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

Orthogonal Projection, Embedding Dimension and Sample Size in Chaotic Time Series from a Statistical Perspective [and Discussion]

B. Cheng, H. Tong, R. J. Bhansali, P. M. Robinson and A. Kleczkowski

Phil. Trans. R. Soc. Lond. A 1994 **348**, 325-341
doi: 10.1098/rsta.1994.0094

Orthogonal projection, embedding dimension and sample size in chaotic time series from a statistical perspective†

By B. Cheng and H. Tong

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

By studying systematically the orthogonal projections, in a particular sense associated with a (random) time series admitting a possibly chaotic skeleton and in a sequence of suitably defined $\mathcal{L}_2$-spaces, we describe a geometric characterisation of the notion of embedding dimension within a statistical framework. The question of sample size requirement in the statistical estimation of the said dimension is addressed heuristically, ending with a pleasant surprise: the curse of dimensionality may be lifted except in the excessively stringent cases.

1. Introduction

In the study of chaos, substantial efforts have been expended on the detection of exotic attractors from data, which may be laboratory-based, field-based or simply computer generated (see Tong & Smith 1992; Drazin & King 1992). In most of these studies, the focus is first and foremost on the estimation of such summarizing invariants as the correlation dimension, the Lyapunov exponents, the information dimension and so on. However, as argued in Cheng & Tong (1992) (and we shall argue even more strongly later), the most natural and logical order is to estimate the dimension of the euclidean space in which the attractor resides, which we shall call the embedding state space, *before* putting the spot-light on the attractor itself. In fact, it is increasingly recognized that the dimension of this space (to be called the embedding dimension) has an impact on the estimation of the correlation dimension and the Lyapunov exponents of the attractor (see the papers cited above). Another powerful reason for not estimating first the correlation dimension of the attractor and its like is the *curse of dimensionality*: roughly k^p data points are needed to yield a useful estimate of the correlation dimension, p, of an attractor, k being a real positive constant. Indeed in one context, Smith (1988) has advised that a sample size as astronomical as 42^p is necessary.

In the same vein, Ruelle (1990) has advanced the rule that if the slope in the Grassberger–Procassia algorithm is measured over at least one decade one finds necessarily

$$\text{correlation dimension} \leqslant 2\log_{10} N, \tag{1.1}$$

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

Phil. Trans. R. Soc. Lond. A (1994) **348**, 325–341
Printed in Great Britain

N being the sample size. (Eckmann & Ruelle (1992) give a similar formula for the estimation of the Lyapunov exponent.)

However, the above *exponential* sample size requirement for the estimation of the correlation dimension seems to have led to some confusion in the literature. For example, Jensen (1992) has questioned the wisdom of adopting 4 as the order of autoregression for the annual sunspot numbers with $N = 289$ on the grounds that 'it is not possible to estimate non-parametrically a function $T_m : \boldsymbol{R}^m \to \boldsymbol{R}$ unless one has of the order 10^m data points'. He then concludes that 'for the sunspot data one has to confine oneself to $m = 1, 2$ and perhaps $m = 3$.' By and large, we agree with his first statement because we have ourselves realized the exponential sample size requirement for the estimation of T_m (Cheng & Tong 1993*b*). That is, the sample size requirements for the estimation of T_m and the correlation dimensions of the attractors defined by T_m are comparable. However, as far as the estimation of the embedding dimension is concerned, confining m to 1 or 2 is unduly pessimistic. For the moment, let us simply announce that under general conditions, the sample size requirement for the estimation of m is not exponential but (roughly) only quadratic in m.

The above announcement might sound too good to be true, especially for those brought up with the curse of dimensionality. However, a moment's reflection should convince us that estimating m is a much less demanding task than estimating either T_m or the correlation dimension of the attractors of T_m. For, we can afford to stand back from the 'microscopic details' of T_m and only concentrate on discerning the 'macroscopic clustering' of the data around some cylinder set. Moreover, the fact that m is integer-valued while the correlation dimension is real-valued reinforces the point.

The rest of the paper is organized as follows. In § 2, we describe cylinder sets associated with appropriate orthogonal projection of a time series. In § 3, we introduce a distance function over the cartesian product $\boldsymbol{N} \times \boldsymbol{N}$ as a measure of the 'goodness of fit' of nonlinear autoregression (NLAR) models of different orders. This formalizes the macroscopic clustering mentioned earlier. This distance function leads, in § 4, to a natural loss function with which consistent estimation of the embedding dimension may be obtained. Section 5 addresses the issue of sample size requirement. Section 6 reports some simulation results. Section 7 gives some concluding discussions.

