AN ADAPTIVE VARIABLE SELECTION FOR NONLINEAR AUTOREGRESSIVE TIME SERIES MODEL

Fueda, Kaoru
Graduate School of Environmental Science, Okayama University

https://doi.org/10.5109/12594

出版情報：Bulletin of informatics and cybernetics. 37, pp.109-121, 2005-12. Research Association of Statistical Sciences
バージョン：
権利関係：

AN ADAPTIVE VARIABLE SELECTION FOR NONLINEAR AUTOREGRESSIVE TIME SERIES MODEL

by

Kaoru Fueda

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.37

FUKUOKA, JAPAN
2005

Bulletin of Informatics and Cybernetics, Vol. 37, 2005

AN ADAPTIVE VARIABLE SELECTION FOR NONLINEAR AUTOREGRESSIVE TIME SERIES MODEL

By

Kaoru Fueda*

Abstract

Estimating an autoregressive function and its derivatives is important to analyze chaotic time series, especially to estimate the Lyapunov exponent. In this article, we propose an adaptive variable selection method to estimate a nonlinear autoregressive function and its derivatives. Our method has a number of advantages. Since an attractor of chaotic time series is bounded, most kernel-based local mean nonparametric methods have the bias near the boundary of the attractor. In contrast, the order of the bias of our method is higher than that of most existing methods. To estimate derivatives of nonlinear function, we approximate it locally using a linear function. Since correlation dimension of the attractor is not integer, local linear regression has multicollinearity at many points and makes the variance of the estimator of the nonlinear function large. Our method reduces this problem by adaptive variable selection in local linear regression problem.

Key Words and Phrases: Nonlinear time series analysis, Nonparametric regression, Local linear smoother, Principal component analysis.

1. Introduction

In an analysis of data from nonlinear autoregressive time series with dynamic noise, a central issue is whether randomness of the data is caused solely by the dynamic noise or by the nonlinearity of the autoregressive model as well. This article investigates the estimation of the Lyapunov exponent for the nonlinear autoregressive time series model to quantify the sensitive dependence on an initial value. Cheng and Tong (1993) considered a nonlinear autoregressive model

$$X_t = F(X_{t-1}, X_{t-2}, \ldots, X_{t-d}) + \varepsilon_t,$$

and related the intuitive geometric reconstruction of phase space in theoretical physics with statistical theory of the determination of order of a nonlinear autoregressive model. They called the order d embedding dimension, and Cheng and Tong (1995) proposed estimators of F and d using Nadaraya-Watson kernel estimator and cross-validation method. Fueda and Yanagawa (2001) introduced the delay time τ to Cheng and Tong's autoregressive model

$$X_t = F(X_{t-\tau}, X_{t-2\tau}, \ldots, X_{t-d\tau}) + \varepsilon_t$$

* Graduate School of Environmental Science, Okayama University, Naka 3-1-1, Tsushima, Okayama 700-8530 Japan. tel +81–86–251–8834 fueda@ems.okayama-u.ac.jp

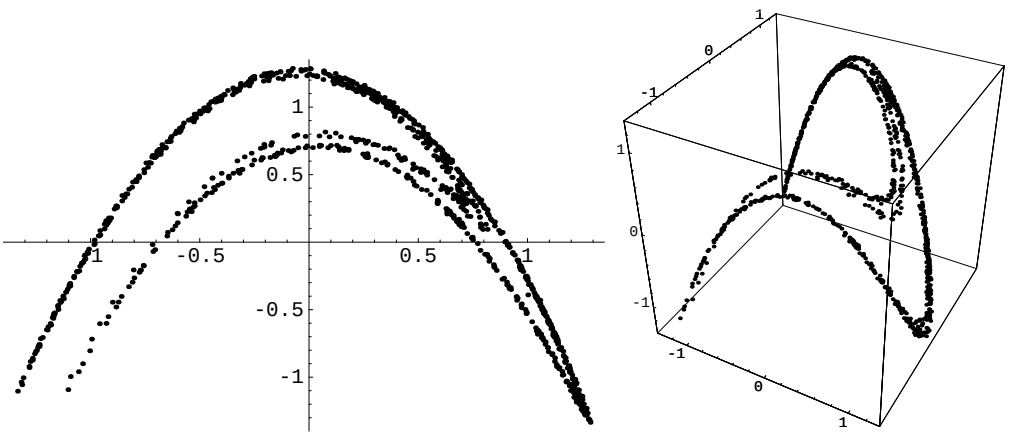

Figure 1: Left:The attractor of Henon map with noise. Right:Plot of $\{(X_{t-2}, X_{t-1}, X_t)\}$

to embed the time series in a lower dimensional space, which is desirable from the view point of *curse of dimensionality*. They proposed estimators of F, d and τ, and proved their consistency. Yonemoto and Yanagawa (2001) pointed out that Fueda and Yanagawa's method often fails to estimate the d and τ, when the size of data is rather small. They proposed a new method of estimation and confirmed that their method works well by simulation.

To consider the distribution of these models, we put

$$\mathcal{X}_t = \left(X_{t-(d-1)}, X_{t-(d-2)}, \ldots, X_{t-1}, X_t\right)^T$$

and we call a set $\{\mathcal{X}_t | d \leq t \leq N\}$ as an attractor. For Henon map:

$$F(X_{t-1}, X_{t-2}) = 1 - 1.4X_{t-1}^2 + 0.3X_{t-2}$$

and $Var(\varepsilon_t) = 0.01^2$, a plot of the attractor and embedding to 3-dimensional space are shown in Fig.1.

