ESTIMATING MEASURES OF SENSITIVITY OF INITIAL VALUES TO NONLINEAR STOCHASTIC SYSTEMS WITH CHAOS

by

J. Fan Q. Yao H. Tong

November 1993

Mimeo Series #2312

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

Estimating Measures of Sensitivity of Initial Values to Nonlinear Stochastic Systems with Chaos *

Jianqing Fan Qiwei Yao Howell Tong

November 30, 1993

Two measures of sensitivity to initial conditions in nonlinear time series are proposed. The notions give some insight into the relationship between the Fisher information in statistical estimation and initial-value sensitivity in dynamical systems. By using the locally polynomial regression, we develop nonparametric estimates for a conditional density function, its square root and its partial derivatives. The proposed procedures are innovative and of interests in their own right. They are also used to estimate the sensitive measures. The asymptotic normality of the proposed estimators have been proved. We also propose a simple and intuitively appealing method for choosing the bandwidths. Two simulated examples are used as illustrations.

KEY WORDS: nonlinear time series, chaos, Lyapunov exponent, sensitivity to initial values, locally polynomial regression, estimation of conditional density.

SHORT TITLE: Estimating Measures of Sensitivity.

*Jianqing Fan is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. His work was partially supported by NSF Grants DMS-9203135 and an NSF Postdoctoral Fellowship. Qiwei Yao is Lecturer and Howell Tong is Professor, Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, UK. They acknowledge partial support of the Science and Engineering Research Council (UK).

1 Introduction

A significant feature of nonlinear systems exhibiting chaos is that a small perturbation in the initial condition could possibly lead to a considerable divergence of the state of the system in the short or the medium term. In a deterministic dynamic system, this phenomenon is usually characterized by the well-known Lyapunov exponents (cf. Eckmann and Ruelle 1985). However, in a stochastic system, it has not been explored systematically. For example, most of the nonlinear time series literature has concentrated on the distributional properties, which are completely determined by the random law in the system (cf. Tong 1990 and references therewith). The possible divergence caused by the perturbation in the initial value is largely neglected. This may well be irrelevant for linear systems but is certainly not the case for nonlinear systems, because the divergence through nonlinear dynamics could be considerable in the time evolution. In the context of nonlinear prediction, the effect of the perturbation in the initial value was illustrated by Yao and Tong (1994).

The goal of this paper is to define two measures on the sensitivity of a nonlinear system to its initial values, as a formalization of a general idea which has already had a significant impact on statistical applications (especially, in nonlinear time series) and the study of noisy chaos (cf. Yao and Tong 1993). The notion of sensitivity measures and dimensionality of attractors of nonlinear time series has gained increasingly attention. See, for example, the recent development by Eckmann and Ruelle (1985), Nychka *et al.* (1992), Smith (1992) and Wolff (1992), Hall and Wolff (1993). Further, we develop some estimates for these measures without assuming any specific form of the model. We measure the sensitivity in terms of the discrepancy of the conditional distributions of the state variables given two different but nearby initial values. The adopted measures are the mutual information based on the Kullback-Leibler information, and the L^2-distance. The locally polynomial regression method (cf. Fan 1992, Fan *et al.* 1993, and Ruppert and Wand, 1994) is adapted in order to estimate the conditional density function, its square root and the sensitivity measures. The size of local neighborhood (i.e. bandwidth) is objectively determined by data via a Residual Square Criterion (RSC) proposed in Fan and Gijbels (1993) together with a plug-in rule (See e.g. Jones, Marron and Sheather, 1993).

The plan of the paper is as follows. In Section 2, we first summarize the idea of the Lyapunov exponents of a deterministic system. Then we derive two sensitivity measures for a stochastic system. Section 3 presents the nonparametric estimators for conditional density functions and the sensitive measures, which are constructed by using the locally polynomial regression. Some methods for bandwidth selection are also suggested. Section 4 reports the simulation results. Some asymptotic results are stated in Section 5. All mathematical proofs are relegated to the appendix.

2 Measures of the sensitivity of a stochastic system

2.1 Lyapunov exponents for a deterministic system

To highlight the essential idea of how the Lyapunov exponents can be used to monitor the sensitivity of a deterministic system on its initial conditions, we consider a one-dimensional discrete-time deterministic system generated by the dynamical equation $Y_t = f(Y_{t-1})$ for $t \geq 1$. Let $\{Y_t(x),\ t \geq 0\}$ denote the trajectory starting at $Y_0 = x \in R$, and x and $x + \delta$ be two nearby initial values. Then, after m iterations

$$Y_m(x+\delta) - Y_m(x) = f^{(m)}(x+\delta) - f^{(m)}(x) \approx \delta \frac{\mathrm{d}}{\mathrm{d}x} f^{(m)}(x) = \delta \prod_{i=1}^{m} \dot{f}\{Y_i(x)\},$$

where $f^{(m)}$ denotes the m-fold composition of f, and $\dot{f}$ denotes the derivative of f. The (local) Lyapunov exponent (at initial value x) is defined as

$$\kappa(x) = \lim_{m\to\infty} \frac{1}{m} \log |\frac{\mathrm{d}}{\mathrm{d}x} f^{(m)}(x)| = \lim_{m\to\infty} \frac{1}{m} \sum_{i=0}^{m-1} \log |\dot{f}\{Y_i(x)\}|,$$

when the limit exists. Hence, we have the approximation

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

which entails that two trajectories with nearby initial values around x could diverge at an exponential rate if $\kappa(x) > 0$. Obviously, the sensitivity of the system is fully monitored by $\kappa(.)$, which is a function of the derivatives of the map $f(\cdot)$.

2.2 Measures of sensitivity for a stochastic system

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

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

for $t \geq 1$, 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 $\mathrm{E}(\, e_t \,|\, X_0, \ldots, X_{t-1}\,) = 0$. If the noise is additive, (2.1) can be written as

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

which includes the nonlinear autoregressive model as a special case. Specifically, suppose that $\{Y_t,\ -\infty < t < \infty\}$ is a one-dimensional strictly stationary time series, which is d-dependent $(d \geq 1)$ in the sense that given $\{Y_i,\ i \leq t\}$, the conditional distribution of Y_{t+1} depends on $\{Y_i,\ i \leq t\}$ only through X_t, where $X_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+1})^T$. Let $f(x) = E(Y_1 | X_0 = x)$. Then Y_t can be expressed as

$$Y_t \;=\; f(X_{t-1}) \;+\; \epsilon_t, \tag{2.3}$$

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 (2.2) holds.

To apply the key ingredient in Section 2.1 to a stochastic system, we investigate how sensitively the conditional expectation $F_m(x) \equiv \mathrm{E}(\, X_m \,|\, X_0 = x\,)$ depends on x. Formally, we can define an index in a similar way as $\kappa(.)$ (cf. Yao and Tong 1994). However, that approach is not very appropriate. Because of the accumulation of noise through the time evolution, the system seems unlikely to have a strong memory of its initial value after a long time, i. e. $F_m(x)$ would be nearly a constant when m is sufficiently large. This suggests that asymptotics are unlikely to yield a practically useful characteristic exponent, unless we are prepared to entertain the assumption that the different trajectories have the same realization of the random noise (cf. Nychka *et al.*1992). Practically, it seems to us that we should seek some indices which capture the divergence caused by a small shift in the initial value in the short or the medium term. Yao and Tong (1994) used the derivative of $F_m(x)$ to monitor the sensitivity of the conditional mean to the initial value, which also played an important role in m-step pointwise prediction in nonlinear time series.

A more informative way is to consider the global deviation of the conditional distribution of X_m given X_0, which has obviously a wider impact on statistical applications as well as the study of noisy chaos (cf. Yao and Tong 1993). To simplify our discussion, we suppose that the system variables as given in (2.1) are bounded. Let $g_m(y|x)$ denote the conditional density of X_m given $X_0 = x$, which is smooth in both x and y. Let x and $x + \delta \in R^d$ be two nearby initial values. There are quite a few measures available for the discrepancy of two densities. In this paper, we adopt the following two indices. The L_2-distance is simply 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

$$D_m(x;\delta) = \delta^T I_{1,m}(x)\delta + o(||\delta||^2), \tag{2.4}$$

where

$$I_{1,m}(x) = \int \dot{g}_m(y|x)\dot{g}_m^T(y|x)\mathrm{d}y, \tag{2.5}$$

$\dot{g}_m(y|x)$ denotes $\partial g_m(y|x)/\partial x$, and $\dot{g}_m^T(y|x)$ denotes its transpose. We also consider the (negative) 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

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

where

$$I_{2,m}(x) = \int \dot{g}_m(y|x)\dot{g}_m^T(y|x)/g_m(y|x)\mathrm{d}y, \tag{2.7}$$

(cf. §2.6 of Kullback 1967). If we treat the initial value x as a parameter vector of the distribution, $I_{2,m}(x)$ 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, the more sensitively the distribution depends on the initial condition.

2.3 Relation with deterministic system

For the additive system (2.2), with small noise (i.e. the variance of noise is small), $I_{1,m}(x)$ and $I_{2,m}(x)$ are dictated by some functions of the derivatives of the map $F(.)$, a feature which is very similar to the deterministic case. To see this, let us suppose that e_t is independent of $\{X_{t-k}$, for all $k \geq 1\}$. For all $t \geq 1$, let e_t have a smooth density function $\frac{1}{\sigma_0} g(\frac{\cdot}{\sigma_0})$, where $g(.)$ has a bounded support, and $\int xg(x)dx = 0$, $\int xx^T g(x)dx = \Sigma$. The assumption of the bounded support is not essential and is introduced only for the sake of brevity of mathematical derivation. It can be removed at the expense of a lengthier derivation.

Under the above assumption, it can be proved that as $\sigma_0 \to 0$,

$$\begin{aligned} g_m(y|x) &= \int g_{m-1}(y|\sigma_0 s + F(x))g(s)\mathrm{d}s \\ &= g_{m-1}(y|F(x)) + \frac{\sigma_0^2}{2}\mathrm{tr}\{\ddot{g}_{m-1}(y|F(x))\Sigma\}\{1+o(1)\} \end{aligned} \tag{2.8}$$

$$\begin{aligned} \dot{g}_m(y|x) &= \dot{F}(x)\dot{g}_{m-1}(y|F(x))\{1+o(1)\} \\ &= -\left(\prod_{i=0}^{m-1}\dot{F}\{F^{(i)}(x)\}\right)\frac{1}{\sigma_0^2}\dot{g}\left(\frac{y-F^{(m)}(x)}{\sigma_0}\right)\{1+o(1)\}, \end{aligned} \tag{2.9}$$

where $\ddot{g}_m(y|x)$ denotes $\partial^2\{g_m(y|x)\}/(\partial x \partial x^T)$, and $F^{(i)}(x)$ denotes the i-fold composition of F starting with $F^{(0)}(x) \equiv x$. It follows from (2.5), (2.7), (2.8) and (2.9) that

$$I_{1,m}(x) = \frac{1}{\sigma_0^3}\left(\prod_{i=0}^{m-1}\dot{F}\{F^{(i)}(x)\}\right) I_{1,0}\left(\prod_{i=0}^{m-1}\dot{F}\{F^{(i)}(x)\}\right)^T\{1+o(1)\}, \tag{2.10}$$

$$I_{2,m}(x) = \frac{1}{\sigma_0^2}\left(\prod_{i=0}^{m-1}\dot{F}\{F^{(i)}(x)\}\right) I_{2,0}\left(\prod_{i=0}^{m-1}\dot{F}\{F^{(i)}(x)\}\right)^T\{1+o(1)\}, \tag{2.11}$$

where $I_{1,0} = \int \dot{g}(y)\dot{g}^T(y)\mathrm{d}y$ and $I_{2,0} = \int \dot{g}(y)\dot{g}^T(y)/g(y)\mathrm{d}y$ which are independent of x. Thus, for model (2.2) with a small additive noise, we can monitor a profile of $I_{i,m}(x)$ $(i = 1, 2)$ by a functional of the derivatives of $F(.)$, which plays a similar role as the Lyapunov exponent and can be estimated by using the locally polynomial regression method (cf. Fan et. al 1993, Yao and Tong 1994). However, for the general model (2.1) or for model (2.2) with considerable noise, instead of going throuhg the functional we have to estimate $I_{1,m}(x)$ and $I_{2,m}(x)$ directly for obvious reasons. Note that σ_0 appears in denominator in both (2.10) and (2.11), which implies that the increase of stochastic noise will reduce the sensitivity of the system to its initial values.

3 Estimating the measures of divergence

When the system is high-dimensional (i.e. $d \geq 1$), the task of estimating the divergence measures is quite horrendous. Therefore, we consider the divergence in the marginal, rather than the joint, conditional distributions of X_m given $X_0 = x$. It is also of practical interests to concentrate on the divergence in one particular component of the system, e. g. the first component as in the time series model (2.3). Thus, our task can be abstractly stated as follows.

