On initial-condition sensitivity and prediction in nonlinear stochastic systems

ON INITIAL-CONDITION SENSITIVITY AND PREDICTION IN NONLINEAR STOCHASTIC SYSTEMS

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

Summary

We describe two different approaches to defining the notion of initial-value sensitivity in a nonlinear stochastic system. The conditional distribution approach is directly relevant to the nonlinear prediction of time series. The kernel regression method based on a locally linear/quadratic fit is adapted to construct, with illustrations, pointwise predictors, predictive intervals and predictive distributions. We propose a bootstrap test for *operationally deterministic* versus stochastic nonlinear modelling, which is based on a newly proposed data-driven bandwidth selector.

1 Introduction

The notion of initial-value sensitivity is well established in and central to the study of deterministic chaos; it is a necessary but not sufficient condition for the generation of deterministic (equivalently low-dimensional) randomness. By contrast, for a stochastic dynamical system, the notion of initial-value sensitivity is not as well established although several different approaches are now available, namely the identical noise realization approach, the local Lyapunov exponent approach, the conditional mean approach and the conditional distribution approach. Their different motivations and interpretations will be briefly discussed in this paper. Clearly the notion is *not* necessary for the generation of stochastic (equivalently high-dimensional) randomness which is usually taken care of by the stochastic driving noise. Therefore the central issue is what we want the notion for. Our approach is partially motivated by its effects on prediction. We now

know that initial-value sensitivity is state-dependent; this fact has profound implications on the practice of prediction. Tong (1995) has stressed that this state-dependency leads to windows of opportunities which are open only to the nonlinear forecasters. Given the substantial recent interests in constructing nonparametric estimates of the point and interval predictors as well as the predictive distributions, we address the important problem of bandwidth selection and propose a practically useful method. We should point out that the importance of the bandwidth parameter goes far beyond the construction of a good kernel estimate. As a matter of fact, it can discriminate between operational determinism and stochastic randomness.

The paper is organized as follows. §2 provides a brief description of two different approaches of the initial-value sensitivity in stochastic systems. Some open problems are mentioned. §3 presents three types of predictors, namely a pointwise predictor, a predictive interval and a predictive distribution. The sensitivity measures studied in §2.2 play an important rôle in monitoring the performance of all the three types of predictors. In §4, a new data-driven bandwidth selector is introduced. Further, a bootstrap test for *operationally deterministic* versus stochastic nonlinear modelling is proposed based on the newly proposed data-driven bandwidth selector.

2 Initial-Value Sensitivity

From the studies in chaos, it has been learned that the sensitivity to initial conditions is an interesting phenomenon in nonlinear systems. This phenomenon has profound implications in nonlinear statistical analysis, especially in nonlinear prediction. Therefore, to define some appropriate measures to quantify the initial-value sensitivity in a stochastic environment is of importance in terms of both theoretical and practical interests. §2.1 and §2.2 present two different approaches to measuring the sensitivity. §2.3 discusses the noise amplification which is a different but relevant phenomenon in nonlinear stochastic systems.

2.1 Identical noise realizations

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

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

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 $E(e_t|X_0, \ldots, X_{t-1}) = 0$. As a start, it is perhaps

quite natural to borrow the concept of the Lyapunov exponent defined originally for deterministic systems. Quite a few publications were devoted to this approach, for example Crutchfield *et al.* (1982), Kifer (1986), Herzel *et al.* (1987), Nychka *et al.* (1992), and Lu (1994). The idea can be described as follows. To see whether a stochastic dynamical system (2.1) is sensitive to its initial values, conceptually we can imagine that the system starts at time $t = 0$ from two nearby initial values say x and $x + \delta$, and the two trajectories always receive the same realization of e_t , for all $t = 1, 2, \ldots$. Then, the divergence of the two trajectories at time m can be approximated, in the case $d = 1$, by

$$|X_m(x + \delta) - X_m(x)| \approx \exp\{m\lambda\}|\delta|, \tag{2.2}$$

where $\lambda = \lim_{n\to\infty} \frac{1}{n} \log |\prod_{i=0}^{n-1} \dot{F}(X_i(x))|$ is called the Lyapunov exponent, and $\dot{F}(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$. It is easy to see that if the system is ergodic, this limit exists and equals to $\mathrm{E}\{\log |\dot{F}(X_1)|\}$. Similar results for the case $d \geq 1$ can be found in Nychka *et al.* (1992), and Lu (1994).

The above approach has the following advantages: it is an obvious analogue to the sensitivity measure in purely deterministic systems, and the asymptotic results are ready to enhance the understanding of the derived measure. However, the asymptotic approximation (2.2) is too rough to use in practice even for the cases with large m , because the RHS expression of (2.2) only offers an approximation of the divergence along the trajectories which simply keep twisting whilst diverging, thereby disqualifying the expression as an approximation for the distance between $X_m(x + \delta)$ and $X_m(x)$. Furthermore, in many practical applications, for example prediction, it is not always possible to justify the assumption of identical noise realization.

The estimation of λ has been discussed by Dechert and Gencay (1990) and Nychka *et al.* (1992) using the neural network model based method, and by Lu (1994) using the locally polynomial regression method.

2.2 Conditional distribution approach

Partially motivated by the studies of nonlinear prediction, a different approach has been adopted by Yao and Tong (1994a,b). They consider the sensitivity of the conditional distribution, or one of its characteristics (e.g. the conditional mean), with respect to initial values. For model (2.1), let $F_m(x) = \mathrm{E}\{X_m|X_0 = x\}$, and $g_m(.|x)$ denote the conditional distribution of X_m given $X_0 = x$.

It follows from Taylor's expansion that

$$F_m(x+\delta) - F_m(x) = \dot{F}_m(x)\delta + o(||\delta||). \tag{2.3}$$

By the chain rule,

$$\dot{F}_m(x) = \mathrm{E}\{\prod_{i=0}^{m-1} \dot{F}(X_i)|X_0 = x\}. \tag{2.4}$$

Roughly speaking, assuming that all the factors on the RHS of the above expression are of comparable size, it seems plausible that $\dot{F}_m(x)$ grows (or decays) exponentially with m. Therefore, it follows from (2.3) that the conditional means diverge (converge) exponentially with m.

The above observation has an important implication in pointwise prediction (cf. §3.1 below). This conditional mean approach appears to be closely related to the approach of Wolff (1992). For example, for the case $d = 1$,

$$\nu_m(x) \equiv \frac{1}{m}\log|\dot{F}_m(x)| = \frac{1}{m}\log|\mathrm{E}\{\prod_{i=0}^{m-1} \dot{F}(X_i)|X_0 = x\}| \tag{2.5}$$

would probably provide the same profile as Wolff's m-step ahead local Lyapunov exponent.