Grassberger and Procaccia (1983a,b) put

$$C_N(r) = \left(\begin{array}{c} N \\ 2 \end{array} \right)^{-1} \sum_{i<j}^{N} I\left(||Y_i - Y_j|| \leq r\right),$$

where I denotes the indicator function, $||\cdot||$ is a L^2-norm and $r > 0$, and they called $C(r) = lim_{N\to\infty} C_N(r)$ the correlation integral and introduced the correlation dimension as

$$p = \lim_{r\to 0} \frac{\log C(r)}{\log r}$$

if the limit exists. The correlation dimension of sample from continuous distribution such as Gaussian distribution is integer, and non-integer correlation dimension is considered as one of the characteristics of chaotic time series. Kawaguchi and Yanagawa (2001) discussed the method of estimating correlation dimension.

In this article we investigate a nonlinear autoregressive time series model with flexible delay time,

$$X_t = F(X_{t-\tau_1}, X_{t-\tau_2}, \ldots, X_{t-\tau_D}) + \varepsilon_t, \tag{1}$$

where D and $\tau_1, \ldots, \tau_D$ are positive integer and ε_t is the dynamic noise. We use the local linear regression proposed by Stone (1977) to estimate the autoregressive function F. Nadaraya-Watson type kernel estimator has the bias near the boundary of support of data and the attractor of chaotic time series is bounded, but local linear regression reduces the bias near the boundary (Wand and Jones (1995)). Fan, Hu and Truong (1994) considered a class of kernel estimators based on local linear regression estimator and showed the asymptotic normality of these estimators. Cleveland (1979) proposed the local polynomial regression estimator, which is the extension of the local linear regression estimator. We also propose cross-validation of weighted sum of squares to select a suitable set of delay times.

Next we discuss the estimation of derivatives of the autoregressive function to estimate the Lyapunov exponent, which was discussed in Eckmann and Ruelle (1985). Let τ be the greatest common denominator of $\tau_1, \ldots, \tau_D$ in model (1). Then there exist integers $c_1, ..., c_D$ such that $\tau_k = c_k\tau, (k = 1, \ldots, D)$. Let a function $G : R^{c_D} \to R^{c_D}$ be

$$G\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{c_D} \end{pmatrix} = \begin{pmatrix} F(x_{c_1}, \ldots, x_{c_D}) \\ x_1 \\ \vdots \\ x_{c_D-1} \end{pmatrix}.$$

Then Lyapunov exponent of deterministic dynamical system

$$X_t = F(X_{t-\tau_1}, X_{t-\tau_2}, \ldots, X_{t-\tau_D}),$$

which is called skeleton of model (1), is defined as

$$\lim_{n\to\infty} \frac{1}{2n} \log |\mu_n(x_0)|$$

when the limit exists, where $\mu_n(x_0)$ is the largest eigenvalue of a positive definite matrix $T_n(x_0)^T T_n(x_0)$ and

$$T_n(x_0) = G'\left(G^{n-1}(x_0)\right) G'\left(G^{n-2}(x_0)\right) \cdots G'\left(G(x_0)\right) G'\left(x_0\right).$$

Yonemoto and Yanagawa (2004) proposed a consistent estimator of Lyapunov exponent. It is necessary to estimate the derivatives of autoregressive function for Yonemoto and Yanagawa's method working well. However, for chaotic time series whose correlation dimension is not integer, the estimator of derivatives tends to unstable.

For above example of Henon map, the number of variables of the autoregressive function should be 2, but as Figure 1 shows, observed data lies on a one dimensional line at some points and variance of the estimator of derivatives at such point tends to be large. Thus we have to reduce the number of variables.

One of the important approaches is the projection pursuit regression proposed by Friedman and Stuetzle (1981). A simple approach is the average derivative estimation proposed by Härdel and Stoker (1989). The sliced inverse regression (SIR) method

proposed by Li (1991) is one of the most powerful method for dimension reduction. However, Xia, *et al.* (2002) pointed out that in autoregressive time series analysis the SIR method requires time reversibility which is the exception rather than rule in time series analysis. The minimum average variance estimation proposed Xia, *et al.* (2002) is a natural extension of Cheng and Tong (1995), which proposed the method of variable selection, to the method of estimation of dimension reduction space.

For chaotic time series model, the direction with which the derivatives should be estimated differs at each part of support of the time series data. In this article, we will propose a new method to select the directions for estimating derivatives. Our approach is inspired by the idea of principal component selection of linear regression problem and local linear smoothers described in Fan and Gijbels (1996).

The remainder of this article is organized as follows. Section 2 states the method of selecting embedding dimension d and delay times $(\tau_1, \tau_2, \ldots, \tau_d)$ and that of estimating the autoregressive function F. Section 3 presents adaptive variable selection to estimate derivatives of the autoregressive function.

2. Estimation of Autoregressive Function

2.1. Estimation method

In this section, we use the method of estimating the autoregressive function. To reduce the bias near the boundary of the attractor and estimate the derivative, we use the local linear regression to estimate the autoregressive function.

Consider the non-linear autoregressive model (1). We assume that $\{X_t\}$ is a discrete-time strictly stationary time series with $E\left[X_t^2\right] < \infty$, and for any t,

$$E\left[\varepsilon_t | \mathcal{A}_1^{t-1}(X)\right] = 0, \text{ almost surely.} \tag{2}$$

and

$$E\left[\varepsilon_t^2 | \mathcal{A}_1^{t-1}(X)\right] = \sigma^2, (\sigma > 0), \text{ almost surely,}$$

where $\mathcal{A}_s^t(X)$ denotes the sigma algebra generated by $(X_s, \ldots, X_t)$, for $s \leq t$. Note that from (1) and (2), it follows that

$$F(X_{t-\tau_1}, \ldots, X_{t-\tau_D}) = E\left[X_t | X_{t-\tau_1}, \ldots, X_{t-\tau_D}\right]$$

with $E\left[\varepsilon_t | X_{t-\tau_1}, \ldots, X_{t-\tau_D}\right] = 0$.