For a $(d+1)$-dimensional random vector (Y, X), where Y is a scalar, X is d-dimensional, let $g(y|x)$ be the conditional density of Y given X and $\dot{g}(y|x)$ denote $\frac{\partial}{\partial x} g(y|x)$. Of interest is to estimate the functions

$$I_1(x) = \int \dot{g}(y|x)\dot{g}^T(y|x)\mathrm{d}y, \tag{3.1}$$

and

$$I_2(x) = \int \dot{g}(y|x)\dot{g}^T(y|x)/g(y|x)\mathrm{d}y, \tag{3.2}$$

based on a sequence of data $(Y_1, X_1), \cdots, (Y_n, X_n)$. Specifically, for univariate time series data $\{x_1, \cdots, x_n\}$, by taking

$$X_i = (x_i, \cdots, x_{i-d})^T, \text{ and } Y_i = x_{i+m},$$

the task reduces to estimating the m-step divergence measures $I_{1,m}(x)$ and $I_{2,m}(x)$ defined respectively in (2.5) and (2.7).

The building blocks for estimating $I_1(x)$ and $I_2(x)$ are $g(y|x)$ and $\dot{g}(y|x)$. This forms the subject of Section 3.1. Let $q(x, y)$ denote $\sqrt{g}(y|x)$. Note that

$$I_2(x) = 4 \int \dot{q}(x,y)\dot{q}(x,y)^T \mathrm{d}y. \tag{3.3}$$

An estimator for $\dot{q}(x, y)$ will also be proposed in Section 3.2.

3.1 Estimating conditional density and its derivative

Estimating the conditional density and its derivatives can be regarded as a nonparametric regression problem. To make this connection, note that

$$E\left(K_{h_2}(Y-y)|X=x\right) \approx g(y|x), \text{as } h_2 \to 0, \tag{3.4}$$

where K is a nonnegative density function and hereafter we always denote $K_h(z) = K(z/h)/h$. The left hand side of (3.4) can be regarded as the regression function of the data $K_{h_2}(Y_i - y)$ on $\{X_i\}$. Recent nonparametric regression theory (see Fan 1992, Fan *et al.* 1993, and Ruppert and Wand 1994) suggests that we use a local polynomial regression to estimate $g(y|x)$ and $\dot{g}(y|x)$. For estimating the first derivative, local quadraticity is preferable (see Fan and Gijbels, 1993). By Taylor's expansion about $x = (x_1, \ldots, x_d)^T \in R^d$, we have

$$\begin{aligned} E\left(K_{h_2}(Y-y)|X=z\right) &\approx g(y|z) \\ &\approx g(y|x) + \dot{g}(y|x)^T(z-x) + \frac{1}{2}(z-x)^T\ddot{g}(y|x)(z-x) \\ &\equiv \beta_0 + \beta_1^T(z-x) + \beta_2^T \mathrm{vec}\{(z-x)\otimes(z-x)^T\}, \end{aligned}$$

where $\ddot{g}(y|x)$ is the Hessian matrix of $g(y|x)$ with respect to x, $\otimes$ denotes the Kronecker product of matrices, $\mathrm{vec}(A) = (a_{11}, a_{22}, \ldots, a_{d,d}, a_{12}, \ldots, a_{1,d}, a_{23}, \ldots, a_{d-1,d})^T \in R^{\frac{d(d+1)}{2}}$ for any $d \times d$ symmetric matrix $A = (a_{ij})$, and

$$\beta_2 = \left(\frac{\partial^2 g(y|x)}{2\partial x_1^2}, \frac{\partial^2 g(y|x)}{2\partial x_2^2}, \ldots, \frac{\partial^2 g(y|x)}{2\partial x_d^2}, \frac{\partial^2 g(y|x)}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^2 g(y|x)}{\partial x_1 \partial x_d}, \frac{\partial^2 g(y|x)}{\partial x_2 \partial x_3}, \ldots, \frac{\partial^2 g(y|x)}{\partial x_{d-1}\partial x_d}\right)^T.$$

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

$$\sum_{i=1}^{n}\left(K_{h_2}(Y_i - y) - \beta_0 - \beta_1^T(X_i - x) - \beta_2^T \mathrm{vec}\{(X_i - x)\otimes(X_i - x)^T\}\right)^2 W_{h_1}(X_i - x), \quad (3.5)$$

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 respectively $g(y|x)$ and $\dot{g}(y|x)$, namely,

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

The least squares theory provides the solution:

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1{}^T, \hat{\beta}_2{}^T)^T = (X^T W X)^{-1} X^T W Y, \quad (3.6)$$

where X is the design-matrix of the least-squares problem (3.5), $W = \mathrm{diag}(W_{h_1}(X_1 - x), \cdots, W_{h_1}(X_n - x))$, and $Y = (K_{h_2}(Y_1 - y), \cdots, K_{h_2}(Y_n - y))^T$.

For simplicity of the presentation, from now on, we treat only a univariate x, i.e. $d = 1$. For the multivariate case, both the theory and the method also hold but with more complicated notation. For the univariate case, simple algebra establishes that

$$\hat{\beta}_j(x, y) = h_1^{-1} \sum_{i=1}^{n} W_j^n \left(\frac{X_i - x}{h_1} \right) K_{h_2}(Y_i - y), \quad j = 0, 1 \tag{3.7}$$

where

$$W_j^n(t) = e_j^T S_n^{-1} (1, h_1 t, h_1^2 t^2)^T \times W(t) \tag{3.8}$$

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

$$S_n = \begin{pmatrix} s_{n,0} & s_{n,1} & s_{n,2} \\ s_{n,1} & s_{n,2} & s_{n,3} \\ s_{n,2} & s_{n,3} & s_{n,4} \end{pmatrix}, \quad s_{n,j} = \sum_{i=1}^{n} (X_i - x)^j W_{h_1}(X_i - x). \tag{3.9}$$

See Fan and Gijbels (1993) for details. We remark that for a fixed h_2, the problem is a standard nonparametric regression problem. The bandwidth h_1 can be selected by using the 'RSC' (Residual Squares Criterion) or the 'refined' criteria proposed in Fan and Gijbels (1993). See Section 3.3 for more detailed discussions.

3.2 Estimating divergence measures

With the derivative of the conditional density estimated by (3.7), a natural estimator for $I_1(x)$ is

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

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

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

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

$$\hat{I}_1(x) = \frac{1}{h_1^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_1^n \left(\frac{X_i - x}{h_1} \right) W_1^n \left(\frac{X_j - x}{h_1} \right) K_{h_2}^*(Y_i - Y_j). \tag{3.10}$$

Analogously, an estimator for $I_2(x)$ can be defined by

$$\hat{I}_2(x) = \int_{-\infty}^{+\infty} \hat{\beta}_1^2(x,y)/\hat{\beta}_0(x,y)\mathrm{d}y, \tag{3.11}$$

with the usual convention $0/0 = 0$. The above integration is typically finite under some mild conditions. However, the estimator (3.11) can not easily be simplified.

An alternative estimator to $I_2(x)$ originates from (3.3). For given bandwidths h_1 and h_2, define

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

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

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

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

is a natural estimate of $q(x,y) = \sqrt{g(y|x)}$ at $(x,y) = (X_t, Y_t)$. Fitting it into the context of locally quadratic regression, we estimate $q(x,y)$, $\dot{q}(x,y) \equiv \frac{\partial}{\partial x} q(x,y)$, and $\ddot{q}(x,y) \equiv \frac{\partial^2}{\partial x^2} q(x,y)$, 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 minimizer of the function

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

H being a probability density function on R^2. Consequently, we estimate $I_2(x)$ by

$$\check{I}_2(x) = 4 \int \{\hat{\dot{q}}(x,y)\}^2 \mathrm{d}y. \tag{3.12}$$

3.3 Selection of bandwidths

While the quality of curve estimation depends sensitively on the choice of the smoothing parameters h_1 and h_2, no final recommendation has been made in the smoothing community. See Jones, Marron and Sheater (1993) for an overview of the current state of the art. For a non-standard problem such as estimating $I_1(x)$ and $I_2(x)$, this issue is even harder. Nevertheless, in this section, we propose a simple and intuitively appealing method for choosing these smoothing parameters. For simplicity of notation, we treat explicitly the one-dimensional case. We first propose a bandwidth choice for estimating the conditional density and then for estimating

$I_1(x)$ and $I_2(x)$ (by using (3.11)). For estimator (3.12), we have not found a systematic way to search for the smoothing parameters h_1 and h_2.

Bandwidth selection for estimating conditional density. As remarked in Section 3.1, for a given bandwidth h_2, the problem (3.5) is a standard nonparametric problem of regressing $Z_i(y) = K_{h_2}(Y_i - y)$ on X_i. Thus, we could use some bandwidth selection techniques in nonparametric regression. A simple and appealing rule is the RSC proposed in Fan and Gijbels (1993). That rule translates into our specific case as follows. Let $\hat{Z}_i(y)$ be the fitted value of the regression problem (3.5) and define the normalized weighted residual sum of squares by

$$\hat{\sigma}^2(x, y; h_1) = \frac{1}{\mathrm{tr}(W - S_n^{-1}T_n)} \sum_{i=1}^{n} \Big(Z_i(y) - \hat{Z}_i(y)\Big)^2 W_{h_1}(X_i - x), \tag{3.13}$$

$S_n = X^T W X$ and $T_n = X^T W^2 X$. See (3.9) for the 3×3-matrix S_n (and similarly T_n) except that $W_{h_1}(X_i - x)$ is replaced by $W_{h_1}^2(X_i - x)$). Define

$$RSC(x, y; h_1) = \sigma^2(x, y; h_1)(1 + 3V_n(x; h_1)), \tag{3.14}$$

where $V_n(x; h_1)$ is the first diagonal element of the matrix $S_n^{-1}T_nS_n^{-1}$. Here, RSC estimates, in a sense, the mean squared error at the point x.

For a given h_2 and y, the proposed bandwidth selection for the estimation of the derivative $\frac{\partial g(y|x)}{\partial x}$ by using (3.5) is

$$\hat{h}_1(y) = adj\ \mathrm{argmin}_h \int RSC(x, y; h)dx, \tag{3.15}$$

where the integration of x is conducted in a region where the curve has to be estimated. Here, the adjusted constant adj depends only on the kernel function W, and is used to adjust the selected bandwidth so that it converges to the theoretical optimal one. From Table 1 of Fan and Gijbels (1993), $adj = 0.7643$ for the Epanechnikov kernel $W(x) = 0.75(1 - x^2)_+$ and $adj = 0.8403$ for the Gaussian kernel $W(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.

The proposed bandwidth (3.15) depends on y. If a constant suffices, we would select

$$\check{h}_1 = adj\ \mathrm{argmin}_h \int\int RSC(x, y; h)dxdy, \tag{3.16}$$

where the integration is conducted on the region of x and y of interest. We could regard (3.16) as minimizing over the average estimated mean squared error for x and y in a region of interest.

Remark that the design matrix in (3.14) does not depend on y. It remains the same for different values of y. This can be used to speed up the computation.

Now, let us turn to h_2. To make the selection rule simple and quick, we use normal referencing rule of Silverman (1986). That rule selects the bandwidth

$$\hat{h}_2 = \left(\frac{8\sqrt{\pi} \int K^2(x)dx}{3(\int x^2 K(x)dx)^2} \right)^{1/5} s_y n^{-1/5}, \tag{3.17}$$

where s_y is the sample standard deviation of the Y-variable. When K is Gaussian kernel, $\hat{h}_2 = 1.06 s_y n^{-1/5}$; for the Epanechnikov kernel, $\hat{h}_2 = 2.34 s_y n^{-1/5}$.

Selecting bandwidths for $I_1(x)$ and $I_2(x)$. For estimating the first derivative, the optimal bandwidth of h_1 is of the order $O(n^{1/7})$ under the assumption that the third derivative with respect to x exists. Thus, for that choice of h_1, there are about $N = O(n^{6/7})$ data points in the neighborhood of $x \pm h_1$. Now, by the theory of Fan (1991) and Hall and Marron (1991), the choice bandwidth h_2 is not very sensitive to $\hat{I}_1(x)$ and $\hat{I}_2(x)$, owing to the integration over y. The choice of order $O(N^{-7/30}) = O(n^{-1/5})$ would be sufficient. To make this order of magnitude meaningful in terms of the scale of y and that of K, we suggest that we use

$$\check{h}_2 = \alpha \left(\frac{8\sqrt{\pi} \int K^2(x)dx}{3(\int x^2 K(x)dx)^2} \right)^{1/5} s_y n^{-1/5}, \tag{3.18}$$

where $\alpha \in [0.5, 1)$ is a specified constant, which makes $\check{h}_2$ smaller than $\hat{h}_2$ in (3.17). The smaller choice of $\check{h}_2$ is natural. For, the integration over y in the definitions of $\hat{I}_1(x)$ and $\hat{I}_2(x)$ reduces the noise level of the estimators, and this allows us to use a smaller bandwidth to reduce the bias in these estimation procedures. The above choice of $\check{h}_1$ and $\check{h}_2$ is also supported by Theorem 5.2. See Remark 5.2 in Section 5 for details.

In the examples that we present in the next section, we will use $\check{h}_2$ in the estimation of I_1 and I_2. Once h_2 is selected, the choice of the bandwidth $\check{h}_1$ is determined by (3.16), which minimizes the average MSE, as explained above.