A more informative way is to consider the global deviation of the conditional distribution of X_m given X_0. We consider the mutual information based on the Kullback-Leibler information, which may be expressed as follows.

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

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

$$K_m(x;\delta) = \delta^\tau I_{1,m}(x)\delta + o(||\delta||^2), \tag{2.6}$$

where

$$I_{1,m}(x) = \int \dot{g}_m(y|x)\dot{g}_m^\tau(y|x)/g_m(y|x)\mathrm{d}y. \tag{2.7}$$

Therefore, if we treat the initial value x as a parameter vector of the distribution, $I_{1,m}(x)$ is the Fisher's information matrix, which represents the information on the initial value $X_0 = x$ contained in X_m. Roughly speaking, (2.6) may be interpreted as saying that the more information X_m contains about the parameter, the more sensitively the distribution depends on the initial condition. The converse is also true. It will be seen that $I_{1,m}(x)$ has a direct application in interval prediction (see §3.3.1 below).

We also consider a simple L_2-distance defined as

$$D_m(x;\delta) = \int \{g_m(y|x+\delta) - g_m(y|x)\}^2\mathrm{d}y.$$

It follows from the Taylor's expansion that

$$D_m(x;\delta) = \delta^\tau I_{2,m}(x)\delta + o(||\delta||^2),$$

where

$$I_{2,m}(x) = \int \dot{g}_m(y|x)\dot{g}_m^\tau(y|x)\mathrm{d}y.$$

In §3.2.2 below, a direct estimator for $I_{2,m}(x)$ can be constructed.

To see that measures $K_m(.;.)$ and $D_m(.;.)$ are more informative than the measure derived from the conditional mean approach, let us consider the following one-dimensional model

$$Y_{t+1} = \alpha Y_t + \sigma(Y_t)\epsilon_{t+1},$$

where α is a real constant, $\sigma(.)$ is positive and differentiable function, and $\{\epsilon_t\}$ are i.i.d. standard normal. Obviously, the conditional mean $F_m(x) = \mathrm{E}\{Y_m|Y_0 = x\} = \alpha^m x$. Therefore, by (2.5), $\nu_m = \log|\alpha|$, a constant, which indicates that when $|\alpha| < 1$, the system is globally as well as locally stable as far as the conditional mean is concerned. However, it is easy to see that $K_m(.;.)$ and $D_m(.;.)$ are no longer constants. For example,

$$I_{1,1}(x) = \frac{1}{\sigma^2(x)}\{\alpha^2 + 2[\dot{\sigma}(x)]^2\}, \qquad I_{2,1}(x) = \frac{1}{4\sqrt{\pi}\sigma^3(x)}\left\{\alpha^2 + \frac{3}{2}[\dot{\sigma}(x)]^2\right\}.$$

Therefore, there is some variation in the sensitivity of the conditional distribution with respect to the initial value x, which is due to the presence of the conditional heteroscedasticity in the model.

All the above measures are defined for fixed m. Intuitively, when $m \to \infty$, the conditional expectation $F_m(x)$ will tend to the unconditional expectation, and both $K_m(x;\delta)$ and $D_m(x;\delta)$ will converge to 0, because the system will eventually lose its memory on its initial state due to the accumulation of random noise in time evolution. However, it will be interesting to see whether the second term in an asymptotic expansion of $K_m(x;\delta)$, or $D_m(x;\delta)$, or even $F_m(x+\delta) - F_m(x)$ will give us an indication of the sensitivity of the system. As far as we know, this still remains an open problem. Also open is the existence of the limit of $\nu_m(x)$ as $m \to \infty$.

To end the discussion on sensitivity, we mention a trivial asymptotic result which indicates that if in model (2.1) the noise is small, the conditional expectation approach will give about the same measure as the Lyapunov exponent defined in §2.1. For simplicity of presentation, we state the result in the case $d = 1$.

Proposition 1. Suppose that a one-dimensional random system is defined by

$$Y_{t+1,m} = f(Y_{t,m}) + \sigma_m \epsilon_{t+1}, \quad t = 1, \ldots, m;\ m = 1, 2, \ldots,$$

where f is bounded and has continuous second derivative, $\{\epsilon_t\}$ are independent and with a common density function on a bounded support. Then as $m \to \infty$,

$$\frac{1}{m} \log |\mathrm{E}\{ \prod_{i=0}^{m-1} \dot{f}(Y_{i,m}) | Y_{0,m} = x\}| \to \lambda$$

provided $\sigma_m = o(m^{-1})$ and

$$\frac{1}{m} \sum_{i=0}^{m-1} \log |\dot{f}\{f^{(i)}(x)\}| \to \lambda,$$

where λ is a finite constant which may depend on x and $f^{(i)}$ denotes the i-fold composition of f.

The estimation of $\dot{F}_m(x)$ will be discussed in §3.1 together with the context of the pointwise prediction. The estimation of $K_m(x;\delta)$ and $D_m(x;\delta)$ will be discussed in §3.2 together with the estimation of predictive distributions.

2.3 Noise amplification

Another interesting feature of a nonlinear system is noise amplification. We can measure it 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 $F(.)$.

Assume that in model (2.1), $F(.)$ has a bounded second derivative, and

$$\mathrm{E}e_t = \mathrm{E}\{e_t | X_{t-1}, X_{t-2}, \ldots\} = 0, \quad \mathrm{Var}(e_t) = \mathrm{Var}\{e_t | X_{t-1}, X_{t-2}, \ldots\} = \Sigma.$$

We also assume that, for all $t \geq 1$, $||e_t|| \leq \zeta$ almost surely, where ζ is a small constant. It can be proved that

$$\mathrm{Var}\{X_m | X_0 = x\} = \Sigma + \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \Lambda\{F^{(k)}(x)\} \right\} \Sigma \left\{ \prod_{k=j}^{m-1} \Lambda\{F^{(k)}(x)\} \right\}^{\tau} + O(\zeta^3),$$

where $F^{(k)}$ denotes the k-fold composition of F, $\Lambda(x) = \partial F(x)/\partial x^{\tau}$, and $O(\zeta^3)$ stands for a $d \times d$ matrix with all entries of the order ζ^3. The special case of $d = 1$ was presented in Yao and Tong

(1994a), which has the simpler form: $\sigma_m^2(x) \equiv \mathrm{Var}(Y_m|Y_0 = x) = \sigma_0^2\mu_m(x)\{1 + o(1)\}$, where $\sigma_0^2 = \mathrm{Var}(e_t)$, and

$$\mu_m(x) = 1 + \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \dot{F}[F^{(k)}(x)] \right\}^2 . \tag{2.8}$$

3 Prediction

Studies on pointwise prediction of nonlinear time series have revealed three distinguishing features of nonlinear prediction: (i) the *dependence* of the prediction accuracy on the current position in the state space; (ii) the *sensitivity* of the predictor to the current state; and (iii) the *non-monotonicity* of the prediction accuracy in multi-step prediction. These results will be reviewed briefly in §3.1. §3.2 discusses the estimation of predictive distributions. As a by-product, the estimation of sensitivity measures derived in §2.2 will also be developed. It will be pointed out in §3.3 that interval predictors constructed in terms of conditional percentiles are not always appropriate for multi-step ahead prediction. Two alternatives will be suggested.