The embedding dimension and the delay times are defined as follows.

DEFINITION 2.1. The time series $\{X_t\}$ is said to have the embedding dimension D with the delay times $\tau_1, \ldots, \tau_D$ if and only if there exists positive integers $D < \infty$ and $\tau_1 < \cdots < \tau_D < \infty$ such that

$$E\left[X_t | X_{t-t_1}, X_{t-t_2}, \ldots, X_{t-t_d}\right] \neq E\left[X_t | X_{t-\tau_1}, X_{t-\tau_2}, \ldots, X_{t-\tau_D}\right] \ a.e.$$

for any $t_1, \ldots, t_d$ such that $\{t_1, \ldots, t_d\} \not\supset \{\tau_1, \ldots, \tau_D\}$, and

$$E\left[X_t | X_{t-t_1}, X_{t-t_2}, \ldots, X_{t-t_d}\right] = E\left[X_t | X_{t-\tau_1}, X_{t-\tau_2}, \ldots, X_{t-\tau_D}\right] \ a.e.$$

for any $t_1, \ldots, t_d$ such that $\{t_1, \ldots, t_d\} \supset \{\tau_1, \ldots, \tau_D\}$.

This is a natural extension of the embedding dimension and the delay time defined in Fueda and Yanagawa (2001).

For simplicity we put

$$F_{t_1,\ldots,t_d}(x_1,\ldots,x_d) = E\left[X_t | X_{t-t_1} = x_1, \ldots, X_{t-t_d} = x_d\right].$$

Denoting the variances of residuals by

$$\sigma^2(t_1,\ldots,t_d) = E\left[(X_t - F_{t_1,\ldots,t_d}(X_{t-t_1},\ldots,X_{t-t_d}))^2\right].$$

By definition of the embedding dimension and delay times, we have the following lemma.

LEMMA 2.2. *For positive integers d and $t_1,\ldots,t_d$,*
i) If $\{t_1,\ldots,t_d\} \supset \{s_1,\ldots,s_{d'}\}$, then

$$\begin{aligned}
&\sigma^2(s_1,\ldots,s_{d'}) - \sigma^2(t_1,\ldots,t_d) \\
&= E\left[(F_{s_1,\ldots,s_{d'}}(X_{t-s_1},\ldots,X_{t-s_{d'}}) - F_{t_1,\ldots,t_d}(X_{t-t_1},\ldots,X_{t-t_d}))^2\right] \\
&\geq 0.
\end{aligned}$$

ii) If $\{t_1,\ldots,t_d\} \supset \{\tau_1,\ldots,\tau_D\}$, then

$$\sigma^2(t_1,\ldots,t_d) = \sigma^2(\tau_1,\ldots,\tau_D).$$

iii) If $\{t_1,\ldots,t_d\} \not\supset \{\tau_1,\ldots,\tau_D\}$, then

$$\sigma^2(t_1,\ldots,t_d) > \sigma^2(\tau_1,\ldots,\tau_D).$$

Proof. i) For simplicity, we write $F_{s_1,\ldots,s_{d'}}(X_{t-s_1},\ldots,X_{t-s_{d'}})$ as $F_{s_1,\ldots,s_{d'}}$ and $F_{t_1,\ldots,t_d}(X_{t-t_1},\ldots,X_{t-t_d})$ as $F_{t_1,\ldots,t_d}$.

$$\begin{aligned}
0 \;&\leq\; E\left[(F_{s_1,\ldots,s_{d'}}(X_{t-s_1},\ldots,X_{t-s_{d'}}) - F_{t_1,\ldots,t_d}(X_{t-t_1},\ldots,X_{t-t_d}))^2\right] \\
&=\; E\left[((X_t - F_{t_1,\ldots,t_d}) - (X_t - F_{s_1,\ldots,s_{d'}}))^2\right] \\
&=\; \sigma^2(t_1,\ldots,t_d) + \sigma^2(s_1,\ldots,s_{d'}) \\
&\quad -2E[(X_t - F_{t_1,\ldots,t_d})(X_t - F_{t_1,\ldots,t_d} + F_{t_1,\ldots,t_d} - F_{s_1,\ldots,s_{d'}})] \\
&=\; \sigma^2(t_1,\ldots,t_d) + \sigma^2(s_1,\ldots,s_{d'}) \\
&\quad -2\left(\sigma^2(t_1,\ldots,t_d) + E[(X_t - F_{t_1,\ldots,t_d})(F_{t_1,\ldots,t_d} - F_{s_1,\ldots,s_{d'}})]\right) \\
&=\; \sigma^2(s_1,\ldots,s_{d'}) - \sigma^2(t_1,\ldots,t_d) \\
&\quad -2E\left[(X_t - F_{t_1,\ldots,t_d})E[F_{t_1,\ldots,t_d} - F_{s_1,\ldots,s_{d'}} | X_{t-t_1},\ldots,X_{t-t_d}]\right] \\
&=\; \sigma^2(s_1,\ldots,s_{d'}) - \sigma^2(t_1,\ldots,t_d).
\end{aligned}$$

ii) From the definition of $\{\tau_1,\ldots,\tau_D\}$ we have

$$F_{t_1,\ldots,t_d}(X_{t-t_1},\ldots,X_{t-t_d}) = F_{\tau_1,\ldots,\tau_D}(X_{t-\tau_1},\ldots,X_{t-\tau_D}) \ a.e.,$$