We do not claim that the any one of the bandwidths (3.15) – (3.18) would be the best choice for all statistical problems that we would encounter. They are basically just quick and simple selection procedures which take the structure and the scale of the data into account. They give us an initial idea as to how much smoothing should be done. However, the theory for estimating I_1 and I_2 is so complicated that there is not much guidance available.

To summarize, we propose that we use the bandwidths $\check{h}_2$ and $\check{h}_1$ to estimate $I_1(x)$ and $I_2(x)$ and use bandwidths $\hat{h}_2$ and $\check{h}_1$ or $\hat{h}_1(y)$ to estimate the conditional densities.

4 Examples

Before proving some asymptotic theorems for the estimators, we illustrate the methods via the following two simulated models.

Example 1. We begin with a simple quadratic model

$$X_t = 0.23X_{t-1}(16 - X_{t-1}) + 0.4\epsilon_t \qquad t \geq 1, \tag{4.1}$$

where ϵ_t, $t \geq 1$, are independent random variables with the same distribution as the random variable η, and η is equal to the sum of 48 independent random variables each uniformly distributed on [-0.25, 0.25]. According to the central limit theorem, ϵ_t can be treated as nearly a standard normal variable. However, it has a bounded support [-12, 12]. Note the bounded support of ϵ_t is necessary for the stationarity of the time series (cf. Chan and Tong 1994). In fact, the skeleton of (4.1) is a transformed logistic map with the coefficient 3.936 (=16×0.246). We have adopted the transformation in order to enlarge the dynamic range of the model. A sample of 1000 is generated form model (4.1). It is well known that the skeleton of this model is chaotic (cf. Hall and Wolff 1993, also see top panel of Figure 1(a)). A typical simulated time series is shown at the bottom panel of Figure 1(a). We consider three cases: $Y_t = X_{t+m}$ for $m = 1, 2, 3$, whose scatter plots are presented in Figures 1(b) – 1(d). As indicated in these figures, there are only a few data at the boundary regions. Thus, we can not expect a very reliable estimate at these regions. As time evolves, the system accumulates more noise, which makes multiple-step prediction harder. Since the signal to noise ratio decreases with the time evolution, we would expect a decrease of the sensitivity to the initial values.

Choosing both kernels K and W to be Gaussian , we have $\hat{h}_2 = 0.98$ by (3.17). By using the RSC-criterion described in Section 3.3, the selected values for $\check{h}_1$ are 0.62 for $m = 1$, 0.70 for $m = 2$, and 0.71 for $m = 3$ respectively (cf. (3.16)). The estimated conditional density functions $\hat{g}_m(y|x) \equiv \hat{\beta}_0(x, y)$ are displayed in Figure 2, which shows that given $X_t = x$, the distribution of X_{t+m} is around $f^{(m)}(x)$, where $f(x) = 0.23x(16 - x)$, and $f^{(m)}$ denotes the

m-th fold composition of f ($m = 1,2,3$). Let $\check{h}_2 = 0.8\hat{h}_2$, we estimate $I_{1,m}(x)$ by using (3.10). The estimated curves are plotted in Figure 3. The sensitivity does vary with the initial value. For example, for $m = 1$, $\hat{I}_1(x)$ attains its minimal value at $x = 8$, monotonically increases when x spreads to both side. From (2.10), we would expect that the sensitivity of the conditional distribution of X_{t+1} to the condition $X_t = x$ is at its weakest at $x = 8$, and increases monotonically when x spreads to both side too. Similar but more complicated conclusions can be made for the cases $m = 2,3$. (Also see Section 4.1 of Yao and Tong 1993, and Example 1 of Yao and Tong 1994.) Figure 4 reports the estimated curves of $I_{2,m}(x)$. We expect that the curves obtained by using (3.11) tend to be somewhat wiggly. This is due to the fact that in (3.11) the estimator $\hat{\beta}_0(x,y)$ appears in the denominator of the integrand. The curves estimated by (3.12) are smoother. However, it remains an open problem as to how to choose the smoothing parameters in using (3.12). For this example, we manually choose bandwidths $(h_1, h_2) = (0.34, 0.68)$, $(0.41, 0.89)$, and $(0.46, 0.85)$ for $m = 1,2$, and 3 respectively. Although the magnitudes of the functions $I_{1,m}(x)$ and $I_{2,m}(x)$ are different, their profiles are somehow similar in the sense that both of them reveal the variation of the strength of the sensitivity of the conditional distribution to its initial value.

(Figures 1 — 4 are about here.)

Example 2. Let us consider the cosine model

$$X_t = 20\cos\left(\frac{\pi X_{t-1}}{10}\right) + \epsilon_t, \tag{4.2}$$

where ϵ_t, $t \geq 1$, are independent standard normal random variables. A sample of 1000 is generated from the above model. The skeleton of this model has a limit point $x = 20$, which converges very fast for a wide range of initial values (see the top panel of Figure 5(a)). The bottom panel of Figure 5(a) indicates a typical simulated data set, which does not show any obvious limit point due to the corruption of the stochastic noise in the model. The scatter plots of X_t against X_{t+m}, for $m = 1,2,3$, are displayed in Figures 5(b) – 5(d). Although the skeleton of this model is not chaotic, Figure 5(b) shows that a small change in X_t, when X_t is around $\pm 5, \pm 15$, will lead a considerable shift in X_{t+1}. On the other hand, X_{t+1} is less sensitive to X_t when X_t is about $0, \pm 10$. Figure 5 also shows that the memory of the system variable on

its initial value decays quickly in the time evolution, which is due to effect of the considerable random noise in the system. For example, it is difficult to trace the trajectory three step ahead (cf. Figure 5(d)).

Choosing both K and W being Gaussian kernel, we have $\hat{h}_2 = 3.65$ by (3.17). By using the RSC-criterion described in last section, the selected values for $\check{h}_1$ are 1.12 for $m = 1$, 1.32 for $m = 2$, and 1.51 for $m = 3$ respectively (cf. (3.16)). The estimated conditional density functions $\hat{g}_m(y|x) \equiv \hat{\beta}_0(x, y)$ are displayed in Figure 6. Let $\check{h}_2 = 0.8\hat{h}_2$, we plot estimated $I_{1,m}(x)$ for $m = 1, 2, 3$ together in Figure 7. It is easy to observe that $I_{1,m}(x)$ decreases sharply as m increases. Further, for fixed m, the sensitivity varies with respect to the initial value, although the variation becames less significant due to the accumulation of considerable random noise when m increases. For example, for $m = 1$, the $\hat{I}_1(x)$ attains its minimal values at $x = 0, \pm 10$. Figure 8 reports the estimated curves of $I_{2,m}(x)$. Similar to Example 1, the curves obtained by using (3.11) are wiggly, and the curves estimated by (3.12) are smoother. In applying (3.12), we use bandwidths $(h_1, h_2) = (0.89, 1.88)$, $(0.94, 2.00)$, $(1.48, 2.14)$ for $m = 1, 2$, and 3 respectively.

(Figure 5 — Figure 8 are about here.)

5 Some asymptotic results

Assume that the sequence of random vectors $\{(X_i, Y_i)\}$ is strictly stationary. Denote by $g(\cdot|\cdot)$ the conditional density of Y_i given X_i and $p(x)$ the marginal density of X_i. Let $\mathcal{F}_i^k$ be the σ-algebra of events generated by the random variables $\{X_j, Y_j, i \le j \le k\}$ and $L_2(\mathcal{F}_i^k)$ denote the collection of all second-order random variables which are $\mathcal{F}_i^k$-measurable. Let

$$\rho(k) = \sup_{U \in L_2(\mathcal{F}_{-\infty}^0), V \in L_2(\mathcal{F}_k^\infty)} \frac{|\mathrm{cov}(U, V)|}{\mathrm{var}^{1/2}(U)\mathrm{var}^{1/2}(V)} \tag{5.1}$$

denote the ρ-mixing coefficient (Kolmogorov and Rozanov, 1960). We first impose some regularity conditions:

(C1) The kernel functions W and K are symmetric and bounded with bounded supports.

(C2) The process $\{X_j, Y_j\}$ is ρ-mixing with $\sum \rho(\ell) < \infty$. Further, assume that there exists a sequence of positive integers $s_n \to \infty$ such that $s_n = o((nh_1h_2)^{1/2})$ and $\{n/(h_1h_2)\}^{1/2}\rho(s_n) \to$

0.

(C3) The function $g(y|x)$ has bounded continuous third order derivatives at point (x,y), and $p(x)$ is continuous at the point x.

(C4) The joint density of the distinct elements of $(X_0, Y_0, X_\ell, Y_\ell)$ $(\ell > 0)$ is bounded by a constant independent of ℓ.

(C5) The bandwidths h_1 and h_2 converge to zero in such a way that $nh_1^3h_2 \to \infty$.

Condition (C1) is imposed for the brevity of proofs, which can be removed at the expense of a longer proof. In particular, the Gaussian kernel is allowed. The assumption on the convergence rate of $\{\rho(\ell)\}$ in (C2) is also for the technical convenience, which is not the weakest possible.

Theorem 5.1. *Under Conditions (C1) — (C5), for* $x \in \{x : p(x) > 0\}$,

$$\sqrt{nh_1h_2}\{\hat{g}(y|x) - g(y|x) - \vartheta_{n,1}\} \xrightarrow{\mathcal{L}} N\left(0,\ \sigma_1^2(x,y)\right), \tag{5.2}$$

$$\sqrt{nh_1^3h_2}\{\hat{\dot{g}}(y|x) - \dot{g}(y|x) - \vartheta_{n,2}\} \xrightarrow{\mathcal{L}} N\left(0,\ \sigma_2^2(x,y)\right). \tag{5.3}$$

Further, $\hat{g}(y|x)$ *and* $\hat{\dot{g}}(y|x)$ *are asymptotically independent in the sense that the random variables on the RHS of (5.2) and (5.3) are joint asymptotic normal with zero covariance. Here,*

$$\vartheta_{n,1} = \frac{1}{2}\mu_K \frac{\partial^2 g(y|x)}{\partial y^2} h_2^2 + o(h_1^3 + h_2^2), \quad \sigma_1^2(x,y) = \frac{g(y|x)\nu_0\nu_K}{p(x)} \frac{\mu_4^2\nu_0 - 2\mu_2\mu_4\nu_2 + \frac{1}{2}\mu_2^2\nu_4}{(\mu_4 - \mu_2^2)^2},$$

$$\vartheta_{n,2} = \frac{\mu_4}{6\mu_2}\frac{\partial^3 g(y|x)}{\partial x^3}h_1^2 + \frac{1}{2}\mu_K \frac{\partial^3 g(y|x)}{\partial x \partial y^2} h_2^2 + o(h_1^2 + h_2^2), \quad \sigma_2^2(x,y) = \frac{g(y|x)\nu_K}{p(x)} \frac{\nu_0\nu_2}{\mu_2^2},$$

and $\mu_K = \int t^2 K(t)\mathrm{d}t$, $\nu_K = \int\{K(t)\}^2\mathrm{d}t$, $\mu_j = \int t^j W(t)\mathrm{d}t$, $\nu_j = \int t^j\{W(t)\}^2\mathrm{d}t$ *(*$j \geq 0$*).*

Remark 5.1. Without the assumption that $W(\cdot)$ is symmetric, the asymptotic biases for $\hat{g}(y|x)$ will be

$$\vartheta_{n,1} = \frac{1}{6}\frac{\partial^3 g(y|x)}{\partial x^3}\frac{\mu_4\mu_3 - \mu_5\mu_2}{\mu_4 - \mu_2^2}h_1^3 + \frac{1}{2}\mu_K\frac{\partial^2 g(y|x)}{\partial y^2}h_2^2 + o(h_1^3 + h_2^2).$$

If our interest is to estimate the conditional density, then the locally linear (instead of the locally quadratic) regression suffices. In that case, the asymptotic normality admits a more symmetric form

$$\sqrt{nh_1h_2}\left(\hat{g}(y|x) - g(y|x) - \frac{h_1^2\mu_2}{2}\frac{\partial^2 g(y|x)}{\partial x^2} - \frac{h_2^2\mu_K}{2}\frac{\partial^2 g(y|x)}{\partial y^2}\right) \xrightarrow{\mathcal{L}} N\left(0, \nu_K\nu_0\frac{g(y|x)}{p(x)}\right),$$

under the assumptions (C1) - (C4) and $nh_1h_2 \to \infty$. Our results and proofs can be readily extended to higher order polynomial regression. We prefer not to present the general theory for the sake of simplicity.