2. Projections and cylinder sets

Let $\{X_t\}$ be a discrete-time stationary time series with $EX_t^2 < \infty$. Let

$$E[X_t | X_{t-1}, \ldots, X_{t-d}]$$

denote the conditional expectation of X_t given $(X_{t-1}, \ldots, X_{t-d})$. Define the residual variance by

$$\sigma^2(d) = E[X_t - E[X_t | X_{t-1}, \ldots, X_{t-d}]]^2. \tag{2.1}$$

Define the generalized partial autocorrelation function by

$$\phi(d) = \{1 - \sigma^2(d+1)/\sigma^2(d)\}^{1/2}. \tag{2.2}$$

Definition 2.1. $\{X_t\}$ is a nonlinear autoregressive process of order d_0, in short NLAR (d_0), if $\exists$ a non-negative integer $d_0 < \infty$ such that $\phi(d_0 - 1) \neq 0$

and $\phi(d) = 0$ for all $d \geqslant d_0$. If no such finite d_0 exists, then $\{X_t\}$ is said to be a nonlinear autoregressive process of infinite order, or NLAR (∞).

For the space $\boldsymbol{R}^\tau$, $\tau \geqslant 1$, define the metric $\| \ \|$ by

$$\| X - Y \| = \{(x_1 - y_1)^2 + \cdots + (x_\tau - y_\tau)^2\}^{1/2}, \tag{2.3}$$

where $X = (x_1, \ldots, x_\tau)^{\mathrm{T}}$ and $Y = (y_1, \ldots, y_\tau)^{\mathrm{T}}$. For the space $\boldsymbol{R}^{\tau+2}$, we define two types of projection $\overline{P}$ and $\underline{P}$ by

$$\overline{P}X = (x_1, \ldots, x_{\tau+1})^{\mathrm{T}} \tag{2.4}$$

and

$$\underline{P}X = (x_2, \ldots, x_{\tau+2})^{\mathrm{T}}, \tag{2.5}$$

where

$$X = (x_1, \ldots, x_{\tau+2})^{\mathrm{T}} \in \boldsymbol{R}^{\tau+2}. \tag{2.6}$$

For two random vectors $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ and $Y = (Y_1, \ldots, Y_m)^{\mathrm{T}}$, define the conditional expectation of X given Y by

$$E[X|Y] = (E[X_1|Y], E[X_2|Y], \ldots, E[X_n|Y])^{\mathrm{T}}. \tag{2.7}$$

Let $Y_t^{(d)} = (X_{t+1}, X_t, X_{t-1}, \ldots, X_{t-d})^{\mathrm{T}}$. Then the conditional expectation

$$\begin{aligned} E[\underline{P}Y_t^{(d)}|X_{t-1}, \ldots, X_{t-d}] &= (E[X_t|X_{t-1}, \ldots, X_{t-d}], X_{t-1}, \ldots, X_{t-d})^{\mathrm{T}} \\ &= (F_d(X_{t-1}, \ldots, X_{t-d}), X_{t-1}, \ldots, X_{t-d})^{\mathrm{T}}, \end{aligned} \tag{2.8}$$

where $F_d(X_{t-1}, \ldots, X_{t-d})$ denotes the conditional expectation of X_t given X_{t-1}, $\ldots$, X_{t-d}.

Lemma 2.1. ***(Projection Lemma).*** *$\{X_t\}$ is a nonlinear autoregression (d_0) if and only if $F_{d_0-1} \neq F_{d_0}$ a.s. and $F_d \equiv F_{d_0}$ a.s. for $d > d_0$.*

Proof. Trivial. ∎

Remark 2.1. Let $\varepsilon_t^{(d)}$ denote the difference $X_t - F_d(X_{t-1}, \ldots, X_{t-d})$. Then we have

$$\underline{P}Y_t^{(d)} = (F_d(X_{t-1}, \ldots, X_{t-d}), X_{t-1}, \ldots, X_{t-d})^{\mathrm{T}} + (\varepsilon_t^{(d)}, 0, \ldots, 0)^{\mathrm{T}}, \tag{2.9}$$

which is a point in the phase space $\boldsymbol{R}^{d+1}$. The larger d is, the more structure has the dynamic associated with the function F_d. However, using $\boldsymbol{R}^{d+1}$, $d > d_0$, yields no further information about the structure of the dynamic associated with F_{d_0} than using $\boldsymbol{R}^{d_0+1}$.