and from Lemma 2.2 i) we have

$$\begin{aligned}
&\sigma^2(t_1,\ldots,t_d) - \sigma^2(\tau_1,\ldots,\tau_D) \\
&= E\left[(F_{t_1,\ldots,t_d}(X_{t-t_1},\ldots,X_{t-t_d}) - F_{\tau_1,\ldots,\tau_D}(X_{t-\tau_1},\ldots,X_{t-\tau_D}))^2\right] \\
&= 0.
\end{aligned}$$

iii) Let $\{s_1, \ldots, s_{d'}\} = \{t_1, \ldots, t_d\} \cup \{\tau_1, \ldots, \tau_D\}$. From the definition of $\{\tau_1, \ldots, \tau_D\}$ we have

$$\begin{aligned} F_{s_1,\ldots,s_{d'}}(X_{t-s_1}, \ldots, X_{t-s_{d'}}) &= F_{\tau_1,\ldots,\tau_D}(X_{t-\tau_1}, \ldots, X_{t-\tau_D}) \ a.e. \\ &\neq F_{t_1,\ldots,t_d}(X_{t-t_1}, \ldots, X_{t-t_d}), \end{aligned}$$

and from Lemma 2.2 i) and ii) we have

$$\begin{aligned} &\sigma^2(t_1, \ldots, t_d) - \sigma^2(\tau_1, \ldots, \tau_D) \\ &= \sigma^2(t_1, \ldots, t_d) - \sigma^2(s_1, \ldots, s_{d'}) + \sigma^2(s_1, \ldots, s_{d'}) - \sigma^2(\tau_1, \ldots, \tau_D) \\ &= E\left[(F_{t_1,\ldots,t_d}(X_{t-t_1}, \ldots, X_{t-t_d}) - F_{s_1,\ldots,s_{d'}}(X_{t-s_1}, \ldots, X_{t-s_{d'}}))^2\right] \\ &> 0. \end{aligned}$$

□

For each $d > 0$, $0 < t_1 < t_2 < \cdots < t_d$, and any given $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)^T \in R^d$, a local linear expansion of $F_{t_1,\ldots,t_d}(\boldsymbol{x})$ at $\boldsymbol{z}$ is

$$F_{t_1,\ldots,t_d}(X_{t-t_1}, \ldots, X_{t-t_d}) \approx \beta_0 + \sum_{j=1}^{d} \beta_j (X_{t-t_j} - z_j), \tag{3}$$

where $\beta_0 = F_{t_1,\ldots,t_d}(\boldsymbol{z})$ and

$$\beta_j = \left. \frac{\partial F_{t_1,\ldots,t_d}(x_1, \ldots, x_d)}{\partial x_j} \right|_{(x_1,\ldots,x_d)=\boldsymbol{z}}, \ j = 1, \ldots, d.$$

Note that the right-hand side of (3) is the tangent plane of $F_{t_1,\ldots,t_d}$ at $\boldsymbol{z}$. The residuals are then

$$X_t - F_{t_1,\ldots,t_d}(X_{t-t_1}, \ldots, X_{t-t_d}) \approx X_t - \left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_{t-t_j} - z_j) \right).$$

Let $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ be the observed data, L be sufficiently large for $\tau_D \leq L$. Following the idea of local linear smoothing estimation, we can estimate $\sigma^2(t_1, \ldots, t_d)$ by exploiting the approximation

$$\begin{aligned} &\sum_{t=L+1}^{N} (X_t - F_{t_1,\ldots,t_d}(X_{t-t_1}, \ldots, X_{t-t_d}))^2 \\ &\approx \sum_{t=L+1}^{N} \left(X_t - \left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_{t-t_j} - z_j) \right) \right)^2 K_{d,h}\left(\mathcal{X}_t - \boldsymbol{z}\right), \end{aligned}$$

where $\mathcal{X}_t = (X_{t-\tau_1}, X_{t-\tau_2}, \ldots, X_{t-\tau_d})^T \in R^d$, $K_{d,h}(\cdot)$ is a weight function such that

$$K_{d,h}(\boldsymbol{z}) = \frac{1}{h^d} K_d\left(\frac{1}{h}\boldsymbol{z}\right), \ \boldsymbol{z} \in R^d,$$

where K_d is a $d-$dimensional kernel function. We consider a minimizing problem of the weighted sum of squares

$$D\left(\boldsymbol{\beta}, \boldsymbol{z}\right) = \sum_{t=L+1}^{N} \left(X_t - \left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_{t-t_j} - z_j) \right) \right)^2 K_{d,h}\left(\mathcal{X}_t - \boldsymbol{z}\right).$$

By solving the normal equation, we calculate the minimizer of the weighted sum of squares as

$$\boldsymbol{\beta}(\boldsymbol{z}) = (\beta_0(\boldsymbol{z}), \beta_1(\boldsymbol{z}), \ldots, \beta_d(\boldsymbol{z}))^T = \left(\boldsymbol{M}(\boldsymbol{z})^T \boldsymbol{W}(\boldsymbol{z}) \boldsymbol{M}(\boldsymbol{z})\right)^{-1} \boldsymbol{M}(\boldsymbol{z})^T \boldsymbol{W}(\boldsymbol{z}) \boldsymbol{Y}, \quad (4)$$

where

$$\boldsymbol{M}(\boldsymbol{z}) = \begin{pmatrix} 1 & X_{L+1-\tau_1} - z_1 & \cdots & X_{L+1-\tau_d} - z_d \\ 1 & X_{L+2-\tau_1} - z_1 & \cdots & X_{L+2-\tau_d} - z_d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N-\tau_1} - z_1 & \cdots & X_{N-\tau_d} - z_d \end{pmatrix},$$