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

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

where $\epsilon_t = Y_t - f(X_{t-1})$. Define $F(X_{t-1}) = (f(X_{t-1}), Y_{t-1}, \ldots, Y_{t-d+1})^T$, $e_t = (\epsilon_t, 0, \ldots, 0)^T$. Then equation (2.1) holds.

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 polynomial fit (cf. Fan 1992) for estimation. 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 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 $m \ge 1$. It is easy to see that the (theoretical) least squares predictor of Y_{n+m} based on $\{Y_t, t \le n\}$ is

$f_m(X_n)$. In practice, we use $\hat{f}_m(X_n)$ as the predictor, where $\hat{f}_m(.)$ is any *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.,

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

where $R_m = o(||\delta||^2)$ as $||\delta|| \to 0$, and $\sigma_m^2(x) = \mathrm{Var}(Y_m|X_0 = x)$ (cf. Yao and Tong 1994a).

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 $\dot{f}_m$.

A few remarks are now in order.

(i) The variation of the conditional variance $\sigma_m^2(x)$ indicates that the accuracy of the prediction depends on initial values.

(ii) When the value $||\dot{f}_m(x)||$ is large (note $\dot{f}_m(x)$ could be considered the first row of the matrix defined in (2.4) if we define F in the way stated below (3.1)), $\{\delta^\tau \dot{f}_m(x)\}^2$ could be large. In this sense, we say that the prediction depends on the initial value x sensitively.

(iii) The m-step ahead prediction is not necessarily more accurate that an $(m+1)$-step ahead prediction. For example, in the case that $d = 1$ and the stochastic noise in the model is small, the conditional variance $\sigma_m^2(x)$ is described by the function $\mu_m(x)$ defined in (2.8). Note that $\mu_{m+1}(x) < \mu_m(x)$ if $\{\dot{f}[f^{(m)}(x)]\}^2 < 1 - 1/\mu_m(x)$, which could imply that $\sigma_{m+1}^2(x) < \sigma_m^2(x)$. (See Fig. 4(c) in Yao and Tong 1994a.)

(iv) In practice, δ is unobservable. However, sometimes we could assume that δ is a random variable independent of $\{Y_t\}$, its density function has the form $\frac{1}{\sigma_\delta} g(\frac{\cdot}{\sigma_\delta})$, where $g(.)$ is a density function on a bounded support in R^d, and $\int g(u)\mathrm{d}u = 0$ and $\int uu^\tau g(u)\mathrm{d}u = \Sigma$. Then, it can be proved that

$$\lim_{n\to\infty} \mathrm{E}\{(Y_t - \hat{f}_m(x))^2|X_0 - \delta = x\} = \mathrm{E}\{\sigma_m^2(x+\delta)\} + \sigma_\delta^2 \dot{f}_m^\tau(x)\Sigma\dot{f}_m(x) + R_m,$$

where $R_m = o(\sigma_\delta^2)$.

(v) Estimators for $f_m(x)$, $\dot{f}_m(x)$ and others may be obtained by using the locally linear regression and have been presented in Yao and Tong (1994a). Further discussion of (3.2) can be found in Yao and Tong (1994a) and Tong (1995). Examples of applications of (3.2) were given in Yao and Tong (1994a).

3.2 Predictive distributions

A more informative way is to estimate the predictive distribution. Let $g_m(.|x)$ be the distribution of Y_m given $X_0 = 0$. Fan, Yao and Tong (1993) proposed the following locally quadratic regression method to estimate predictive distribution g_m, which also leads to estimates for the sensitive measures derived in §2.2.

3.2.1 Estimation of $g_m(.|x)$

Let $K(.)$ be a density function, and $K_h(z) = K(z/h)/h$. Note that

$$E\left(K_{h_2}(Y_m - y)|X_0 = x\right) \approx g_m(y|x), \quad \text{as } h_2 \to 0.$$

The LHS of the above expression can be regarded as the regression function of the data $\{K_{h_2}(Y_{m+t} - y)\}$ on $\{X_t\}$. By Taylor's expansion about $x = (x_1, \ldots, x_d)^\tau \in R^d$, we have

$$\begin{aligned} E\left(K_{h_2}(Y_m - y)|X_0 = z\right) &\approx g_m(y|z) \\ &\approx g_m(y|x) + \dot{g_m}(y|x)^\tau (z - x) + \frac{1}{2}(z - x)^\tau \ddot{g_m}(y|x)(z - x) \\ &\equiv \beta_0 + \beta_1^\tau (z - x) + \beta_2^\tau \text{vec}\{(z - x)(z - x)^\tau\}, \end{aligned}$$

where $\dot{g}_m(y|x) = \frac{\partial g_m(y|x)}{\partial x^\tau}$, $\ddot{g}_m(y|x)$ is the Hessian matrix of $g_m(y|x)$ with respect to x, $\text{vec}(A) = (a_{11}, a_{22}, \ldots, a_{d,d}, a_{12}, \ldots, a_{1,d}, a_{23}, \ldots, a_{d-1,d})^T \in R^{d(d+1)/2}$ for any $d \times d$ symmetric matrix $A = (a_{ij})$, and

$$\begin{aligned} \beta_2 = \Bigg(&\frac{\partial^2 g_m(y|x)}{2\partial x_1^2}, \frac{\partial^2 g_m(y|x)}{2\partial x_2^2}, \ldots, \frac{\partial^2 g_m(y|x)}{2\partial x_d^2}, \frac{\partial^2 g_m(y|x)}{\partial x_1 \partial x_2}, \ldots, \\ &\frac{\partial^2 g_m(y|x)}{\partial x_1 \partial x_d}, \frac{\partial^2 g_m(y|x)}{\partial x_2 \partial x_3}, \ldots, \frac{\partial^2 g_m(y|x)}{\partial x_{d-1} \partial x_d} \Bigg)^\tau. \end{aligned}$$