Theorem 5.2. *Under Conditions (C1) — (C5), if* $nh_1^3h_2^2 \to \infty$, *for* $x \in \{x : p(x) > 0\}$, $\sqrt{nh_1^3}\{\hat{I}_1(x) - I_1(x) - \vartheta_n\} \xrightarrow{\mathcal{L}} N(0, \sigma^2)$, *where*

$$\begin{aligned}
\vartheta_n &= h_1^2 \frac{\mu_4}{3\mu_2} \int \left\{ \frac{\partial g(y|x)}{\partial x} \frac{\partial^3 g(y|x)}{\partial x^3} \right\} \mathrm{d}y + h_2^2 \mu_K \int \left\{ \frac{\partial g(y|x)}{\partial x} \frac{\partial^3 g(y|x)}{\partial x \partial y^2} \right\} \mathrm{d}y + o(h_1^2 + h_2^2), \\
\sigma^2 &= \frac{4\nu_2}{\mu_2^2 p(x)} \left[\int \left\{ \frac{\partial g(y|x)}{\partial x} \right\}^2 g(y|x) \mathrm{d}y - \left\{ \int \frac{\partial g(y|x)}{\partial x} g(y|x) \mathrm{d}y \right\}^2 \right].
\end{aligned}$$

Remark 5.2. The choice of h_2 for estimating $I_1(x)$ is not as sensitive as that for estimating the conditional density. In fact, for h_2 in the range that $(nh_1^3)^{-1/4} >> h_2 >> (nh_1^3)^{-1/2}$, the asymptotic bias and variance of $I_1(x)$ remain the approximately the same; i.e. the term $O(h_2^2)$ in ϑ_n becomes negligible. Thus, the optimal choice of bandwidth is $h_1 = cn^{-1/7}$ and $n^{-1/7} >> h_2 >> n^{-2/7}$, where

$$c = \left\{ \frac{27\mu_2^2\sigma^2}{4\mu_4^2} \left(\int \frac{\partial g(y|x)}{\partial x} \frac{\partial^3 g(y|x)}{\partial x^3} \mathrm{d}y \right)^{-2} \right\}^{-1/7}.$$

Appendix — Proofs

Proof of Theorem 5.1. Let $m(x,y) = E\{K_{h_2}(Y_i - y)|X_i = x\}$, $\beta \equiv (m_0(x,y), m_1(x,y), m_2(x,y))^T$ $\equiv (m(x,y), \frac{\partial}{\partial x}m(x,y), \frac{\partial^2}{\partial x^2}m(x,y))^T$, and $H = \mathrm{diag}(1, h_1, h_1^2)$. It follows from (3.6) that

$$H(\hat{\beta} - \beta) = H(X^TWX)^{-1}X^TW(Y - X\beta) = S_n^{*-1}\{(t_{n,0}, t_{n,1}, t_{n,2})^T + (\gamma_{n,0}, \gamma_{n,1}, \gamma_{n,2})^T\}, \quad \text{(A.1)}$$

where

$$\begin{aligned}
t_{n,j} &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i - x}{h_1} \right)^j W_{h_1}(X_i - x) \{K_{h_2}(Y_i - y) - m(X_i, y)\}, \\
\gamma_{n,j} &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i - x}{h_1} \right)^j W_{h_1}(X_i - x) \{m(X_i, y) - m(x,y) - m_1(x,y)(X_i - x) - m_2(x,y)\frac{(X_i - x)^2}{2}\}, \\
s_{n,j}^* &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i - x}{h_1} \right)^j W_{h_1}(X_i - x),
\end{aligned}$$

and S_n^* is a 3×3 matrix with the (i,j)-element s_{i+j-2}^*. Let S, and Σ be 3×3 matrices with the (i,j)-element μ_{i+j-2}, and ν_{i+j-2} respectively, and $\gamma = (\mu_3, \mu_4, \mu_5)^T$. In the sequel, we will establish

(a) $S_n^* \xrightarrow{m.s.} p(x)S$, in the mean square sense.

(b) $h_1^{-3}(\gamma_{n,0}, \gamma_{n,1}, \gamma_{n,2})^T \xrightarrow{m.s.} \frac{1}{6}p(x)\frac{\partial^3 g(y|x)}{\partial x^3}\gamma$, in the mean square sense.

(c) $\sqrt{nh_1h_2}(t_{n,0}, t_{n,1}, t_{n,2}) \xrightarrow{\mathcal{L}} N(0, g(y|x)p(x)\nu_0\nu_K\Sigma)$.

Combining these with (A.1), we have

$$\sqrt{nh_1h_2}\left\{H(\hat{\beta}-\beta) - \frac{1}{6}h_1^3\frac{\partial^3 g(y|x)}{\partial x^3}S^{-1}\gamma\right\} \xrightarrow{\mathcal{L}} N\left(0, p(x)^{-1}g(y|x)\nu_0\nu_K S^{-1}\Sigma S^{-1}\right). \quad (A.2)$$

It follows from Taylor expansion that $m_j(x,y) = \frac{\partial^j g(y|x)}{\partial x^j} + \frac{1}{2}h_2^2\mu_K\frac{\partial^{j+2}g(y|x)}{\partial x^j \partial y^2} + o(h_2^2)$. Using this expansion and considering the marginal distribution of (A.2), we obtain the result.

We are now in a position to establish the conclusions (a) – (c). Conclusions (a) and (b) can be proved by computing the means and the variances of $s_{n,j}^*$ and $\gamma_{n,j}$. We only prove (a). By Taylor's expansion:

$$Es_{n,j}^* = E\left(\frac{X_1 - x}{h_1}\right)^j W_{h_1}(X_1 - x) = p(x)\mu_j + o(1).$$

The variance of $s_{n,j}^*$ can be calculated by using the stationarity and mixing conditions, and is of size $(nh_1)^{-1}$. Since these calculations are similar to those given in the proof of (A.8), we omit the details.

To prove (c), we consider arbitrary linear combinations of $t_{n,j}$ with constant coefficients η_j $(j = 0, 1, 2)$. Let

$$\begin{aligned} Q_n &= \sqrt{nh_1h_2}(\eta_0 t_{n,0} + \eta_1 t_{n,1} + \eta_2 t_{n,2}) \\ &= n^{-1/2}\sum_{i=1}^{n}\sqrt{h_1h_2}D_{h_1}(X_i - x)\{K_{h_2}(Y_i - y) - m(X_i, y)\}, \end{aligned} \quad (A.3)$$

where $D(u) = (\eta_0 + \eta_1 u + \eta_2 u^2)W(u)$. Write $Q_n = n^{-1/2}\sum_{i=0}^{n-1} Z_{n,i}$.

We now employ the small-block and large-block arguments. Following the proof of Masry and Fan (1993), partition the set $\{1, \cdots, n\}$ into $2k+1$ subsets with large blocks of size $r = r_n$

and small block of size $s = s_n$. Put $k = k_n = \lfloor \frac{n}{r_n+s_n} \rfloor$. Define the random variables

$$\eta_j = \sum_{i=j(r+s)}^{j(r+s)+r-1} Z_{n,i}, \quad \xi_j = \sum_{i=j(r+s)+r}^{(j+1)(r+s)-1} Z_{n,i}, \quad \zeta_k = \sum_{i=k(r+s)}^{n-1} Z_{n,i}.$$

Then,

$$Q_n = n^{-1/2}\{\sum_{j=0}^{k-1} \eta_j + \sum_{j=0}^{k-1} \xi_j + \zeta_k\} \equiv n^{-1/2}\{Q'_n + Q''_n + Q'''_n\}.$$

We will show that as $n \to \infty$,

$$\frac{1}{n}E(Q''_n)^2 \to 0, \quad \frac{1}{n}E(Q'''_n)^2 \to 0 \tag{A.4}$$

$$\left| E\left[\exp(itQ'_n)\right] - \prod_{j=0}^{k-1} E\left[\exp(it\eta_j)\right] \right| \to 0 \tag{A.5}$$

$$\frac{1}{n}\sum_{j=0}^{k-1} E\left(\eta_j^2\right) \to \sigma^2, \qquad \frac{1}{n}\sum_{j=0}^{k-1} E\left(\eta_j^2 I\{|\eta_j| \geq \varepsilon\sigma\sqrt{n}\}\right) \to 0 \tag{A.6}$$

for every ε, where $\sigma^2 = p(x)g(y|x)\nu_0\nu_D$ with $\nu_D = \int D^2(u)du$. (A.4) implies that Q''_n and Q'''_n are asymptotically negligible, (A.5) implies that the summands $\{\eta_j\}$ in Q'_n are asymptotically independent, and (A.6) is the standard Lindeberg-Feller conditions for asymptotic normality of Q'_n under independence. Expressions (A.4) – (A.6) entail the following asymptotic normality: $Q_n \xrightarrow{\mathcal{L}} N(0, \sigma^2)$.

We first choose the block sizes. Condition (C2) implies that there exist constants $q_n \to \infty$ such that $q_n s_n = o(\sqrt{nh_1h_2})$ and $q_n\{n/(h_1h_2)\}^{1/2}\rho(s_n) \to 0$. Define the large block size $r_n = \lfloor (nh_1h_2)^{1/2}/q_n \rfloor$. Then, it can easily be shown that

$$s_n/r_n \to 0, \quad r_n/n \to 0, \quad r_n/(nh_1h_2)^{1/2} \to 0, \text{ and } \frac{n}{r_n}\rho(s_n) \to 0. \tag{A.7}$$

We first establish the following approximation:

$$\text{var}(Z_{n,0}) = p(x)g(y|x)\nu_D\nu_0(1 + o(1)), \quad \sum_{\ell=1}^{n-1} |\text{cov}(Z_{n,0}, Z_{n,\ell})| = o(1). \tag{A.8}$$

The first part follows directly from Taylor's expansion:

$$\text{var}(Z_{n,0}) = h_1h_2ED^2_{h_1}(X_1 - x)\{K^2_{h_2}(Y_1 - y) - m^2(X_1, y)\} = p(x)g(y|x)\nu_D\nu_0 + o(1).$$

To prove the second conclusion, by a change of variable and Condition (C4), $|\mathrm{cov}(Z_{n,0}, Z_{n,\ell})| \le c_n$ for some constant sequence $c_n \to 0$. Also, by (5.1), $|\mathrm{cov}(Z_{n,0}, Z_{n,\ell})| \le \mathrm{var}(Z_{n,0})\rho(\ell)$. Let $d_n \to \infty$ such that $c_n d_n \to 0$. Then, we have

$$\sum_{\ell=1}^{n-1} |\mathrm{cov}(Z_{n,0}, Z_{n,\ell})| \le c_n d_n + \mathrm{var}(Z_{n,0}) \sum_{\ell=d_n+1}^{\infty} \rho(\ell) \to 0.$$

We now establish (A.4). First of all, by stationarity and (A.8),

$$\mathrm{var}(\xi_j) = s\mathrm{var}(Z_{n,0}) + 2s\sum_{j=1}^{s-1}(1 - j/s)\mathrm{cov}(Z_{n,0}, Z_{n,j}) = s\sigma^2(1 + o(1)),$$

and

$$E(Q_n'')^2 = \sum_{j=0}^{k-1} \mathrm{var}(\xi_j) + \sum_{\substack{i=0 \\ i\neq j}}^{k-1} \sum_{j=0}^{k-1} \mathrm{cov}(\xi_i, \xi_j) \equiv F_1 + F_2. \tag{A.9}$$

By (A.8), $F_1 = O(ks) = o(n)$. Now, we consider F_2. We first note that with $n_j = j(r+s) + r$,

$$F_2 = \sum_{\substack{i=0 \\ i\neq j}}^{k-1} \sum_{j=0}^{k-1} \sum_{\ell_1=0}^{s-1} \sum_{\ell_2=0}^{s-1} \mathrm{cov}(Z_{n,n_i+\ell_1}, Z_{n,n_j+\ell_2}),$$

but since $i \neq j$, $|n_i - n_j + \ell_1 - \ell_2| \ge r$ so that

$$|F_2| \le 2 \sum_{\ell_1=0}^{n-r-1} \sum_{\ell_2=\ell_1+r}^{n-1} |\mathrm{cov}(Z_{n,\ell_1}, Z_{n,\ell_2})|.$$

By stationarity and (A.8),

$$|F_2| \le 2n \sum_{j=r}^{n-1} |\mathrm{cov}(Z_{n,0}, Z_{n,j})| = o(n).$$

This together with (A.9) proves the first part of (A.4). For the second part of (A.4), using a similar argument together with (A.8), we obtain that

$$\begin{aligned} \frac{1}{n}E(S_n''')^2 &\le \frac{1}{n}(n - k(r+s))\mathrm{var}(Z_{n,0}) + 2\sum_{j=1}^{n-1} |\mathrm{cov}(Z_{n,0}, Z_{n,j})| \\ &\le \frac{r_n + s_n}{n}\sigma^2 + o(1) \to 0. \end{aligned}$$

Equation (A.5) can be proved as follows. Note that η_a is $\mathcal{F}_{i_a}^{j_a}$-measurable with $i_a = a(r+s)+1$ and $j_a = a(r+s)+r$. Hence, applying Volkonskii and Rozanov's Lemma with $V_j = \exp(it\eta_j)$ and using the fact that α-mixing coefficients are bounded by ρ-mixing coefficients: $\alpha(k) \le \rho(k)/4$, we have

$$\left| E\exp(itQ') - \prod_{j=0}^{k-1} E[\exp(it\eta_j)] \right| \le 4k\rho(s_n - 1) \sim 4\frac{n}{r_n}\rho(s_n - 1),$$

which tends to zero by (A.7).

