,m}(X_n), \tau_{1-\omega/2,m}(X_n)]$ could be considered a reasonable interval predictor extended from the conditional expectation. Yao & Tong (1992) have pointed out that, in a special case, the above asymmetric least squares approach can be used to estimate conditional percentiles directly.

Theorem 3.1. *Under conditions (A 1)–(A 6) listed in the appendix,*

(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

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

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

(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

$$\mu_3 = \tfrac{1}{2}\sigma_0^2 \operatorname{tr}\{\ddot{\tau}_{\omega,m}(x)\} + o(1), \quad \mu_4 = \frac{1}{2\sigma_0^2}\int uu^{\mathrm{T}}\ddot{\tau}_{\omega,m}(x)uK(u)\,\mathrm{d}u + o(1),$$

$$\sigma_3^2 = \frac{1}{p(x)\gamma^2} \int K^2(u)\,\mathrm{d}u\, \mathrm{Var}\{\dot{Q}_\omega(Y_m - \tau_{\omega,m}(x)) | X_0 = x\},$$

$$\Sigma_4 = \frac{1}{p(x)\sigma_0^2\gamma^2} \int uu^{\mathrm{T}} K^2(u)\,\mathrm{d}u\, \mathrm{Var}\{\dot{Q}_\omega(Y_m - \tau_{\omega,m}(x)) | X_0 = x\},$$

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

Yao & Tong (1992) have proved (ii) of Theorem 3.1 in the special case $d = 1$. The technically more involved multidimensional case contains no fundamentally new ideas for the current version. Theorem 3.1(i) can be proved in a similar way (also see Fan *et al.* 1992).

Theorem 3.1 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 3.1(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 $\{Y_t, t \leqslant n\}$, noting the remarks before Theorem 3.1,

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

We can monitor the prediction error caused by the stochastic noise by the width of the interval. To monitor the initial-value sensitivity of the predictive intervals by the estimates of the derivatives of the conditional percentiles or expectiles. The estimates presented in the next subsection also offer measures for the sensitivity.

(*c*) *Estimates of* $I_m(x)$

Fan *et al.* (1993) has discussed the estimation of $I_m(x)$. For simplicity, let us discuss the first order case, i.e. $d = 1$, noting that the generalization to the higher order cases is straightforward. Notice that

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

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

For given bandwidths h_1 and h_2, let

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

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

for $1 \leqslant i \leqslant n$. 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 $(x,y) = (X_t, Y_t)$. We estimate $q_m(x,y)$ and

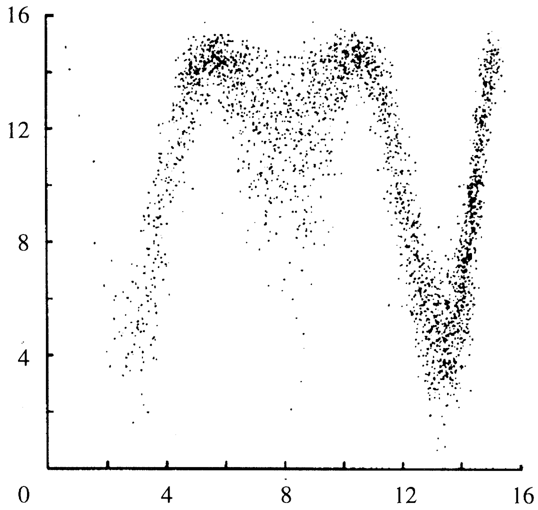

Figure 2. The scatter plots of Y_{t+3} against Y_t.

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 $(\hat{a}, \hat{b})$ is the minimizer of the function

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

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

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

For further details, we refer to Fan *et al.* (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 simulated and real datasets to illustrate the finite-sample behaviour. The estimators used in pointwise prediction are constructed by using locally linear regression method (cf. Yao & Tong 1994). We use gaussian kernel in our estimation. Other examples may be found in Yao & Tong (1994).

(a) *Logistic map*

We begin with the simple one-dimensional model,

$$Y_t = 0.230Y_{t-1}(16 - Y_{t-1}) + 0.4\epsilon_t,$$

where $\{\epsilon_t\}$ is a sequence of independent random variables with the standard normal distribution truncated in the interval $[-12, 12]$. A sample of 1200 is generated. We consider the three-step-ahead prediction only, i.e. $m = 3$, to save space.