Example 2.1. Consider the stochastic logistic map

$$X_t = \alpha X_{t-1}(1 - X_{t-1}) + \varepsilon_t \quad (0 \leqslant \alpha \leqslant 4). \tag{2.10}$$

Here

$$Y_t^{(1)} = (X_{t+1}, X_t, X_{t-1})^{\mathrm{T}}, \quad E[\underline{P}Y_t^{(1)}|X_{t-1}] = (F_1(X_{t-1}), X_{t-1})^{\mathrm{T}}.$$

The mapping $F_1 : X_{t-1} \to X_t$ in phase space $\boldsymbol{R}^2$ is a parabola. In phase space $\boldsymbol{R}^3$, we have

$$Y_t^{(2)} = (X_{t+1}, X_t, X_{t-1}, X_{t-2})^{\mathrm{T}},$$

$$E[\underline{P}Y_t^{(2)}|X_{t-1}, X_{t-2}] = (F_2(X_{t-1}, X_{t-2}), X_{t-1}, X_{t-2})^{\mathrm{T}}.$$

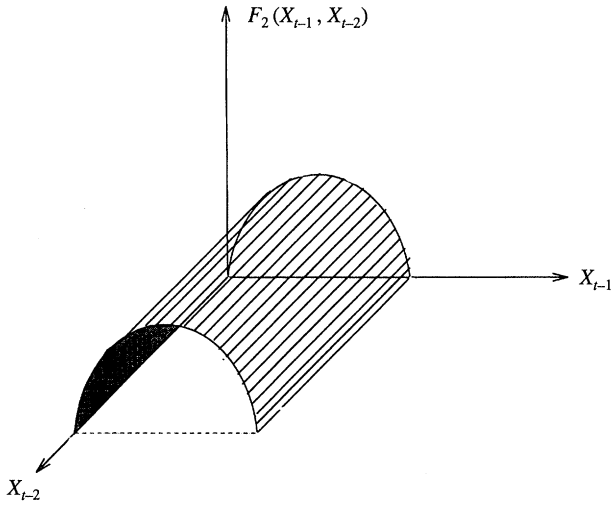

Figure 1. The mapping $F_2 : (X_{t-1}, X_{t-2})^{\mathrm{T}} \to X_t$ in phase space $\boldsymbol{R}^3$.

Comparing the dynamic defined by F_2 in phase space $\boldsymbol{R}^3$ with that defined by F_1 in phase space $\boldsymbol{R}^2$, we can clearly see the cylinder set structure. We shall see that cylinder sets indicate redundant information.

3. Distance function and embedding dimension

Let $Z_t^{(d)} = (X_{t-1}, \ldots, X_{t-d})$ and recall that $E[X_t|Z_t^{(d)}]$ is denoted by $F_d(Z_t^{(d)})$. Define

$$\mathcal{L}_2(Z_t^{(d)}) = \{F(Z_t^{(d)}) | F \text{ is measurable on } \boldsymbol{R}^d \text{ such that } E|F(Z_t^{(d)})|^2 < \infty\}.$$

Then

$$\mathcal{L}_2(Z_{t-1}^{(1)}) \subset \mathcal{L}_2(Z_t^{(2)}) \subset \cdots \subset \mathcal{L}_2(Z_t^{(d)}) \subset \cdots$$

and $F_d(Z_t^{(d)})$ is the orthogonal projection of X_t in $\mathcal{L}_2(Z_t^{(d)})$. For integers $0 < d_1 \leqslant d_2$, we have

$$F_{d_1}(Z_t^{(d_1)}) \in \mathcal{L}_2(Z_t^{(d_1)}) \subset \mathcal{L}_2(Z_t^{(d_2)})$$

and

$$F_{d_2}(Z_t^{(d_2)}) \in \mathcal{L}_2(Z_t^{(d_2)}).$$

Therefore we may consider the closeness between $F_{d_1}(Z_t^{(d_1)})$ and $F_{d_2}(Z_t^{(d_2)})$ in $\mathcal{L}_2(Z_t^{(d_2)})$ by

$$\Delta(d_1, d_2) = E[F_{d_1}(Z_t^{(d_1)}) - F_{d_2}(Z_t^{(d_2)})]^2. \tag{3.1}$$

Notice that F_d is uniquely determined once d is given. Thus $\Delta(.,.)$ is well defined and $\Delta(d_1, d_2)$ is the (squared) distance between the orthogonal projection of X_t in $\mathcal{L}_2(Z_t^{(d_1)})$ and that in $\mathcal{L}_2(Z_t^{(d_2)})$.