$$\boldsymbol{W}(\boldsymbol{z}) = diag\left[K_{d,h}(\mathcal{X}_{L+1} - \boldsymbol{z}), \ldots, K_{d,h}(\mathcal{X}_N - \boldsymbol{z})\right],$$

and $\boldsymbol{Y} = (X_{L+1}, X_{L+2}, \ldots, X_N)^T$. Then we put $\hat{F}_{t_1,\ldots,t_d}(\boldsymbol{z}) = \beta_0(\boldsymbol{z})$ as an estimator of $F_{t_1,\ldots,t_d}(\boldsymbol{z})$. This local linear estimator is similar to Nadaraya-Watson kernel estimator; however, the bias of this estimator is smaller than the one of Nadaraya-Watson kernel estimator. See ex. Fan and Gijbels (1996).

2.2. Delay Time Selection

As a criterion to select delay times of the autoregressive function, we shall select d and $t_1, t_2, \ldots, t_d$ which minimize

$$\frac{1}{N-L} \sum_{t=L+1}^{N} D\left(\boldsymbol{\beta}(\mathcal{X}_t), \mathcal{X}_t\right).$$

Many methods were proposed to estimate this value. Akaike (1973), Takeuchi (1976) and Konishi and Kitagawa (1996) estimated the bias of the estimator of Kullback-Leibler divergence and proposed information criteria. Taniguchi and Kakizawa (2000) generalized Takeuchi's information criterion replacing Kullback-Leibler divergence with general distance. Other method is a simple and traditional cross-validation, which delete one sample to make unbiased estimator. We now extend the cross-validation method of Cheng and Tong (1993) and Fueda and Yanagawa (2001). A similar extension may be effected by using the approach of Auestad and Tjϕstheim (1990), which is asymptotically equivalent to the cross-validation method.

Put

$$CV(d, t_1, \ldots, t_d) = \frac{1}{N-L} \sum_{t=L+1}^{N} \left(X_t - \hat{F}^{(-t)}_{t_1,\ldots,t_d}(\mathcal{X}_t) \right)^2,$$

where $\hat{F}^{(-t)}_{t_1,\ldots,t_d}(\mathcal{X}_t) = \beta_0^{(-t)}$ and $\boldsymbol{\beta}^{(-t)} = \left(\beta_0^{(-t)}, \beta_1^{(-t)}, \ldots, \beta_d^{(-t)}\right)^T \in R^{d+1}$ is the minimizer of a weighted sum of squares

$$\sum_{s=L+1, s\neq t}^{N} \left(X_s - \left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_{s-t_j} - X_{t-t_j}) \right) \right)^2 K_{d,h}\left(\mathcal{X}_s - \mathcal{X}_t\right).$$

Thus we estimate the embedding dimension and the delay times as the minimizer of $CV(d, t_1, \ldots, t_d)$, and denote them by $\hat{D}$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}$.

To investigate the asymptotic properties of the estimator, we need the following assumptions:

1. $\{X_t\}$ is strictly stationary and ergodic.
2. The kernel K_d is a compactly supported, bounded kernel such that $\int \boldsymbol{z}\boldsymbol{z}^T K_d(\boldsymbol{z}) d\boldsymbol{z} = \mu_2(K) I$, where $\mu_2(K_d) \neq 0$ is scalar and I is the $d \times d$ identity matrix. In addition, the kernel K_d is spherically symmetric, that is $K(\boldsymbol{z})d\boldsymbol{z} = -K(\boldsymbol{z})d\boldsymbol{z}$ for all $z \in R^d$.
3. Let $f_X(x_1, \ldots, x_d)$ be the stationary joint density of $(X_{t-t_1}, \ldots, X_{t-t_d})$ and $\mathrm{supp}(f_X)$ be the support of f_X. For $\boldsymbol{x} \in \mathrm{supp}(f_X)$, f_X is continuously differentiable at $\boldsymbol{x}$ and all second-order derivatives of $F_{\tau_1,\ldots,\tau_D}$ are continuous at $\boldsymbol{x}$.
4. $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.
5. There is a convex set $\mathcal{S}$ with nonnull interior such that

$$\inf_{\boldsymbol{x}\in\mathcal{S}} f(\boldsymbol{x}) > 0.$$

THEOREM 2.3. *Suppose that the assumption 1-5 hold.*
1. For $t_1, \ldots, t_d$ such that $\{t_1, \ldots, t_d\} \supset \{\tau_1, \ldots, \tau_D\}$, and $\boldsymbol{z} = (z_{t_1}, \ldots, z_{t_d})^T \in R^d$ such that $f_X(\boldsymbol{z}) > 0$,

$$E[\hat{F}_{t_1,\ldots,t_d}(z_{t_1}, \ldots, z_{t_d}) - F_{\tau_1,\ldots,\tau_D}(z_{\tau_1}, \ldots, z_{\tau_D})|\mathcal{X}] = \frac{1}{2}\mu_2(K) tr\left(\frac{\partial^2 F}{\partial z \partial z^T}\right) h^2 + o_p(h^2) \tag{5}$$

and

$$Var(\hat{F}_{t_1,\ldots,t_d}(z_{\tau_1}, \ldots, z_{\tau_D})|\mathcal{X}) = \frac{R(K)\sigma^2}{nh^d f(\boldsymbol{z})} + o_p(n^{-1}h^{-d}) \tag{6}$$