The scatter plots of Y_{t+3} against Y_t are displayed in figure 2, which show obvious change of the variability of Y_{t+3} with respect to Y_t. For example, the variability of

Figure 3

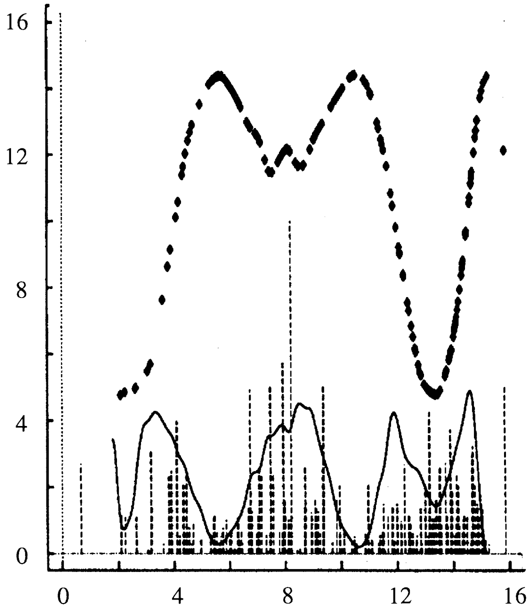

Figure 4

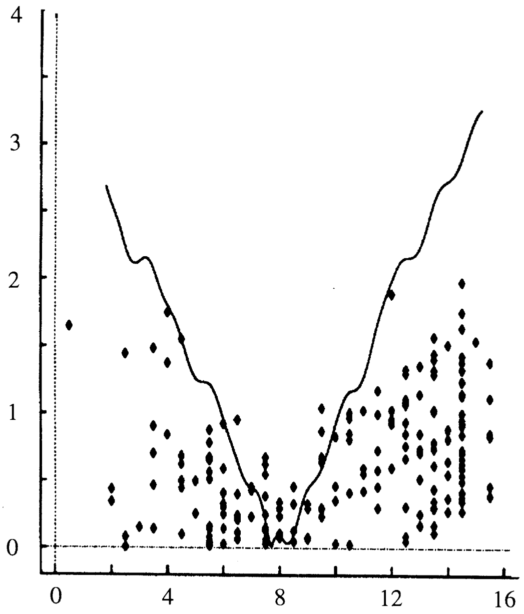

Figure 3. The plots of the 200 three-step predicted values and the corresponding absolute prediction errors against their initial values, as well as the estimated conditional variance $\hat{\sigma}_3^2(x)$ $(h = 0.2)$. Diamonds, predicted values; impulses, absolute prediction errors; solid curve, $\hat{\sigma}_3^2(x)$.

Figure 4. The plots of the 200 three-step predicted values and the corresponding absolute prediction errors against their rounded initial values, and the estimated function $|\hat{\lambda}_3(x)|(h = 0.2)$. Diamonds, absolute prediction errors; solid curve, $|\hat{\lambda}_3(x)|$.

Y_{t+3} is at its largest when $Y_t \approx 8$, and at its smallest when $Y_t \approx 5.6$ and 10.4. We use the first 1000 observations to estimate the unknown functions. The predicted values for the last 200 observations together with their absolute prediction errors and estimated conditional variance $\hat{\sigma}_3^2(x)$ are plotted in figure 3. Since rounding errors in the calculation are below 10^{-6}, the accuracy is dominated by the conditional variance. Figure 3 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 figure 2.

Suppose that we disturb the initial value x by rounding it to the nearest value from among $[x], [x] + 0.5$, and $[x] + 1$, where $[x]$ denotes the integer part of x. Hence, $|\delta| \leqslant 0.5$. Figure 4 shows that for one-step-ahead prediction, the absolute prediction error increases as $|\hat{\lambda}_1(x)|$ increases, which is consistent with the asymptotic conclusion presented in (3.2).

Figure 5 presents 200 real values and the predictive intervals constructed by the estimated conditional percentiles $\hat{\xi}_{\alpha,3}(.)$ obtained by the multidimensional *downhill simplex method* (cf. § 10.5 of Press *et al.* 1992). The width of the interval varies with respect to the initial value. For example, the width attains its maximum around $x = 8$, and its minimum about $x = 5.6$ and 10.4. Notice that the presented intervals contain the predicted values with realtive frequency 0.9 as they are supposed to do. The predictive intervals constructed by the estimated conditional expectile $\hat{\tau}_{\omega,3}(.)$ are displayed in figure 6.