Considerations of this nature suggest the following least squares problem: let $\hat{\beta}_0$ and $\hat{\beta}_1$ and $\hat{\beta}_2$ minimize

$$\sum_{t=1}^{n-m} \left(K_{h_2}(Y_{t+m} - y) - \beta_0 - \beta_1^\tau (X_t - x) - \beta_2^\tau \text{vec}\{(X_t - x)(X_t - x)^\tau\}\right)^2 W_{h_1}(X_t - x), \qquad (3.3)$$

where W is a nonnegative function, which serves as a kernel function, and h_1 is the bandwidth, controlling the size of the local neighborhood. Then, clearly $\hat{\beta}_0$ and $\hat{\beta}_1$ estimate $g_m(y|x)$ and $\dot{g}_m(y|x)$ respectively, namely,

$$\hat{g}_m(y|x) = \hat{\beta}_0 \text{ and } \dot{g}_m(y|x) = \hat{\beta}_1.$$

The least-square theory provides the solution:

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1^\tau, \hat{\beta}_2^\tau)^\tau = (X^\tau W X)^{-1} X^\tau W Y, \tag{3.4}$$

where X is the design-matrix of the least-square problem (3.3), $W = \text{diag}(W_{h_1}(X_1 - x), \cdots,$ $W_{h_1}(X_{n-m} - x))$, and $Y = (K_{h_2}(Y_{1+m} - y), \cdots, K_{h_2}(Y_n - y))^\tau$. Fan, Yao and Tong (1993) has proved that the estimators $\hat{g}_m(y|x)$ and $\dot{g}_m(y|x)$ are asymptotically normal, and discussed the issue of the selection of bandwidths h_1 and h_2.

If we use locally constant fitting, *i.e.* let β_1 and β_2 be 0 in (3.3), the least squares approach will lead to the conventional kernel estimator for the conditional density function (cf. Rosenblatt 1969).

3.2.2 Estimation of $I_{1,m}(x)$ and $I_{2,m}(x)$

For simplicity of presentation, from now on we treat only a univariate case, i.e. $d = 1$ and $X_t = Y_t$. However, both the theory and the method generalize in an obvious way to the multivariate case but with more complicated notation. For the univariate case, the estimates in (3.4) can be expressed as

$$\hat{\beta}_j(x, y) = h_1^{-1} \sum_{t=1}^{n-m} W_j^n \left(\frac{Y_t - x}{h_1} \right) K_{h_2}(Y_{t+m} - y), \quad j = 0, 1, \tag{3.5}$$

where

$$W_j^n(t) = e_j^T S_n^{-1} (1, h_1 t, h_1^2 t^2)^T \times W(t),$$

with e_j being the unit vector with $(j+1)^{th}$ element 1 and

$$S_n = \begin{pmatrix} s_{n,0} & s_{n,1} & s_{n,2} \\ s_{n,1} & s_{n,2} & s_{n,3} \\ s_{n,2} & s_{n,3} & s_{n,4} \end{pmatrix}, \quad s_{n,j} = \sum_{t=1}^{n-m} (Y_t - x)^j W_{h_1}(Y_t - x).$$

With the derivative of the conditional density estimated by (3.5), a natural estimator for $I_{2,m}(x)$ is

$$\begin{aligned} \hat{I}_{2,m}(x) &= \int \hat{\beta}_1^2(x, y) \mathrm{d}y \\ &= \frac{1}{h_1^2} \sum_{i=1}^{n-m} \sum_{j=1}^{n-m} W_1^n \left(\frac{Y_i - x}{h_1} \right) W_1^n \left(\frac{Y_j - x}{h_1} \right) \int K_{h_2}(Y_{i+m} - y) K_{h_2}(Y_{j+m} - y) \mathrm{d}y. \end{aligned}$$

Assume that the kernel $K(\cdot)$ is symmetric. Then,

$$\int K_{h_2}(Y_i - y) K_{h_2}(Y_j - y) \mathrm{d}y = K_{h_2}^*(Y_i - Y_j),$$

where $K^* = K * K$ is a convolution of the kernel function K with itself. Thus, the proposed estimator can be expressed as

$$\hat{I}_{2,m}(x) = \frac{1}{h_1^2} \sum_{i=1}^{n-m} \sum_{j=1}^{n-m} W_1^n \left(\frac{X_i - x}{h_1} \right) W_1^n \left(\frac{X_j - x}{h_1} \right) K_{h_2}^*(Y_{i+m} - Y_{j+m}). \tag{3.6}$$

The asymptotic normality for the above estimator has been established (cf Fan, Yao and Tong 1993).

Analogously, an estimator for $I_{1,m}(x)$ can be defined by

$$\hat{I}_{1,m}(x) = \int \hat{\beta}_1^2(x,y)/\hat{\beta}_0(x,y) \mathrm{d}y, \tag{3.7}$$

with the usual convention $0/0 = 0$. The above integration is typically finite under some mild conditions. However, the estimator (3.7) cannot be simplified easily.

An alternative estimator to $I_{1,m}(x)$ originates from the fact that

$$I_{1,m}(x) = 4 \int \left\{ \frac{\mathrm{d}\sqrt{g_m(y|x)}}{\mathrm{d}x} \right\}^2 \mathrm{d}y.$$

For given bandwidths h_1 and h_2, define

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

$$C(Y_i) = \#\{Y_t, 1 \le t \le n-m, \ : \ |Y_t - Y_i| \le h_1\},$$

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

$$Z_t \equiv [C(Y_t, Y_{t+m})/\{C(Y_t)\, h_2\}]^{1/2}$$

is a natural estimate of $q(x,y) \equiv \{g_m(y|x)\}^{1/2}$ at $(x,y) = (Y_t, Y_{t+m})$. Fitting it into the context of locally quadratic regression, we may estimate $q(x,y)$, and its first and second order partial derivatives with respect to x which are denoted by $\dot{q}(x,y)$ and $\ddot{q}(x,y)$ respectively, by using $\hat{q}(x,y) = \hat{a}$, $\hat{\dot{q}}(x,y) = \hat{b}$ and $\hat{\ddot{q}}(x,y) = \hat{c}$, where $(\hat{a}, \hat{b}, \hat{c})$ are the minimizers of the function

$$\sum_{t=1}^{n-m} \{Z_t - a - b(Y_t - x) - c(Y_t - x)^2/2\}^2 H\left(\frac{Y_t - x}{h_1}, \frac{Y_{t+m} - y}{h_2} \right),$$