2. $\lim_{n\to\infty} Pr\left\{\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}\} \not\supset \{\tau_1, \ldots, \tau_D\}\right\} = 0.$

Proof.
1. Since $\{t_1, \ldots, t_d\} \supset \{\tau_1, \ldots, \tau_D\}$, we have $F_{t_1,\ldots,t_d}(z_{t_1}, \ldots, z_{t_d}) = F_{\tau_1,\ldots,\tau_D}(z_{\tau_1}, \ldots, z_{\tau_D})$ from Lemma 2.2 i). Then, we have (5) and (6) from Theorem 2.1 of Ruppert and Wand (1994).
2. From ergodicity of $\{X_t\}$, we have

$$CV(d, t_1, \ldots, t_d) = \sigma(t_1, \ldots, t_d) + O_P(n^{-1/2}).$$

Thus for $\{t_1, \ldots, t_d\} \not\supset \{\tau_1, \ldots, \tau_D\}$, from Lemma 2.2 iii)

$$Pr\left\{\{t_1, \ldots, t_d\} = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}\}\right\} \leq Pr\{CV(t_1, \ldots, t_d) \leq CV(\tau_1, \ldots, \tau_D)\} \to 0$$

as $n \to \infty$. □

3. Estimation of Derivatives

In this section, we discuss the estimation of derivatives of the autoregressive function. The most important character of chaos is the sensitive dependence on initial value and noise, which is measured by the derivatives of the autoregressive function. Usually, the local linear regression $\hat{\boldsymbol{\beta}}(\boldsymbol{z})$ given in (4) gives the estimator of the autoregressive function as

$$\frac{\partial \hat{F}}{\partial z_i}(\boldsymbol{z}) = \hat{\beta}_i(\boldsymbol{z}) \text{ for } i = 1, 2, \ldots, d,$$

(Fan and Gijbels (1996)). However, as Figure 1 shows, observed data looks like to lie on a one dimensional line at some points. This locally collinear phenomenon occurred because the correlation dimension of chaotic time series is not integer. Though the common way to reduce collinearity is variable selection, we can delete neither X_{t-1} nor X_{t-2} for Henon map: $F(X_{t-1}, X_{t-2}) = 1 - 1.4X_{t-1}^2 + 0.3X_{t-2}$. To get a stable estimator of derivatives, we introduce adaptive variable selection based on the principal component analysis.

3.1. Directions for Estimating Derivatives

Let $\hat{D}$ and $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}$ be the embedding dimension and delay times which were selected in section 2. For simplicity we omit $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}$ from $F_{\hat{\tau}_1,\ldots,\hat{\tau}_{\hat{D}}}$. Let $\hat{F}(\boldsymbol{z})$ be the estimator of $F(\boldsymbol{z})$ estimated in section 2 for $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)^T \in R^d$. we consider a weighted sum of squares

$$D_1(\boldsymbol{\beta}, \boldsymbol{z}) = \sum_{t=L+1}^{N} \left(X_t - \hat{F}(\boldsymbol{z}) - \left(\sum_{j=1}^{d} \beta_j (X_{t-t_j} - z_j) \right) \right)^2 K_{d,h}(\mathcal{X}_t - \boldsymbol{z}), \qquad (7)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T \in R^d$ is a vector of coefficient. A standard way to solve this minimization problem is rewriting

$$D_1(\boldsymbol{\beta}, \boldsymbol{z}) = \left|\left| \boldsymbol{W}(\boldsymbol{z})^{1/2} \boldsymbol{Y}_1(\boldsymbol{z}) - \boldsymbol{W}(\boldsymbol{z})^{1/2} \boldsymbol{M}_1(\boldsymbol{z}) \boldsymbol{\beta} \right|\right|^2,$$

where

$$\boldsymbol{M}_1(\boldsymbol{z}) = \begin{pmatrix} X_{L+1-\tau_1} - z_1 & \cdots & X_{L+1-\tau_d} - z_d \\ X_{L+2-\tau_1} - z_1 & \cdots & X_{L+2-\tau_d} - z_d \\ \vdots & \ddots & \vdots \\ X_{N-\tau_1} - z_1 & \cdots & X_{N-\tau_d} - z_d \end{pmatrix},$$

$$\boldsymbol{Y}_1(\boldsymbol{z}) = \left(X_{L+1} - \hat{F}(\boldsymbol{z}), X_{L+2} - \hat{F}(\boldsymbol{z}), \ldots, X_N - \hat{F}(\boldsymbol{z}) \right)^T$$

and

$$\boldsymbol{W}(\boldsymbol{z})^{1/2} = diag\left[K_{d,h}(\mathcal{X}_{L+1} - \boldsymbol{z})^{1/2}, \ldots, K_{d,h}(\mathcal{X}_N - \boldsymbol{z})^{1/2} \right],$$

and $||\cdot||$ is a L^2-norm. However the variance of estimator is large if the determinant of $\boldsymbol{M}_1(\boldsymbol{z})^T \boldsymbol{W}(\boldsymbol{z}) \boldsymbol{M}_1(\boldsymbol{z})$ is very small. Nonaka, Ando and Konishi (2003) introduced a penalty term to regularize the normal equation. In this article, we investigate the following principal component selection for the multicollinearty problem.