Figure 5

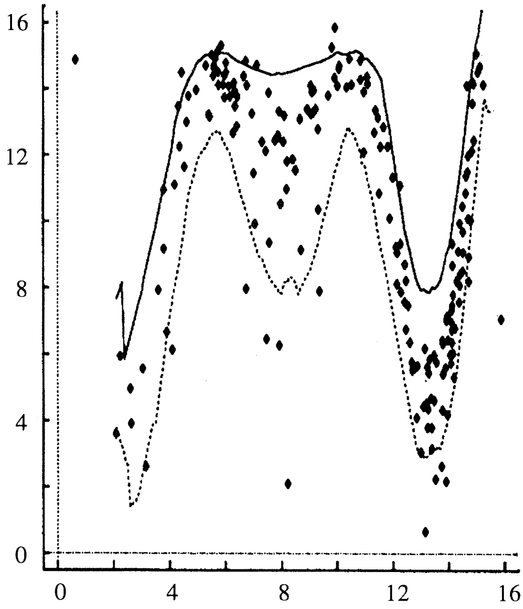

Figure 6

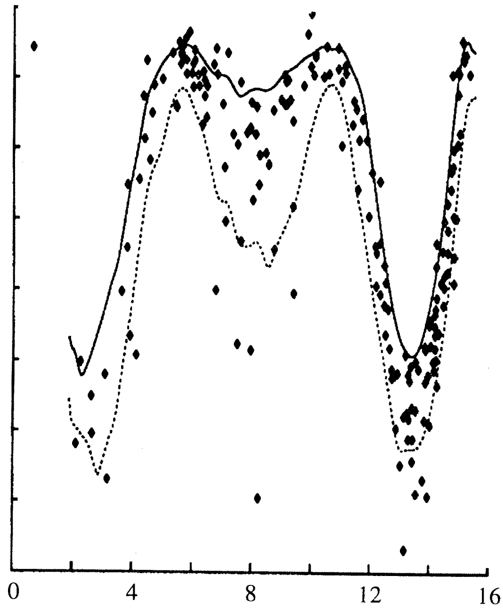

Figure 5. The predictive interval $[\hat{\xi}_{0.05,3}(x), \hat{\xi}_{0.95,3}(x)]$ $(h = 0.42)$, and 200 real values. Solid curve, $\hat{\xi}_{0.95,3}(x)$; dotted curve, $\hat{\xi}_{0.05,3}(x)$; diamonds, real values.

Figure 6. The predictive interval $[\hat{\tau}_{0.05,3}(x), \hat{\tau}_{0.95,3}(x)]$ $(h = 0.2)$, and 200 real values. Solid curve, $\hat{\tau}_{0.95,3}(x)$; dotted curve, $\hat{\tau}_{0.05,3}(x)$; diamonds, real values.

To monitor the sensitivity of the predictive interval to the initial value, we plot the three sensitive measures for $m = 1, 2$ in figure 7. The profiles of the Fisher information,

$$I_m(x), \quad \{(\dot{\hat{\xi}}_{0.05,m}(x))^2 + (\dot{\hat{\xi}}_{0.95,m}(x))^2\}^{1/2}, \quad \{(\dot{\hat{\tau}}_{0.05,m}(x))^2 + (\dot{\hat{\tau}}_{0.95,m}(x))^2\}^{1/2},$$

are generally quite similar.

(*b*) *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 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.

We also perform the interval prediction using conditional percentiles, namely

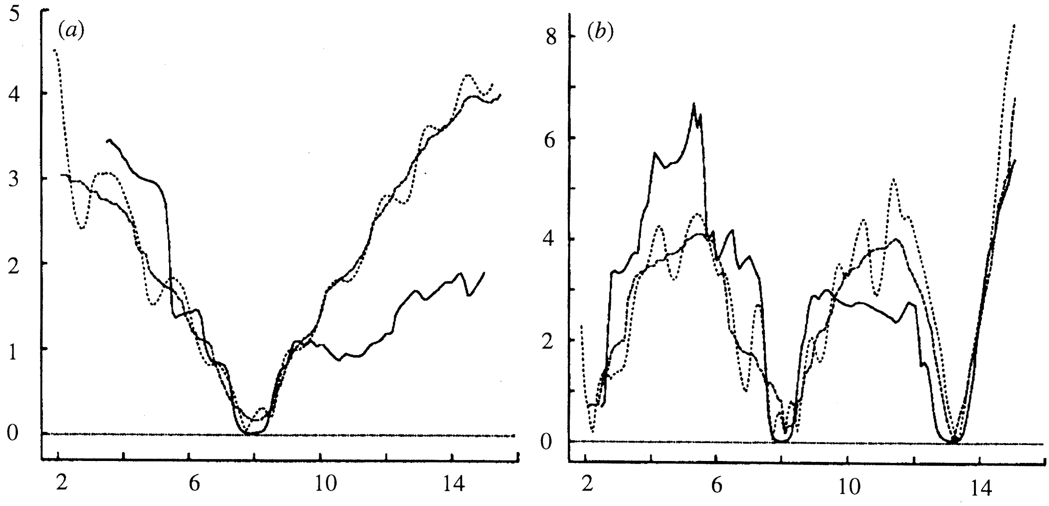

Figure 7. The estimated Fisher information $\hat{I}_m(x)$, and the derivatives of conditional percentiles and expectiles. (*a*) $m = 1$ ($h_1 = 0.61, h_2 = 0.24$ for $\hat{I}_1(x)$); (*b*) $m = 2$ ($h_1 = 0.57, h_2 = 0.22$ for $\hat{I}_2(x)$). Solid curve, $\hat{I}_m(x)$; dashed curve, $\{(\dot{\hat{\xi}}_{0.05,m}(x))^2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