H being a probability density function on R^2. Consequently, we estimate $I_{1,m}(x)$ by

$$\hat{I}_{1,m}(x) = 4 \int \{\hat{\dot{q}}(x,y)\}^2 \, \mathrm{d}y.$$

3.3 Interval predictors

3.3.1 Sensitivity to initial values

For model (3.1), the conditional distribution of Y_{n+m} given $\{Y_t,\ t \le n\}$ depends on $\{Y_t,\ t \le n\}$ only through X_n. Given the distribution, we could construct a predictive interval for Y_{n+m} based on the past data X_n only. Suppose that $\Omega_m(X_n)$ is such an interval with the coverage probability $1-\alpha$, i.e.

$$P\{Y_m \in \Omega_m(x)|X_0 = x\} = 1-\alpha. \tag{3.8}$$

Inspired by the studies in pointwise prediction, a natural question is how sensitively the cover probability depends on the initial value x. The following Proposition 2 indicates that the sensitivity can be monitored by the measure $I_{1,m}(x)$ introduced in §2.2.

Proposition 2. Let $g_m(.|x)$ be the conditional density function of Y_m given $X_0 = x \in R^d$, and all the second partial derivatives of $g_m(y|x)$ respect to x are bounded. Then for any bounded $\Omega_m(.)$ satisfying (3.8) and $\delta \in R^d$,

$$\begin{aligned} & |P\{Y_m \in \Omega_m(x)|X_0 = x+\delta\} - (1-\alpha)| \\ \le\ & (1-\alpha)^{\frac{1}{2}}\{\delta^\tau I_{1,m}(x)\delta\}^{\frac{1}{2}} + O(||\delta^2||) \ \le\ ||\delta||(1-\alpha)^{\frac{1}{2}}[\mathrm{tr}\{I_{1,m}(x)\}]^{\frac{1}{2}} + O(||\delta^2||), \end{aligned}$$

where $I_{1,m}(x)$ is defined as in (2.7).

Proof. It follows from (3.8) that

$$\begin{aligned} & P\{Y_m \in \Omega_m(x)|X_0 = x+\delta\} = \int_{\Omega_m(x)} g_m(y|x+\delta)\mathrm{d}y \\ =\ & \int_{\Omega_m(x)} \{g_m(y|x) + \delta^\tau \dot{g}_m(y|x)\}\mathrm{d}y + O(||\delta||^2) = 1-\alpha + \int_{\Omega_m(x)} \delta^\tau \dot{g}_m(y|x)\mathrm{d}y + O(||\delta||^2). \end{aligned}$$

By Cauchy-Schwarz inequality that

$$\begin{aligned} \left|\int_{\Omega_m(x)} \delta^\tau \dot{g}_m(y|x)\mathrm{d}y\right| &\le \left\{\int_{\Omega_m(x)} g_m(y|x)\mathrm{d}y \int \frac{\{\delta^\tau \dot{g}_m(y|x)\}^2}{g_m(y|x)}\mathrm{d}y\right\}^{\frac{1}{2}} \\ &\le \{(1-\alpha)\delta^\tau I_{1,m}(x)\delta\}^{\frac{1}{2}}. \end{aligned}$$

3.3.2 Percentiles and expectiles

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 percentile of Y_m given $X_0 = x \in R^d$ is defined as

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

where the loss function

$$R_\alpha(y) = \begin{cases} (1-\alpha)|y| & y \le 0, \\ \alpha|y| & y > 0. \end{cases} \tag{3.9}$$

In fact, 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 $1-\alpha$.

Similar to §3.2, 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

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

where $K(.)$ is a probability density function on R^d, and h is a bandwidth.

An alternative approach is to change the loss function (3.9) 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). Note that $\tau_{\omega,m}(x)$ reduces to $E(Y_m|X_0 = x)$ when $\omega = \frac{1}{2}$. It can be proved that

$$\omega = \frac{E\{\, |Y_m - \tau_{\omega,m}(x)|\ I_{\{Y_m \le \tau_{\omega,m}(x)\}} \mid X_0 = x\}}{E\{\, |Y_m - \tau_{\omega,m}(x)| \mid X_0 = x \,\}}.$$

Now, $\tau_{\omega,m}(x)$ can also be used to construct a predictive interval: given $\{Y_t,\ t \le n\}$, 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 and Tong 1995a). 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 as a reasonable interval predictor extended from the conditional expectation. Yao and Tong (1995a) has pointed out that, in a special case, the above asymmetric least squares approach can be used to estimate conditional percentiles directly.

In practice, we use the following two kinds of intervals to predict Y_{n+m} from $\{Y_t,\ t \leq n\}$,

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

The asymptotic normality of the estimators for $\xi_{\alpha,m}(x)$, $\dot{\xi}_{\alpha,m}(x)$, $\tau_{\omega,m}(x)$, and $\dot{\tau}_{\omega,m}(x)$ were presented in Yao and Tong (1994b).

3.3.3 Two alternative predictors

For $m > 1$, the m-step ahead predictive distribution $g_m(.|x)$ is not always symmetric and unimodal. When $g_m(.|x)$ is asymmetric or multi-modal, the predictive intervals constructed in §3.3.2 lose their appeal. To cope with such cases, we now give alternatives based on the predictive distribution as follows.

Maximum density region: For a given $x \in R^d$ and an $\alpha \in (0,1)$, the maximal density region is defined as

$$D_m(x) = \{y | g_m(y|x) \geq l_m(x,\alpha)\},$$

where $l_m(x,\alpha) > 0$ is determined by

$$\int_{D_m(x)} g_m(y|x)\mathrm{d}y = 1-\alpha.$$

Mode-based interval: If the density function $g_m(.|x)$ has a unique mode $\theta_m(x)$, i.e. $g_m(\theta_m(x)|x) > g_m(y|x)$ for all $y \neq \theta_m(x)$, then the mode based interval is defined as $[\theta_m(x) - b_m(x,\alpha),\ \theta_m(x) + b_m(x,\alpha)]$, where $b_m(x,\alpha) > 0$ is determined by

$$\int_{\theta_m(x)-b_m(x,\alpha)}^{\theta_m(x)+b_m(x,\alpha)} g_m(y|x)\mathrm{d}y = 1-\alpha.$$

Note that the maximum density region is no longer an interval if the density function $g_m(.|x)$ is multi-modal. In practice, we estimate $g_m(.|x)$ using the method described in §3.2.1 first and then we estimate the above two 'end points' based on the estimated $g_m(.|x)$.