We now show the first part of (A.6). By stationarity and (A.8), $\mathrm{var}(\eta_j) = \mathrm{var}(\eta_0) = r\sigma^2(1+o(1))$. This implies that

$$\frac{1}{n}\sum_{j=0}^{k-1} E(\eta_j^2) = \frac{k_n r_n}{n}\sigma^2(1+o(1)) \sim \frac{r_n}{r_n+s_n}\sigma^2 \to \sigma^2.$$

It remains to establish the second part of (A.6). Using the fact that $D(\cdot)$ is bounded, we have $|Z_{n,j}| \le C(h_1h_2)^{-1/2}$ for some constant C. This and (A.7) entail $\max_{0\le j\le k-1}|\eta_j|/\sqrt{n} \le Cr_n(nh_1h_2)^{-1/2} \to 0$. Hence, when n is large the set $\{|\eta_j| \ge \sigma\varepsilon\sqrt{n}\}$ becomes an empty set, namely the second part of (A.6) holds. □

Proof of Theorem 5.2. We adopt the notation introduced in the proof of Theorem 1. Let $\xi_{n,j}(x,y) = (t_{n,j}+\gamma_{n,j})/h_1$. To prove Theorem 2, we need the following asymptotic results, which will be proved later.

(d) $E\int\{\xi_{n,1}(x,y)\}^2\mathrm{d}y = O\{h_1^4 + (nh_1^3h_2)^{-1}\} = o(h_1^2 + (nh_1)^{-1/2})$,

(e) $\sqrt{nh_1^3}\{\int \xi_{n,1}(x,y)m_1(x,y)\mathrm{d}y - \frac{\mu_4 p(x)}{6}\int \frac{\partial g(y|x)}{\partial x}\frac{\partial^3 g(y|x)}{\partial x^3}dyh_1^2 + o(h_1^2)\} \xrightarrow{\mathcal{L}} N(0,\sigma_0^2)$, where $\sigma_0^2 = \nu_2 p(x)\,[\int\{\dot g(y|x)\}^2 g(y|x)dy - \{\int \dot g(y|x)g(y|x)\mathrm{d}y\}^2]$.

By (A.1), we have that $\hat\beta_1(x,y) - m_1(x,y) = (0,1,0)S_n^{*-1}(\xi_{n,0},\xi_{n,1},\xi_{n,2})^T$. It follows from (a) that

$$\begin{aligned}
&\hat I_1(x) - \int\{m_1(x,y)\}^2\mathrm{d}y \\
=\ &\int\{\hat\beta_1(x,y) - m_1(x,y)\}^2\mathrm{d}y + 2\int m_1(x,y)\{\hat\beta_1(x,y) - m_1(x,y)\}\mathrm{d}y \\
=\ &\left\{\frac{1}{p^2(x)\mu_2^2}\int \xi_{n,1}^2(x,y)\mathrm{d}y + \frac{2}{p(x)\mu_2}\int \xi_{n,1}(x,y)m_1(x,y)\mathrm{d}y\right\}(1+o_p(1)). \qquad \text{(A.10)}
\end{aligned}$$

Since $\int\{m_1(x,y)\}^2\mathrm{d}y = I_1(x) + h_2^2\mu_K\int\left\{\frac{\partial g(y|x)}{\partial x}\frac{\partial^3 g(y|x)}{\partial x\partial y^2}\right\}\mathrm{d}y$, Theorem 5.2 follows immediately from (d), (e), and (A.10).

The proof of (d) is similar to that of (a), and is omitted here [The conclusion (d) is basically the MISE for the derivative estimation]. To prove (e), we define that

$$U(x_1,y_1;x,y) = h_1^{-2}(x_1-x)W_{h_1}(x_1-x)\{K_{h_2}(y_1-y) - m(x,y) - m_1(x,y)(x_1-x) - m_2(x,y)(x_1-x)^2/2\}$$

and that $V(x_1,y_1) = \int U(x_1,y_1;x,y)m_1(x,y)dy$. Then, $\int \xi_{n,1}(x,y)m_1(x,y)\mathrm{d}y = n^{-1}\sum_{i=1}^n V(X_i,Y_i)$. It can be shown via Taylor's expansion that

$$EV(X_1,Y_1) = \frac{p(x)\mu_4}{6}\int\int \frac{\partial g(y|x)}{\partial x}\frac{\partial^3 g(y|x)}{\partial x^3}dyh_1^2 + o(h_1^2)$$

and that

$$EU^*(X_1,Y_1;x,y)U^*(X,Y;x,y+h_2z) = h_1^{-3}h_2^{-1}g(y|x)p(x)\nu_2\int K(u)K(u+z)du(1+o(1)),$$

where $U^*(x_1,y_1;x,y) = h_1^{-2}(x_1-x)W_{h_1}(x_1-x)K_{h_2}(y_1-y)$. Thus,

$$\begin{aligned}
\mathrm{var}(V(X_1,Y_1)) &= EV^2(X_i,Y_i) + O(h_1^4) \\
&= h_2\int\int EU^*(X_1,Y_1;x,y)U^*(X_1,Y_1;x,y+h_2z)m_1(x,y)m_1(x;y+h_2z)dydz \\
&\quad -E[\int h_1^{-2}(X_1-x)W_{h_1}(X_1-x)m(X_1,y)m_1(x,y)dy]^2(1+o(1)) \\
&= h_1^{-3}p(x)\nu_2[\int m_1^2(x,y)g(y|x)dy - \{\int m_1(x,y)g(y|x)dy\}^2](1+o(1)).
\end{aligned}$$

Now, using the big-small block arguments as in the proof of Theorem 5.1, we establish e). □

References

[1] Chan, K.S. and Tong, H. (1994) A note on noisy chaos. *J. R. Statis. Soc.* **B**, 56. To appear.

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

[3] Fan, J. (1991). On the estimation of quadratic functionals. *Annals of Statistics*, **19**, 1273-1294.

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

[5] Fan, J. and Gijbels, I. (1993). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Institute of Statistics Mimeo Series # 2301*

[6] Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. *Institute of Statistics Mimeo Series # 2302*

[7] Hall, P. and Marron, J.S. (1991). Lower bounds for bandwidth selection in density estimation. *Prob. Theory Rel. Fields.*, **90**, 149-173.

[8] Hall, P. and Wolff, C.L. (1993). Properties of invariant distributions and Lyapunov exponents for chaotic logistic maps. *Manuscript.*

[9] Jones, M.C., Marron, J.S. and Sheather, S.J. (1993). Progress in data based bandwidth selection for kernel density estimation. Department of Statistics, University of North Carolina. *Mimeo Series # 2088.*

[10] Masry, E. and Fan, J. (1993). Local polynomial Estimating regression functions for mixing processes. *Manuscript.*

[11] Kolmogorov, A.N. and Rozanov, Yu. A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory Prob. Appl.*, **52**, 204-207.

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

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

[14] Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, to appear.

[15] Silverman, N.B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[16] Smith, R. (1992). Estimating dimension in noisy chaotic time series. *J. R. Statist. Soc. B*, **54**, 329-351.

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

[18] Volkonskii, V.A. and Rozanov, Yu. A. (1959). Some limit theorems for random functions. *Theory Prob. Appl.*, **4**, 178-197.

[19] Wollf, R. (1992). Local Lyapunov exponents: looking closely at chaos. *J. R. Statist. Soc. B*, **54**, 353-371.

[20] Yao, Q. and Tong, H. (1993). On prediction and chaos in stochastic systems. Technical Report, University of Kent, an invited paper to be presented at the Royal Society Discussion meeting on Chaos and Forecasting, March 2-3, 1994.

[21] Yao, Q. and Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *J. R. Statis. Soc.* **B**, **56**. (To appear).

Figure Captions

Figure 1 (a) Plot of the skeleton of model (4.1) and a simulated time series. Top panel is the plot of the skeleton $x_t = 0.23x_{t-1}(16 - x_{t-1})$, which is a chao, and the bottom panel is the plot of t against $X_t - 16$; (b) - (d) are the scatter plots of X_t against X_{t+m} for (b) m=1; (c) m=2; (d) m=3.

Figure 2 The estimated $g_m(y|x)$: the conditional density function of Y_{t+m} given $Y_t = x$ for logistic model (4.1). (a) $m = 1$; (b) $m = 2$; (c) $m = 3$.

Figure 3 The estimated curves of $I_{1,m}(x)$ for logistic model (4.1) by (3.10). (a) $m = 1$; (b) $m = 2$; (c) $m = 3$.

Figure 4 The estimated curves of $I_{2,m}(x)$ for logistic model (4.1). Solid curve — estimated by (3.11); dashed curve — estimated by (3.12). (a) $m = 1$; (b) $m = 2$; (c) $m = 3$.

Figure 5 (a) Plot of the skeleton of model (4.1) and a simulated time series. Top panel is the plot of the skeleton of $x_t = 20\cos(\pi x_{t-1}/10)$, which converges to its limit point very fast, the bottom panel is the plot of t against $X_t - 45$; (b) - (d) are the scatter plots of X_t against X_{t+m} for (b) m=1; (c) m=2; (d) m=3.

Figure 6 The estimated $g_m(y|x)$: the conditional density function of Y_{t+m} given $Y_t = x$ for cosine model (4.2). (a) $m = 1$; (b) $m = 2$; (c) $m = 3$.

Figure 7 The estimated curve of $I_{1,m}(x)$ for cosine model (4.2): Solid curve — $m = 1$; dashed curve — $m = 2$; dotted curve — $m = 3$.

Figure 8 The estimated curves of $I_{2,m}(x)$ for cosine model (4.2). Solid curve — estimated by (3.11); dashed curve — estimated by (3.12). (a) $m = 1$; (b) $m = 2$; (c) $m = 3$.

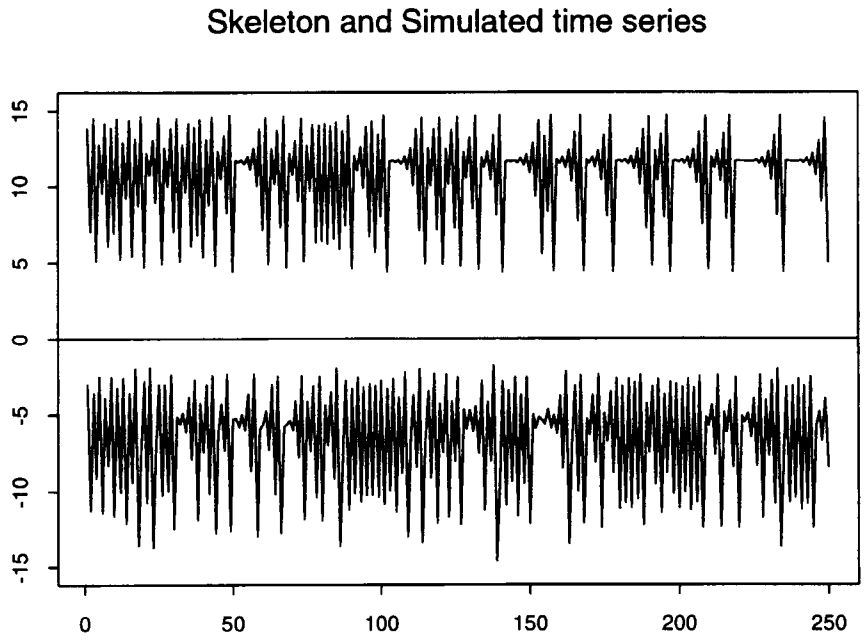

Figure 1(a)

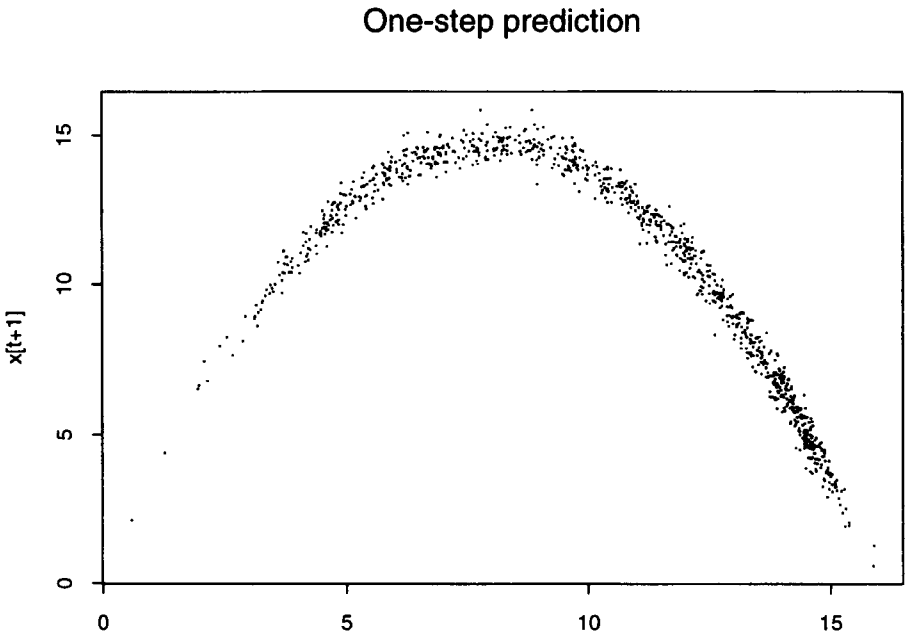

Figure 1(b)