Since $\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{M}_1(\boldsymbol{z})$ is a non-negative definite matrix, there is an orthogonal matrix $\boldsymbol{V}=(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d)$ such that

$$\boldsymbol{V}^T\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{M}_1(\boldsymbol{z})\boldsymbol{V}=diag\left[\lambda_1,\ldots,\lambda_d\right],$$

where $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d\geq 0$. For this matrix $\boldsymbol{V}$, we may rewrite

$$D_1\left(\boldsymbol{\beta},\boldsymbol{z}\right)=D_1\left(V\boldsymbol{\beta}',\boldsymbol{z}\right)=\left|\left|\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{Y}_1-\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{M}_1(\boldsymbol{z})\boldsymbol{V}\boldsymbol{\beta}'\right|\right|^2$$

where $\boldsymbol{\beta}'=\left(\beta_1',\ldots,\beta_d'\right)^T=\boldsymbol{V}^T\boldsymbol{\beta}$. The minimizer of D_1 is given by

$$\hat{\boldsymbol{\beta}}(\boldsymbol{z})=Vdiag\left[1/\lambda_1,\ldots,1/\lambda_d\right]V^T\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{Y}_1.$$

However, due to multicollinearty, if some of λ's are sufficiently small so we may regard λ's as

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p>0\fallingdotseq\lambda_{p+1}\fallingdotseq\cdots\fallingdotseq\lambda_d,$$

then the minimizer $\hat{\boldsymbol{\beta}}(\boldsymbol{z})$ is unstable. In such case we have

$$\boldsymbol{V}^T\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{M}_1(\boldsymbol{z})\boldsymbol{V}\fallingdotseq diag\left[\lambda_1,\ldots,\lambda_p,0,\ldots,0\right],$$

and

$$\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{M}_1(\boldsymbol{z})\boldsymbol{v}_i\fallingdotseq 0\text{ for }i=p+1,\ldots,d.$$

So we have

$$D_1\left(V\boldsymbol{\beta}',\boldsymbol{z}\right)\fallingdotseq\left|\left|\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{Y}_1-\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{M}_1(\boldsymbol{z})\left(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\right)\begin{pmatrix}\beta_1'\\ \vdots\\ \beta_p'\end{pmatrix}\right|\right|^2$$

and note that the right hand side does not depend on $\beta_{p+1}',\ldots,\beta_d'$. To get a stable minimizer, we omit the principal components $\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{M}_1(\boldsymbol{z})\boldsymbol{v}_i$ $(i=p+1,\ldots,d)$ and put $\beta_{p+1}'=\cdots=\beta_d'=0$. Now we have a new design matrix $\boldsymbol{W}(\boldsymbol{z})^{1/2}\boldsymbol{M}_1(\boldsymbol{z})\left(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\right)$ and call $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p$ principal directions. Then we state minimizing problem $D_1\left(\left(\beta_1',\ldots,\beta_p'\right)^T,\boldsymbol{z}\right)$ with respect to p parameters $\beta_1',\ldots,\beta_p'$. The minimizer is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}'(p,\boldsymbol{z}) &= \left(\beta_1',\ldots,\beta_p'\right)^T\\ &= \left(\begin{pmatrix}\boldsymbol{v}_1^T\\ \vdots\\ \boldsymbol{v}_p^T\end{pmatrix}\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{M}_1(\boldsymbol{z})\left(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\right)\right)^{-1}\begin{pmatrix}\boldsymbol{v}_1^T\\ \vdots\\ \boldsymbol{v}_p^T\end{pmatrix}\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{Y}_1\\ &= diag\left[1/\lambda_1,\ldots,1/\lambda_p\right]\begin{pmatrix}\boldsymbol{v}_1^T\\ \vdots\\ \boldsymbol{v}_p^T\end{pmatrix}\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{Y}_1.\end{aligned}$$

So we have the minimizer of $D_1\left(\boldsymbol{\beta},\boldsymbol{z}\right)$ as

$$\begin{aligned}\hat{\boldsymbol{\beta}}(p,\boldsymbol{z}) &= \left(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\right)\hat{\boldsymbol{\beta}}'(p,\boldsymbol{z})\\ &= \left(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p\right)diag\left[1/\lambda_1,\ldots,1/\lambda_p\right]\begin{pmatrix}\boldsymbol{v}_1^T\\ \vdots\\ \boldsymbol{v}_p^T\end{pmatrix}\boldsymbol{M}_1(\boldsymbol{z})^T\boldsymbol{W}(\boldsymbol{z})\boldsymbol{Y}_1.\end{aligned}$$

3.2. Number of Principal Directions

Finally we should select the number of principal directions p. In this section we again use cross-validation. Since the cross-validation method for non-linear problem incorporates many computational costs, we apply the one for weighted least square linear regression problem. Let

$$CV(p, \boldsymbol{z}) = \sum_{t=L+1}^{N} K_{d,h}(\mathcal{X}_t - z) \left(\frac{X_t - \hat{X}_t}{1 - h_{tt}} \right)^2,$$

where $\hat{X}_t = \mathcal{X}_t \hat{\boldsymbol{\beta}}(p, \boldsymbol{z})$ and
$h_{tt} = K_{d,h}(\mathbf{0})(\mathcal{X}_t - \boldsymbol{z})$

$$\times (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) \left(\begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_p^T \end{pmatrix} \boldsymbol{M}_1(\boldsymbol{z})^T \boldsymbol{W}(\boldsymbol{z}) \boldsymbol{M}_1(\boldsymbol{z}) (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) \right)^{-1} \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_p^T \end{pmatrix} (\mathcal{X}_t - \boldsymbol{z})^T.$$

For each $\boldsymbol{z}$, denote the minimizer of $CV(p, \boldsymbol{z})$ as $\hat{p}(\boldsymbol{z})$, then the estimators of the derivatives of $F_{\hat{\tau}_1, \hat{\tau}_{\hat{D}}}$ are given by

$$\frac{\partial \hat{F}}{\partial \boldsymbol{x}}(\boldsymbol{z}) = \hat{\beta}(\hat{p}(\boldsymbol{z}), \boldsymbol{z}).$$