4 Bandwidths

All our estimates discussed so far are based on the nonparametric kernel regression with locally linear fit or locally quadratic fit. Of great importance in nonparametric kernel regression is the bandwidth choice. In practice, a good automatically selected bandwidth is always a useful starting point.

For independent observations, the most frequently used technique is to choose h automatically by cross-validation. However, it has been well documented that if the observations are dependent, especially if the errors of the model are dependent, the cross-validation will not always produce good bandwidths (cf. for example, Altman 1990, Hart 1991, and Hart 1994). In order to cope with the possible dependence among data, different modifications for cross-validation method have been suggested. (See, for example, Härdle and Vieu 1992, Marron 1987, Chu and Marron 1991, Hart 1994 and etc.) Yao and Tong (1995b) suggested a new method to modify the cross-validation method to cope with the dependence in the data, which will be reported briefly in §4.1.

Moreover, based on the data-driven bandwidth selector, a bootstrap test can be constructed to test whether in (2.1) the noise term e_t is sufficiently small so that the model could be treated as a deterministic system, or rather an *operationally deterministic* system (see §4.2 below).

To simplify the notation, we only consider the locally linear regression estimates for one-dimensional models. However, the methods and the theory readily extend to other kernel regression methods and to higher dimensional models. The model and the estimator which will be discussed in this section can be described as follows.

Suppose that $\{X_t,\ Y_t\}$ is a strictly stationary process, and both X_t and Y_t are scalar. Suppose that our interest is to estimate the regression function $f(x) = \mathrm{E}\{Y_1|X_1 = x\}$. Write

$$Y_t = f(X_t) + \epsilon_t, \tag{4.1}$$

where $\epsilon_t = Y_t - \mathrm{E}\{Y_t|X_t\}$. In the case that $X_t = Y_{t-1}$, $f(.)$ is the auto-regression function for the time series $\{Y_t\}$. Given the observations $\{(X_t,\ Y_t);\ 1 \le t \le n\}$, the locally linear regression estimator of $f(x)$ is $\hat{f}_{n,h}(x) = \hat{a}$, where

$$(\hat{a}, \hat{b}) = \mathrm{argmin}_{(a,b)\in R^2} \sum_{t=1}^{n} \{Y_t - a - b(X_t - x)\}^2 K\left(\frac{X_t - x}{h}\right), \tag{4.2}$$

where $K(.)$ is a density function, and h is the bandwidth.

4.1 A data-driven bandwidth selector

We shall omit all details of the assumptions on the mixing condition of the processes and the other regularity conditions required to ensure the validity of the results discussed below. They can be found in Yao and Tong (1995b).

It follows from Theorem 2 of Yao and Tong (1995b) that

$$\mathrm{E}\{\hat{f}_{n,h}(x) - f(x)\}^2 \approx \frac{h^4}{4}\sigma_0^4\{\ddot{f}(x)\}^2 + \frac{1}{nhp(x)}\sigma^2(x)\int K^2(u)\mathrm{d}u,$$

where $\sigma_0^2 = \int u^2 K(u)\mathrm{d}u$, $\sigma^2(x) = \mathrm{Var}(Y_1|X_1 = x)$, and $p(x)$ is the marginal density function of X_1. To minimize the above approximated MES, the bandwidth should be equal to

$$h_n(x) = \frac{1}{n^{1/5}}\left(\frac{\sigma^2(x)\int K^2(u)\mathrm{d}u}{p(x)\sigma_0^4\{\ddot{f}(x)\}^2}\right)^{1/5}.$$

The above bandwidth cannot be directly applied in practice because it depends on various unknown functions. However, it does indicate that a reasonable bandwidth is of the order

$$h \propto n^{-\frac{1}{5}}, \tag{4.3}$$

which will play an important rôle in the following proposed method for choosing h.

We split the sample into two pieces $\{(X_t, Y_t),\ 1 \le t \le n_1\}$, and $\{(X_t, Y_t),\ n_1 < t \le n\}$ for some $n_1 < n$. We estimate $f(.)$ using the first n_1 observation. The estimator given by (4.2) is denoted by $\hat{f}_{n_1,h}(.)$. We choose h such that $\hat{f}_{n_1,h}(.)$ gives the best prediction for Y_t for $n_1 < t \le n$ in the sense that h_{n_1} minimizes

$$\mathrm{ECV}_{n_1}(h) \equiv \frac{1}{n - n_1}\sum_{t=n_1+1}^{n}\{Y_t - \hat{f}_{n_1,h}(X_t)\}^2 w(X_t) \tag{4.4}$$

over $h \in H_{n_1}$, where $w(.) \ge 0$ is a weight function, and

$$H_n = [an^{-\frac{1}{5}-\varepsilon_0},\ bn^{-\frac{1}{5}+\varepsilon_0}].$$

In the above expression, $0 < a < b < \infty$, and $\varepsilon_0 \in (0, \frac{1}{150})$ are some constants.

In the light of (4.3), for the estimator $\hat{f}_{n,h}$ which is based on the whole sample, we propose the bandwidth choice

$$\hat{h}_n = \hat{h}_{n_1}\left(\frac{n_1}{n}\right)^{1/5}. \tag{4.5}$$

The above approach could still be viewed as a generalization of cross-validation. In order to overcome the drawbacks of the ordinary cross-validation for dependent data, we leave the last $n - n_1$ observations out.

Note that the best (pointwise) prediction for Y_t based on X_t is $f(X_t) = \mathrm{E}\{Y_t|X_t\}$. To justify the above approach, we compare the $\hat{h}_n$ with the bandwidth which minimizes the average squared errors of the *fictitious* post samples $\{(X_t, Y_t),\ t = n+1, \ldots, n_2\}$ $(n_2 > n)$

$$M_n(h) = \frac{1}{n_2 - n} \sum_{t=n+1}^{n_2} \{\hat{f}_{n,h}(X_t) - f(X_t)\}^2 w(X_t).$$