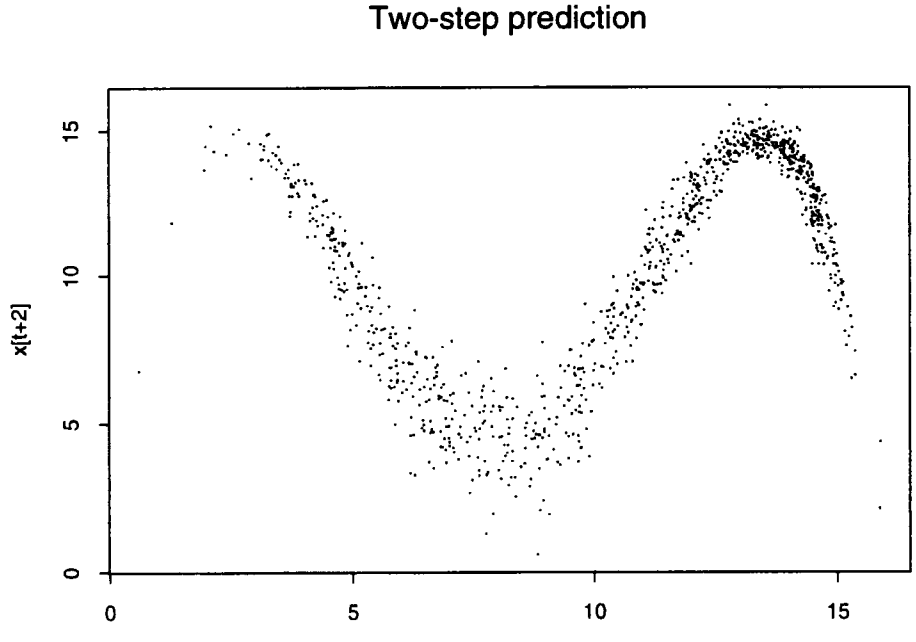

Figure 1(c)

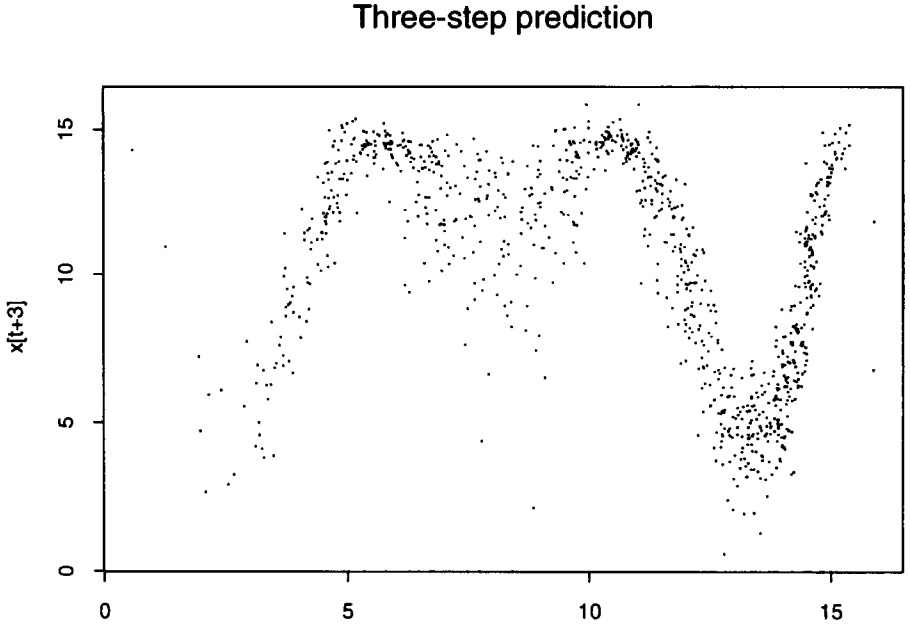

Figure 1(d)

Conditional density of X(t+1) given X(t)

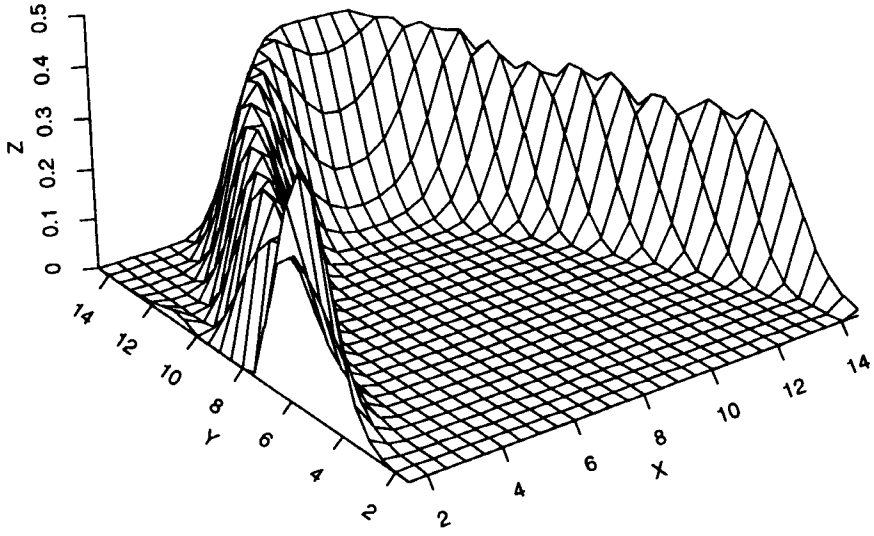

Figure 2(a)

Conditional density of X(t+2) given X(t)

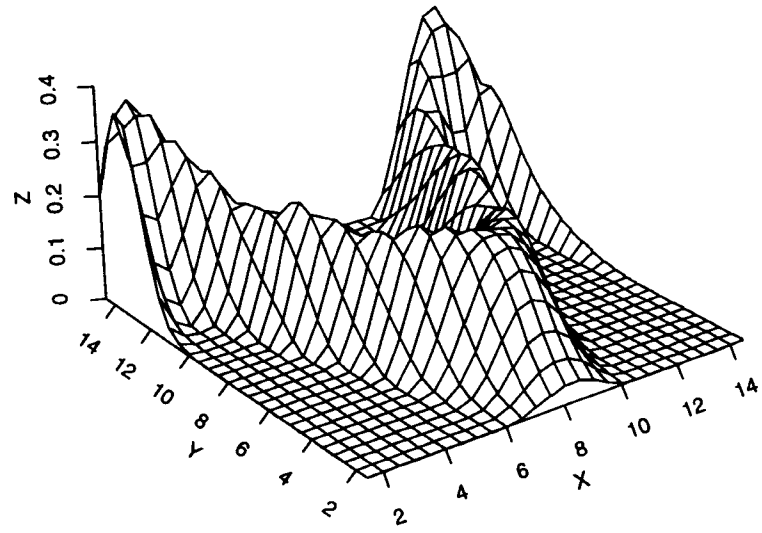

Figure 2(b)

Conditional density of X(t+3) given X(t)

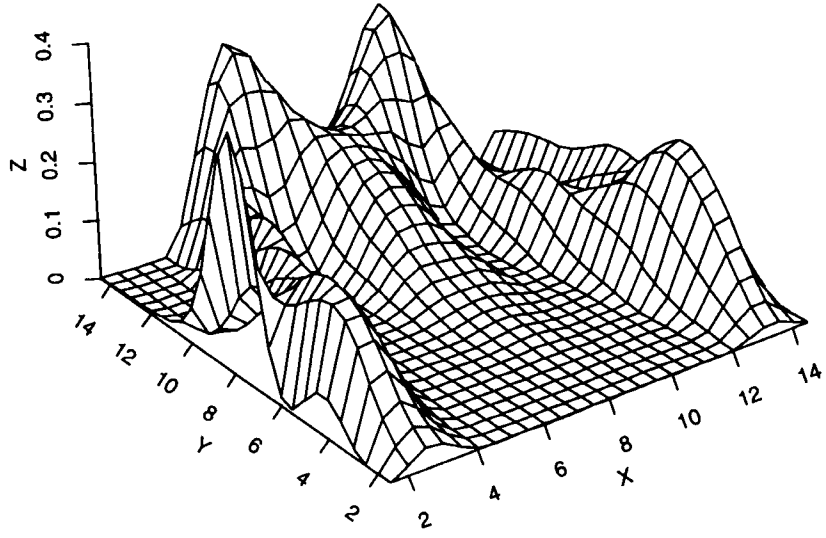

Figure 2(c)

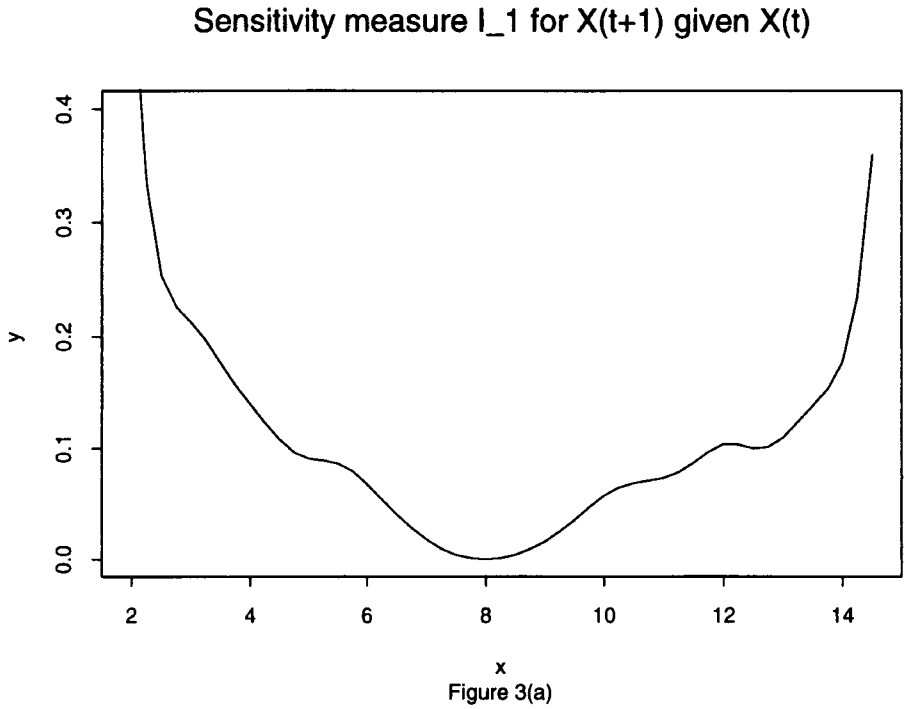

Figure 3(a)

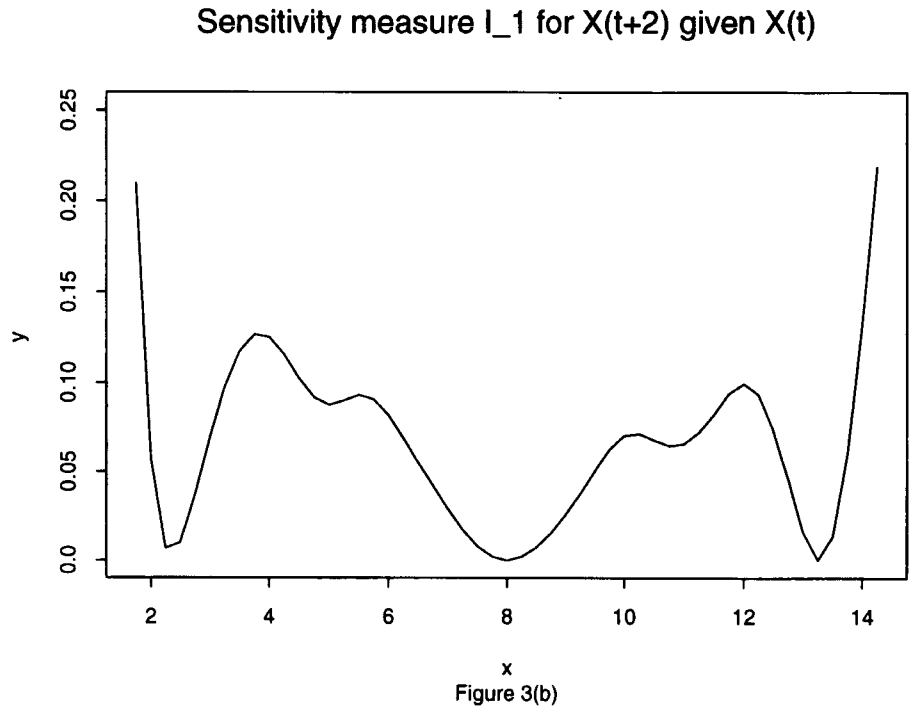

Figure 3(b)