4. Discussion

In this article, though consistency of $(\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}})$ is not proved, $\hat{F}_{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}}$ is a consistent estimator of F. However, if $\hat{D}$ is much larger than D, the variance of the estimator $\hat{F}_{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{D}}}$ is large. It means we should not select a $\hat{D}$ that is too large. Fueda and Yanagawa (2001) proved the consistency of their estimator of embedding dimension and delay time using cross-validation with kernel estimator, and Xia, *et al.* (2002) proved consistency of their estimator of the dimension of the effective dimension reduction (EDR) space using cross-validation with local linear estimator. However they required a very complicated assumption and their rates of convergence are very slow. It is not surprising because Stone (1974) showed that cross-validation criterion and Akaike's information criterion (AIC) are asymptotically equivalent for model selection, and Fujikoshi (1985) showed AIC is not consistent for estimating the true model.

Yonemoto and Yanagawa (2001)'s improvement in Fueda and Yanagawa (2001)'s estimator may works well for this estimator. That is, let

$$CV^* = \min CV(t_1, \ldots, t_d)$$

and

$$\mathcal{T} = \{(d, t_1, \ldots, t_d) | CV(t_1, \ldots, t_d) < (1 + \varepsilon) CV^*\},$$

where $\varepsilon = 0.05$ or 0.1. Then Yonemoto and Yanagawa (2001)'s improvement suggests to use

$$D^* = \min\{d | (d, t_1, \ldots, t_d) \in \mathcal{T}\}$$

and

$$(\tau_1^*, \ldots, \tau_{D^*}^*) = \operatorname{argmin}\{d | (d, t_1, \ldots, t_d) \in \mathcal{T}\}.$$

However, their performance has not checked yet.

Acknowledgement

The author would like to deeply thank Professor Takashi Yanagawa for his contributions to the structure of this work and for other excellent recommendations. I am most grateful to the referee for constructive comments and advice. I also wish to thank Professor Yutaka Tanaka for his advice on multivariate analysis, and Professor Howell Tong for his lecture and stimulating discussion.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *2nd international symposium on information theory,* Ed. B. N. Petrov and F. Csaki, 267-281. Budapest: Akademiai Kiado. (Reproduced (1992) in *Breakthroughs in statistics 1,* Ed. S. Kotz and N. L. Johnson, 610-624. New York: Springer-Verlag.)

Auestad, B. and Tjϕstheim, D. (1990). Identification of nonlinear time series: first order characterization and order determination, *Biometrika,* **77**, 669-688.

Cheng, B. and Tong, H. (1993). On residual sums of squares in non-parametric autoregression, *Stochastic Process. Appl.* **48,** 157-174.

Cheng, B. and Tong, H. (1995). Orthogonal projection, embedding dimension and sample size in chaotic time series from a statistical perspective, In Tong, H., editor, *Chaos and Forecasting,* World Scientific, Singapore, London.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatter plots, *J. Amer. Statist. Assoc.,* **74,** 829-836.

Eckmann, J. P. and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.,* **57,** No. 3, Part I, 617–656.

Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications,* Monographs on statistics and applied probability 66, Chapman and Hall, London.

Fan, J., Hu, T. and Truong, Y. K. (1994). Robust non-parametric function estimation, *Scand. J. Statist.,* **21,** 433-446.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *J. Am. Statist. Ass.*, **76**, 817-823.

Fueda, K. and Yanagawa, T. (2001). Estimating the embedding dimension and delay time from chaotic time series with dynamic noise, *Journal of the Japan Statistical Society,* **31, 1,** 27-38.

Fujikoshi, Y. (1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria, *J. Multiv. Anal.,* **17**, 27-37.

Grassberger, P. and Procaccia, I. (1983a). Characterization for strange attractor, *Physical Review Letters,* **50, 5,** 346-349.

Grassberger, P. and Procaccia, I. (1983b). Measuring the strangeness of strange attractors, *Physica,* **9D, 5,** 189-208.

Härdel, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by method of average derivatives, *J. Am. Statist. Ass.*, **84**, 986-995.

Kawaguchi, A. and Yanagawa, T. (2001). Estimating correlation dimension in chaotic time series, *Bulletin of Informatics and Cybernetics,* **33, 1,** 63-71.

Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection, *Biometrika,* **83,** 875-890.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion), *J. Am. Statist. Ass.*, **86**, 316-342.

Nonaka, Y., Ando, T. and Konishi, S. (2003). Nonlinear regression modeling using regularized local likelihood, *Bulletin of the Computational Statistics of Japan,* **16,** 43-57.

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression, *Ann. Statist.,* **22**, 1346-1370.

Stone, C. J. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *J. R. Statist. Soc.* **B, 36**, 111-147.

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

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models, *Mathematical Science,* **153,** 12-18.

Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series,* Springer-Verlag, New York.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing,* Monographs on statistics and applied probability 60, Chapman and Hall, London.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space, *J. R. Statist. Soc.* **B, 64**, Part 3, 363-410.

Yonemoto, K. and Yanagawa, T. (2001). Estimating the embedding dimension and delay time of chaotic time series by an autoregressive model, *Bulletin of informatics and cybernetics,* **33, 1,** 53-62.

Yonemoto, K. and Yanagawa, T. (2004). Estimating the Lyapunov Exponent from Chaotic Time Series with Dynamic Noise, MHF Preprint Series, MHF2004-1, Faculty of Mathematics, Kyushu University, 2004.

Received October 29, 2003
Revised September 27, 2004