If n/n_2 converges to a positive constant, Theorem 3 of Yao and Tong (1995b) shows that

$$M_n(h) \sim \frac{h^4}{4}\sigma_0^2 \int \{\ddot{f}(x)\}^2 p(x)w(x)\mathrm{d}x + \frac{1}{nh}\int \sigma^2(x)w(x)\mathrm{d}x \int K^2(u)\mathrm{d}u, \tag{4.6}$$

uniformly for $h \in H_n$. It is easy to see that the minimizer of the RHS of the above expression is

$$h_n = \frac{\alpha}{n^{1/5}} \left\{ \frac{\int \sigma^2(x)w(x)\mathrm{d}x}{\int \{\ddot{f}(x)\}^2 p(x)w(x)\mathrm{d}x} \right\}^{1/5}, \tag{4.7}$$

where $\alpha > 0$ is a constant which only depends on the kernel $K(.)$. Yao and Tong (1995b) proved that if n_1/n converges to some positive constant,

$$\frac{\hat{h}_n - h_n}{h_n} \xrightarrow{P} 0. \tag{4.8}$$

Some simulation studies on $\hat{h}_n$ have been reported in Yao and Tong (1995b). Perhaps the most obvious drawback of the method is that the data have not been used in the most efficient way, comparing with the ordinary cross validation method. However, it preserves the dependence on the data while selecting the bandwidth. Apparently, it saves considerable computational time comparing to other cross validation methods, which will also be taken into account when we construct a bootstrap test based on an automatic bandwidth selector.

4.2 Tests for operationally deterministic systems

To distinguish between deterministic chaos and nonlinear stochastic systems is always an interesting and somehow intriguing challenge (cf. Farmer and Sidorowich 1987, Sugihara and May 1990, for example). Casdagli (1992) constructed an ingenious forecasting algorithm using the k nearest neighbours, and he claimed that 'a small value of k corresponds to a deterministic approach to modelling. The largest value of k corresponds to fitting a stochastic linear autoregressive model. Intermediate values of k correspond to fitting non-linear stochastic models'. Unfortunately, as far as we know there is to-date no theoretical justification for his data-analytic technique. In fact, similar features can be observed within the framework of kernel regression. Moreover, a theoretical justification has been given by Yao and Tong (1995b), which we summarize below.

Strictly speaking, our theorems do not apply to the purely deterministic system (i.e. $\epsilon_t = 0$ $a.s.$ in (4.1)).

For model (4.1), an estimator $\hat{f}_{n,h}(.)$ is derived as in (4.2). In the light of (4.6), the asymptotic mean squared errors of the estimator $\hat{f}_{n,h}(.)$ is

$$\mathrm{MSE}_n(h) = \frac{h^4}{4}\sigma_0^2 \int \{\ddot{f}(x)\}^2 p(x)w(x)\mathrm{d}x + \frac{1}{nh}\int \sigma^2(x)w(x)\mathrm{d}x \int K^2(u)\mathrm{d}u.$$

To minimize $\mathrm{MSE}_n(h)$, we consider three cases: (i) use $h = h_n \approx 0$ when the noise is small enough (i.e. $\sigma^2(x)$ is small enough) such that the second term of the RHS of the above expression can be ignored; (ii) use $h = h_n = \infty$ when the model is linear (i.e. $\ddot{f}(x) \equiv 0$); (iii) use $h = h_n \in (0, \infty)$ when the model is nonlinear and stochastic, where h_n is given as in (4.7). However, in practice, we always choose h in a properly specified interval $[h_l,\ h_u] \subset (0, \infty)$. Based on the above observation, we calibrate model (4.1) as *operationally deterministic* if $\mathrm{MSE}_n(h)$ is monotonically increasing as a function of h over $[h_l,\ h_u]$.

Suppose we use $h = \hat{h}_n$ given as in (4.5) in estimating (4.2), i.e., $\hat{h}_n = \hat{h}_{n_1}(n_1/n)^{1/5}$, where $\hat{h}_{n_1}$ is the minimizer of $\mathrm{ECV}_{n_1}(h)$ defined in (4.4). Similar to (4.6), it can be proved that

$$\mathrm{ECV}_{n_1}(h) - \int \sigma^2(x)w(x)\mathrm{d}x \sim \mathrm{MSE}_{n_1}(h),$$

uniformly for $h \in H_{n_1}$. Since $\hat{h}_n$ is a consistent estimate of h_n (cf. (4.8)), a *small* value of $\hat{h}_n$ will indicate that model (4.1) is operationally deterministic.

Of course, it remains to decide how small is *small* in this context, for the purpose of which Yao and Tong (1995b) has suggested the following bootstrap test.

Bootstrap test:

1. For the given data $\{(X_t, Y_t), 1 \le t \le n\}$, choose $H_n = [a_l s_n n^{-1/5}, a_u s_n n^{-1/5}]$, where s_n is the sample standard deviation of $\{X_t\}$, and $0 < a_l < a_u < \infty$ are constant. Obtain the estimate $\hat{h}_n$ by (4.5) . Specify an interval $[h_l,\ h_u]$ which contains $\hat{h}_n$ as an inner point.

2. Estimate f using (4.2) with $h = \hat{h}_n$, and calculate the residuals $\hat{\epsilon}_t = Y_t - \hat{f}_{n,\hat{h}_n}(X_t)$ for $t = 1, \ldots, n$.

3. Draw n samples $\{\epsilon_t^*,\ t = 1, \ldots, n\}$ from the residuals $\{\hat{\epsilon}_t\}$ using the standard bootstrap method, and form the bootstrap sample $\{(X_t, Y_t^*),\ 1 \le t \le n\}$ with

$$Y_t^* = \hat{f}_{n,\hat{h}_n}(X_t) + \epsilon_t^*.$$

4. Obtain an estimate $\hat{h}_n^*$ from the sample $\{(X_t, Y_t^*),\ 1 \le t \le n\}$ using the same method as in Step 1 with $[h_l,\ h_u]$ instead of H_n.

5. Repeat Steps 3 and 4 N times, and count the frequency of occurrence of the event that $\hat{h}_n^* \le \hat{h}_n$. The relative frequency α (= frequency/N) is taken as the evidence for the model (4.1) being operationally deterministic.

Remark 1. If $\hat{h}_n$ in Step 1 is very small (i.e. $\hat{h}_n \approx a_l s_n n^{-1/5}$), the search for $\hat{h}_n^*$ around $\hat{h}_n$ in Step 4 should be conducted on finer grids.

Remark 2. In the above test, values of α near to 1 provide evidence of the system being operationally deterministic; values of α around 0.5 provide evidence of the system being nonlinear and stochastic (i.e. $\ddot{f}(x) \not\equiv 0$).

Remark 3. In the above test, small values of α may be taken to indicate that the model is linear, or simply 'white noise' (i.e. $f(x) \equiv 0$).