Table 1. *Point prediction of the Canadian lynx data (on natural log scale)*

year	true value	error ($\hat{f}_1$)	$\Vert \hat{\lambda}_1 \Vert$	error ($\hat{f}_2$)	$\hat{\sigma}_2^2$	$\Vert \hat{\lambda}_2 \Vert$
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

$[\hat{\xi}_{0.05,m}, \hat{\xi}_{0.95,m}]$. The results are reported in table 2. The bandwidth is chosen as 0.57 for one-step prediction and 0.51 for two-step prediction. 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 for the case $m = 1$ (except for the year 1925).

This research was partly supported by the Science and Engineering Research Council (U.K.).

Appendix A. The regularity conditions

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

Table 2. *Interval prediction of the Canadian lynx data (on natural log scale)*

		predictive interval	
year	true value	$m = 1$	$m = 2$
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]

(A 1) The joint density of distinct elements of (X_1, Y_1, X_k, Y_k) is bound by a constant independent of k.

(A 2) X_t has the probability density function p, and $|p(x) - p(y)| \leqslant C \parallel x - y \parallel$ for any $x, y \in R^d$.

(A 3) The precess $\{Y_t\}$ is ρ-mixing, i.e.

$$\rho_j = \sup_{U \in \mathrm{Im}_{-\infty}^{0}, V \in \mathrm{Im}_j^{\infty}} \mathrm{Corr}(U, V) \to 0, \quad \text{as} \quad j \to \infty,$$

where Im_i^j is the σ-field generated by $\{Y_k, k = i, \ldots, j\}$. Further, we assume that $\sum_{k=1}^{\infty} \rho_k < \infty$.

(A 4) $K(\cdot)$ is a continuous density function with a bounded support in R^d, and

$$\int xK(x)\,\mathrm{d}x = 0, \qquad \int xx^{\mathrm{T}}K(x)\,\mathrm{d}x = \sigma_0^2 I_d,$$

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

(A 5) The bandwidth $h \to 0$, $nh^{2+d} \to \infty$, and $(\log n)/(nh^d) \to 0$.

(A 6) For any compact subset $B \in R^d$, there exists a constant c such that for any x, $y \in B$,

$$\left| \int z^2 g_m(z|x)\,\mathrm{d}z - \int z^2 g_m(z|y)\,\mathrm{d}z \right| \leqslant ||x - y||,$$

where $g_m(y|x)$ denotes the conditional density of Y_m given X_0.

References

Bloomfield, P. & Steiger, W. L. 1983 *Least absolute deviations.* Boston: Birkhäuser.

Chan, K. S. & Tong, H. 1994 A note on noisy chaos. *Jl R. statist. Soc.* B **56**, 301–311.

Deissler, R. J. & Farmer, J. D. 1989 Deterministic noise amplifiers. *Tech. Rep.* LA-UR-89-4236, Los Alamos Laboratory, U.S.A.

Eckmann, J. P. & Ruelle, D. 1985 Ergodic theory of chaos and strange attractors. *Rev. mod. Phys.* **57**, 617–656.

Fan, J. 1992 Design-adaptive nonparametric regression. *J. Am. statist. Ass.* **87**, 998–1004.

Fan, J., Hu, T. C. & Truong, Y. K. 1992 Robust nonparametric function estimation. *Tech. Rep.* 035-92, Mathematics Science Research Institute, Berkeley.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. & Engel, J. 1993 Local polynomial fitting: a standard for nonparametric regression. *Tech. Rep.*, University of North Carolina.

Fan, J., Yao, Q. & Tong, H. 1993 Estimating measures of sensitivity of initial values to nonlinear stochastic systems with chaos. *Tech. Rep.*, University of Kent.

Kullback, S. 1967 *Information theory and statistics.* New York: Dover.

Neway, W. K. & Powell, J. K. 1987 Asymmetric least squares estimation and testing. *Econometrica* **55**, 819–847.

Press, W. H., Flannery, B. P., Tenkolsky, S. A. & Vetterling, W. T. 1992 *Numerical recipes.* Cambridge University Press.

Tong, H. 1990 *Non-linear time series: a dynamical system approach.* Oxford University Press.

Yao, Q. & Tong, H. 1992 Asymmetric least squares regression estimation: a nonparametric approach. *Tech. Rep.*, University of Kent.

Yao, Q. & Tong, H. 1994 Quantifying the influence of initial values on nonlinear prediction. *Jl R. statist. Soc.* B **56**. (In the press.)