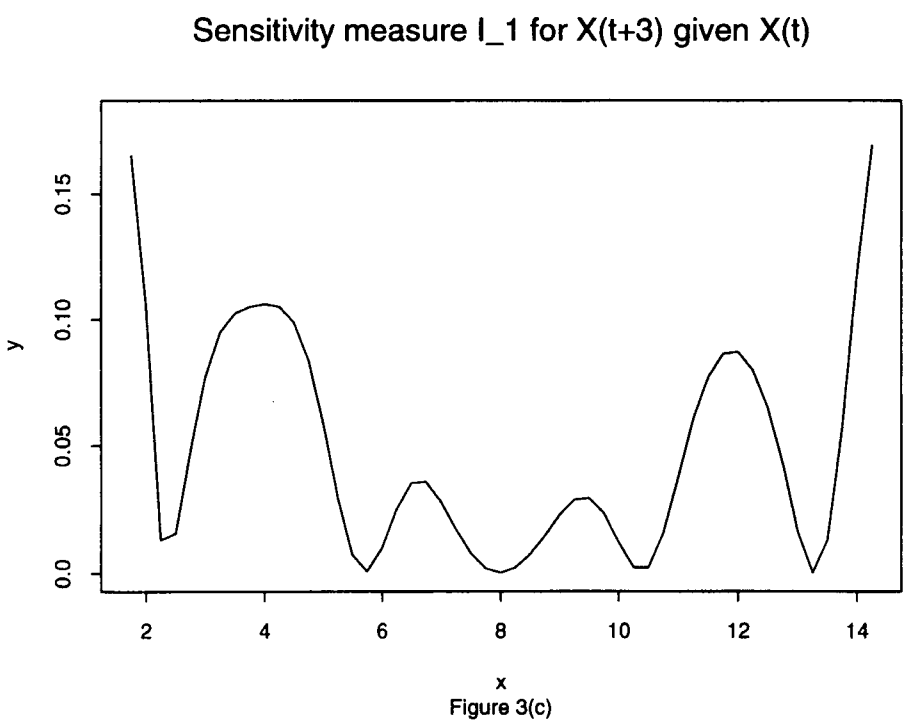

Figure 3(c)

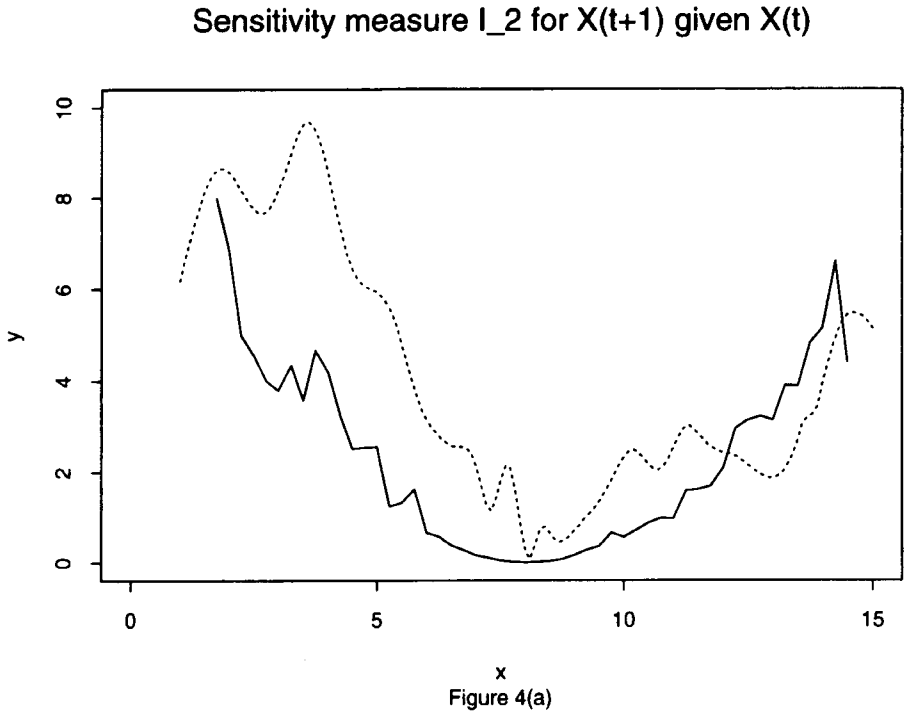

Figure 4(a)

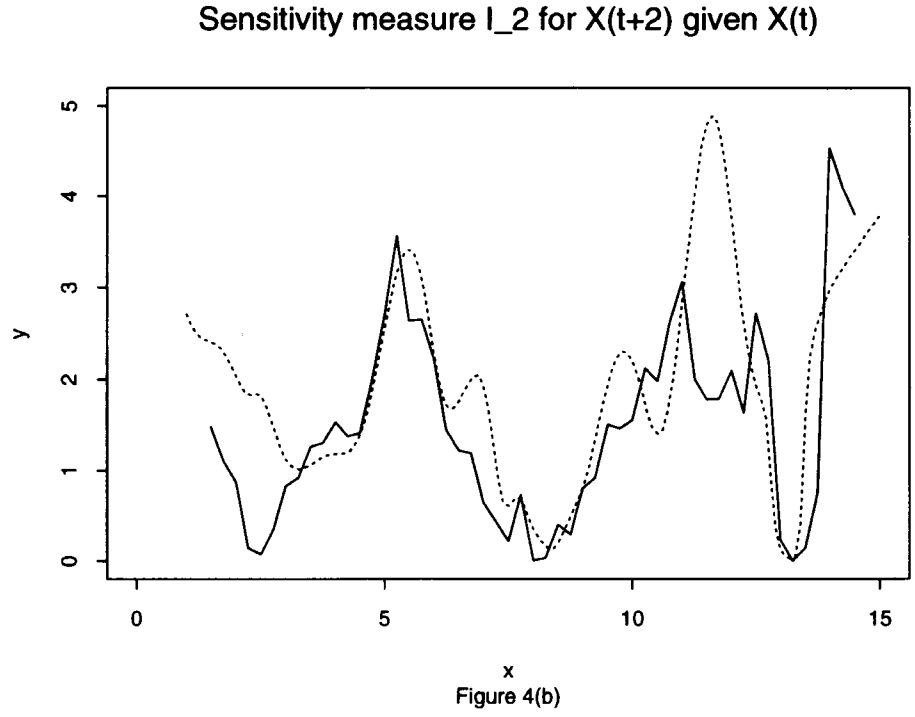

Figure 4(b)

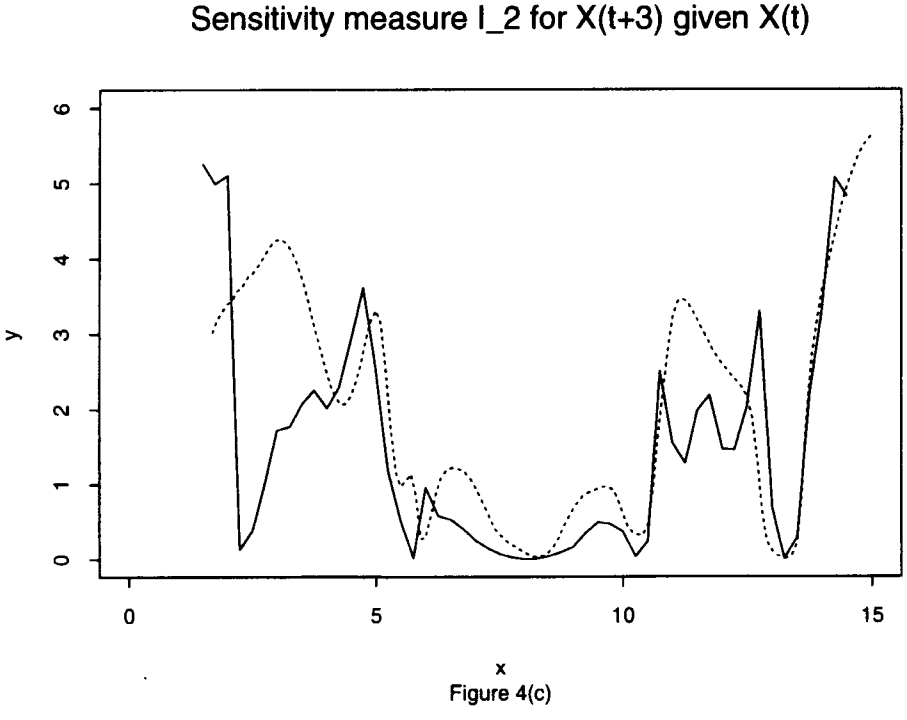

Figure 4(c)

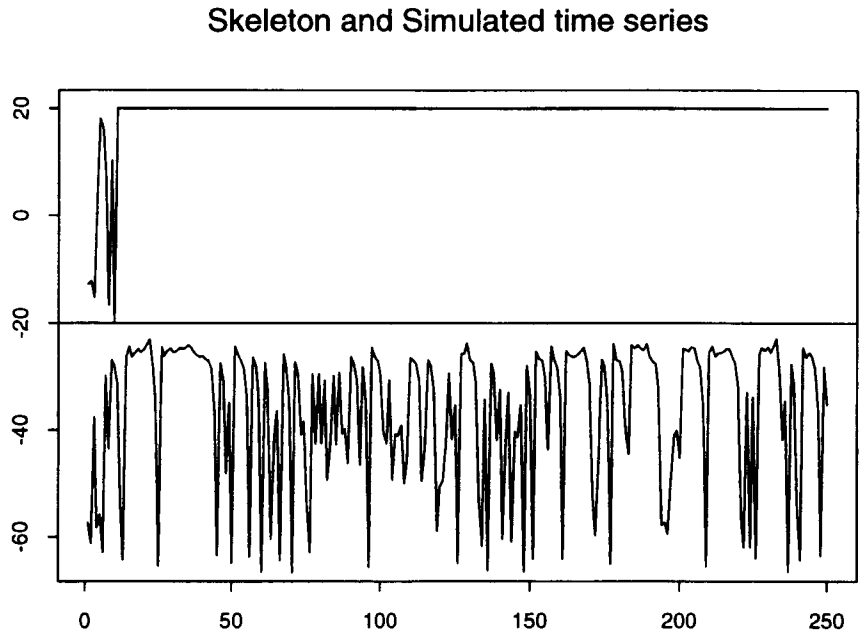

Figure 5(a)

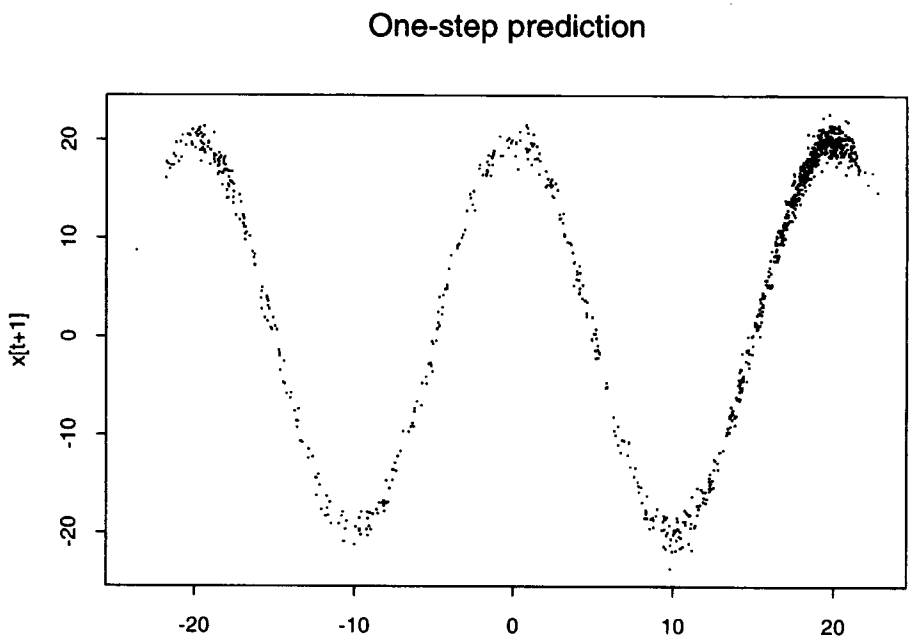

Figure 5(b)

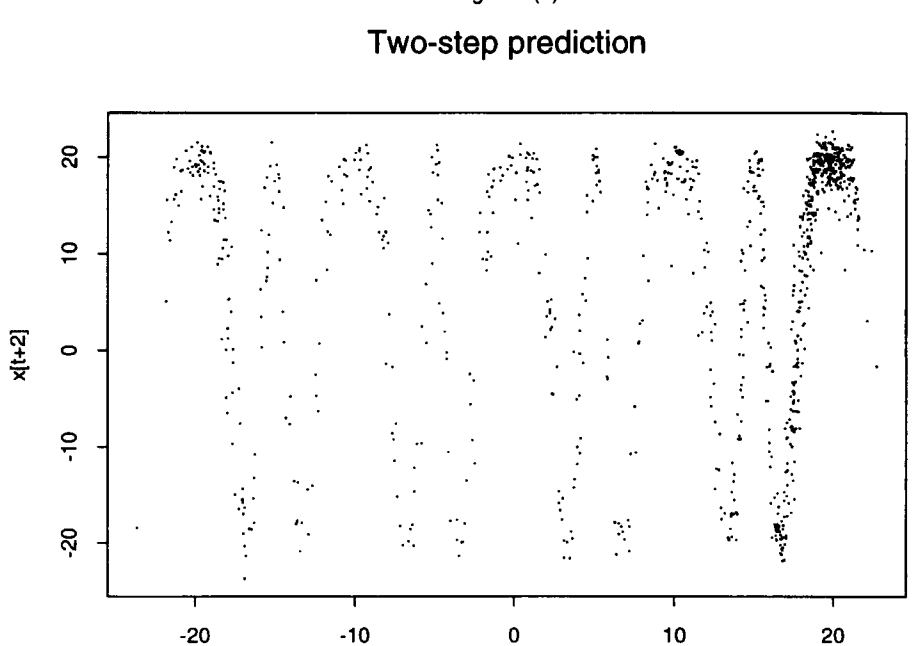

Figure 5(c)