Remark 4. In order to provide further evidence of the system being operationally deterministic, it might be worth exploring the following alternative. In Step 5 above, we count instead the frequency of occurrence of the event that $\hat{h}_n^*$ is the smallest value in the interval $[h_l,\ h_u]$ without incurring a computation overflow.

Remark 5. The above bootstrap test can be defined in terms of any reasonable data-driven bandwidth selectors.

Remark 6. The above test was not formulated in the standard setup of testing statistical hypotheses. The α-value can not be regarded as either a significance level or the power of the test.

To illustrate the above test, we report simulation studies for two examples as follows. We always set $n = 500$ for the sample size, $n_1 = 350$ for the estimation of $\hat{h}_n$, $N = 70$ for the number of bootstrap replications. We let $a_l = 0.003$ and $a_u = 3$, and search for $\hat{h}_n$ among 50 grids in H_n defined in Step 1. We choose that $h_l = \hat{h}/2$ or $\hat{h}/3$, and $h_u = \hat{h} + a$ in Step 4, where $a > 0$ is a constant. For each model, we carry out 50 repetitions of Monte Carlo experiments. We always choose $K(.)$ being the Gaussian kernel, and $w(.)$ the indicator function of the 90% inner samples.

Example 1. Let

$$Y_t = 0.246 Y_{t-1}(16 - Y_{t-1}) + \sigma \epsilon_t, \tag{4.9}$$

Table 1. The bootstrap tests for model (4.9)

σ	α-value	Mean($\hat{h}_n$)	Variance($\hat{h}_n$)	Mean($\hat{h}_n^*$)	Variance($\hat{h}_n^*$)
0.07	0.618	0.1838	0.0022	0.1719	0.0013
0.04	0.640	0.1568	0.0010	0.1396	0.0007
0.01	0.622	0.0870	0.0000	0.0819	0.0001
0.005	0.962	0.0870	0.0000	0.0660	0.0001

Table 2. The bootstrap tests for the modelling of $Y_t = X_{t+m}$ on X_t, where X_t is determined by (4.10)

m	α-value	Mean($\hat{h}_n$)	Variance($\hat{h}_n$)	Mean($\hat{h}_n^*$)	Variance($\hat{h}_n^*$)
1	1.000	0.1032	0.0000	0.0624	0.0000
3	1.000	0.1032	0.0000	0.0640	0.0000
5	0.938	0.1032	0.0000	0.0836	0.0002
7	0.171	1.4716	3.8387	1.7208	3.5506
9	0.353	3.1152	3.0196	3.4750	3.0845
11	0.337	3.2297	3.1245	3.7506	3.4099

where $\sigma > 0$ is a constant, and ϵ_t, $t \geq 1$, are independent random variables with the same distribution as the random variable 0.5η, 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 almost standard normal. However, it has a bounded support $[-12,\ 12]$. The simulation has been carried out for the cases with σ equal to four different values between 0.07 and 0.005. The average α-values in 50 repetitions of Monte Carlo experiments are reported in Table 1, which show that the bootstrap test has no difficulties in identifying the model being nonlinear and stochastic when $\sigma \geq 0.01$. But when $\sigma = 0.005$, the test shows that we could treat (4.9) as a deterministic model. The means and variances of $\hat{h}_n$ in the 50 repetitions of Monte Carlo experiments, together with their bootstrap counterparts (in 3500 ($= 50 \times 70$) replications), are also included in the table.

Example 2. For the transformed standard logistic model (with coefficient 4)

$$X_{t+1} = 0.25X_t(16 - X_t), \tag{4.10}$$

we apply the test to the model of $Y_t = X_{t+m}$ on X_t, for $m = 1, 3, \ldots, 11$. The results are reported in Table 2. We can see that the bootstrap test has no difficulties in confirming that we can model X_{t+m} as a deterministic function of X_t for $m \leq 5$. However, for $m \geq 7$ the α-values are considerably smaller than 0.5, which shows that now it will be difficult to model X_{t+m} as a deterministic function of X_t with the given data. Further, a nonlinear stochastic model like (4.1) with $\ddot{f}(x) \not\equiv 0$ is not suitable for those cases either (cf. Remarks 2,3).

BIBLIOGRAPHY

Altman N.S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, **85**, 749-759.

Casdagli M. (1992). Chaos and deterministic versus stochastic non-linear modelling. *J. Roy. Statist. Soc.* **B**, **54**, 303-328.

Chu C.K. and J.S. Marron (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, **19**, 1906-1918.

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

Dechert W.D. and R. Gencay (1990). Estimating Lyapunov exponents with multilayer feedforward network learning. Technical Report, Department of Economics, University of Houston, USA.

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

Fan J., Q. Yao and H. Tong (1993). Estimating measures of sensitivity of initial values to nonlinear stochastic systems with chaos. Technical Report, Univ. of Kent.

Farmer J.D. and J.J. Sidorowich (1987). Predicting chaotic time series. *Phys. Rev. Lett.*, **59**, 845-848.

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

Härdle W. and P. Vieu (1992). Kernel regression smoothing of time series. *J. Time Series Anal.*, **13**, 209-232.

Hart J.D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc.* **B**, **53**, 173-187.

Hart J.D. (1994). Automated kernel smoothing of dependent data by using time series crossing-validation. *J. Roy. Statist. Soc.* **B**, **56**, 529-542.

Herzel H., W. Ebeling and T. Schumeister (1987). Nonuniform chaotic dynamics and effects of noise in biochemical system. *Z. Naturforsch.*, **42**, 136-142.

Lu Z.Q. (1994). Estimating Lyapunov exponents in chaotic time series with locally weighted regression. Ph.D dissertation, University of North Carolina at Chapel Hill.

Marron J.S. (1987). Partitioned cross-validation. *Econometric Rev.*, **6**, 271-284.

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

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

Rosenblatt M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis II* (Edited by P.R. Krishnaiah), 25-31. Academic Press, New York.

Sugihara G. and R.M. May (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement errors in time series. *Nature*, **344**, 734-741.

Tong H. (1995). A personal overview of nonlinear time series analysis from a chaos perspective (with discussion). *Scand. J. Statist.* (To appear).

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

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

Yao Q. and H. Tong (1994b). On prediction and chaos in stochastic systems. *Phil. Trans. Roy. Soc. (Lond.)* **A**, 357-369.

Yao Q. and H. Tong (1995a). Asymmetric least squares regression estimation: a nonparametric approach. *J. Nonparametric Statist.* (To appear.)

Yao Q. and H. Tong (1995b). A bandwidth selector and an informal test for operational determinism. Technical Report, Univ. of Kent.