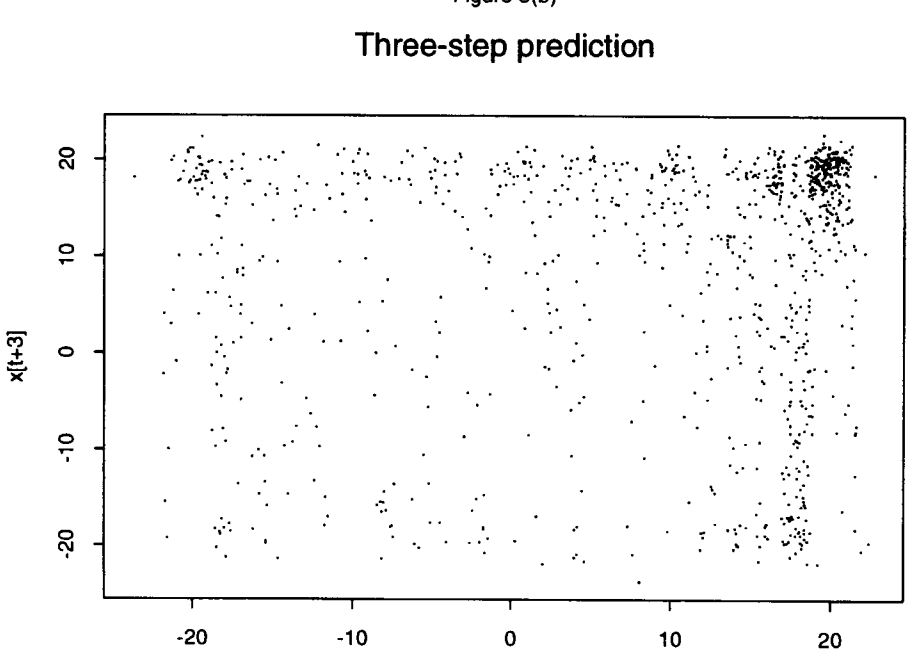

Figure 5(d)

Conditional density of X(t+1) given X(t)

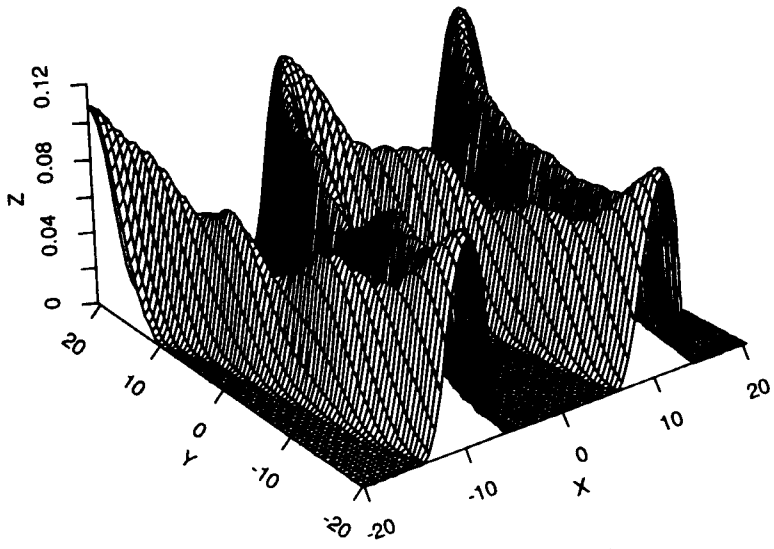

Figure 6(a)

Conditional density of X(t+2) given X(t)

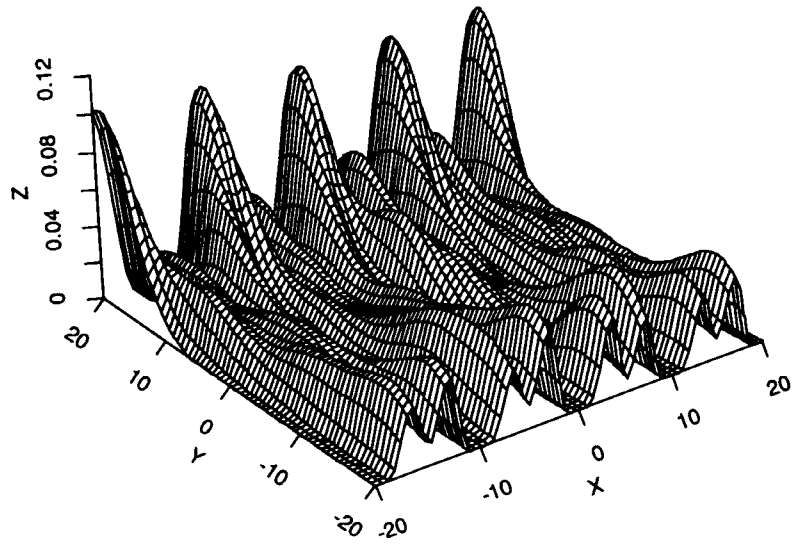

Figure 6(b)

Conditional density of X(t+3) given X(t)

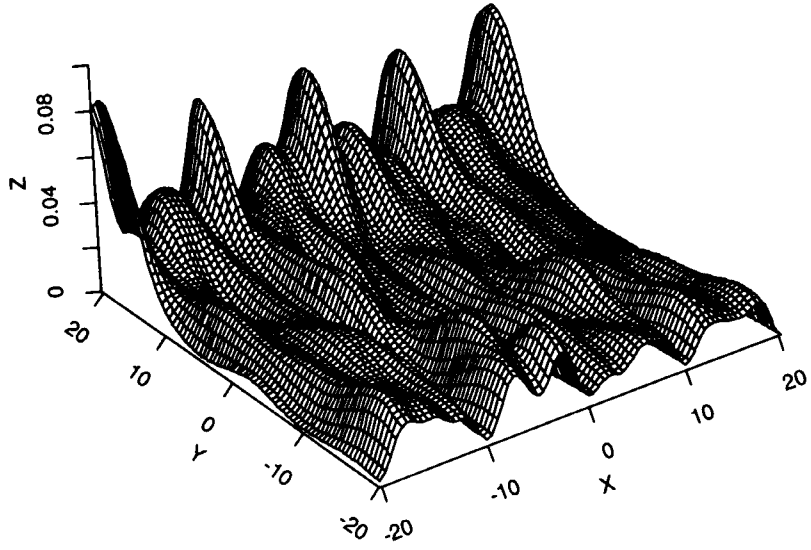

Figure 6(c)

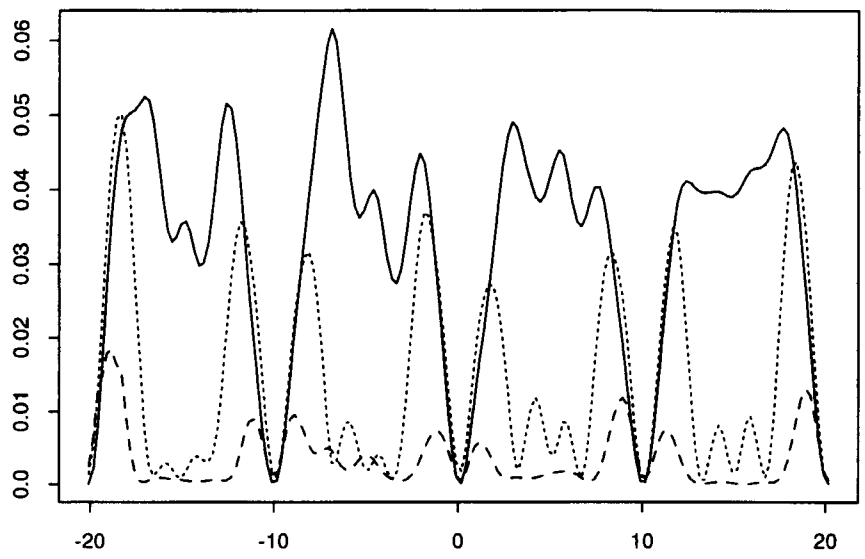

Figure 7

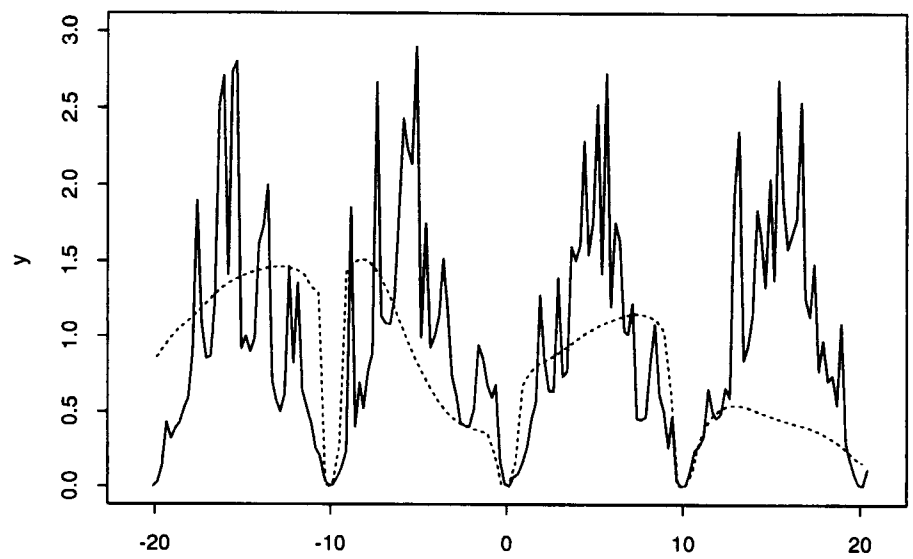

Figure 8(a)

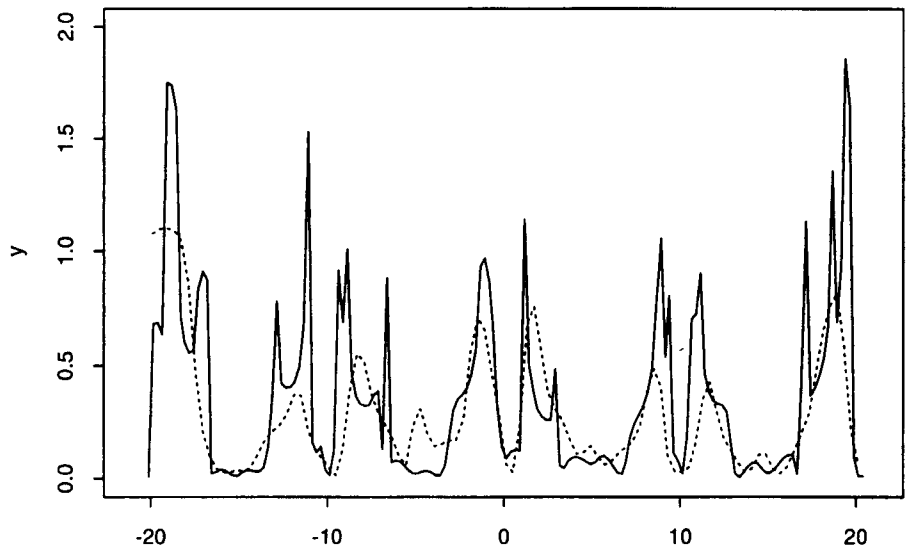

Figure 8(b)

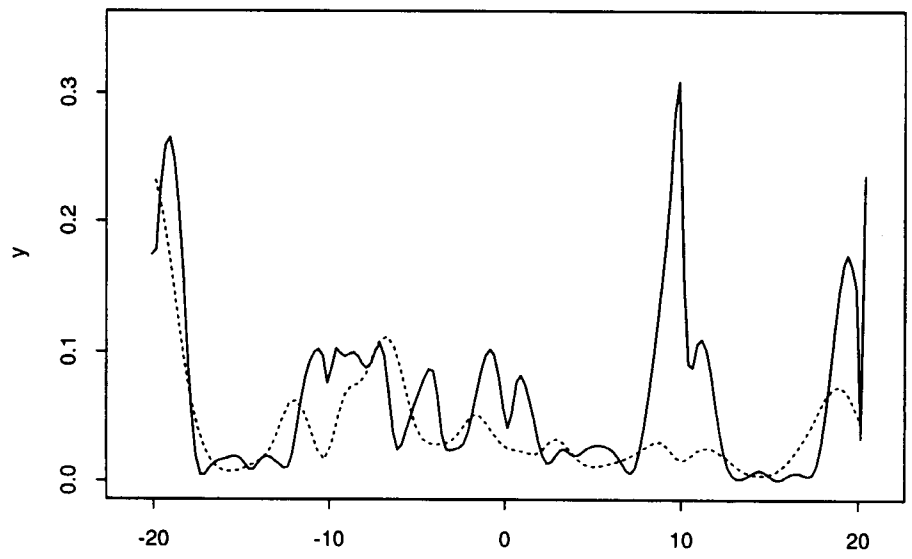

Figure 8(c)