Definition 3.1. The time series $\{X_t\}$ is said to have embedding dimension d_0 ($d_0 \geqslant 1$) if and only if

$$\begin{array}{lll} \text{(i)} & \Delta(d, d_0) \neq 0 & \text{for all} \quad d < d_0, \\ \text{(ii)} & \Delta(d, d_0) = 0 & \text{for all} \quad d \geqslant d_0. \end{array}$$

Proposition 3.1.

(i) $\Delta^{1/2}(.,.)$ *is a properly defined distance function on* $\boldsymbol{N} \times \boldsymbol{N}$, *i.e.*

$\Delta^{1/2}(d_1, d_2) = \Delta^{1/2}(d_2, d_1)$,

$\Delta^{1/2}(d, d) = 0$,

$\Delta^{1/2}(d_1, d_3) \leqslant \Delta^{1/2}(d_1, d_2) + \Delta^{1/2}(d_2, d_3)$.

(ii) If for each $d \geqslant 1$, F_d *has bounded first partial derivatives on* $\boldsymbol{R}^d$, *then*

$$\Delta(d_2, d_1) \leqslant c|d_2 - d_1|,$$

where c *is a constant.*

(iii) For $d_1 \leqslant d_2 \leqslant d_3$,

$$\Delta(d_2, d_3) \leqslant \Delta(d_1, d_3).$$

That is for fixed d_3, $\Delta(d, d_3)$ *is a decreasing function in* d.

(iv) For any $d_1 \leqslant d_2$, *we have*

$$\Delta(d_1, d_2) = \sigma^2(d_1) - \sigma^2(d_2). \tag{3.2}$$

(v) $\sum_{d=1}^{\infty} \Delta(d, d+1) < \infty$.

(vi) There are infinitely many d *for which*

$$\Delta(d, d+1) \leqslant \tilde{\kappa}/d,$$

where $\tilde{\kappa}$ *is a constant.*

(vii) $\forall d \leqslant D < \infty, \exists \kappa_D$, $0 < \kappa_D < \infty$, *such that*

$$\Delta(d, d+1) \leqslant \kappa_D/d.$$

Proof.

(i) The first two assertions are obvious. The third assertion follows from the Minkowski's inequality. Without loss of generality, let $d_1 \leqslant d_3$. If $d_2 \leqslant d_1$, then from (iii) we have $\Delta(d_1, d_3) \leqslant \Delta(d_2, d_3)$ and the assertion is true.
There are two cases left: $d_2 \geqslant d_3 \geqslant d_1$ and $d_3 \geqslant d_2 \geqslant d_1$. For the former case, the assertion follows from (iii). For the latter case, the assertion follows from (iv).

(ii) Suppose that $d_1 \leqslant d_2$. Then

$$\Delta(d_1, d_2) = E\{F_{d_2}(Z_t^{(d_2)}) - E[F(Z_t^{(d_2)})|Z_t^{(d_1)}]\}^2.$$

Let

$$\zeta = F_{d_2}(Z_t^{(d_1)}, \underbrace{0, \ldots, 0}_{d_2 - d_1}).$$

Obviously, $\zeta \in \mathcal{L}_2(Z_t^{(d_2)})$. Since $E[F_{d_2}(Z_t^{(d_2)})|Z_t^{(d_1)}]$ is the orthogonal projection of $F_{d_2}(Z_t^{(d_2)})$ in $\mathcal{L}_2(Z_t^{(d_1)})$, we have

$$\Delta(d_1, d_2) \leqslant E[F_{d_2}(Z_t^{(d_2)}) - \zeta]^2 = E[G(x^*_{d_1+1})X_{t-d_1-1} + \cdots + G(x^*_{d_2})X_{t-d_2}]^2,$$

where $G(x_j^*)$ denotes the partial derivative of F_{d_2} with respect to the jth component and evaluated at x_j^*. Now,

$$E\left[\sum_{j=d_1+1}^{d_2} G(x_j^*)X_{t-j}\right]^2 \leqslant \text{const.} \times E\left[\sum_{j=d_1+1}^{d_2} G^2(x_j^*)X_{t-j}^2\right]$$
$$\leqslant \text{const.} \times (d_2 - d_1),$$

and hence (ii) is proved.

(iii)

$$\Delta(d_2, d_3) = E[F_{d_3}(Z_t^{(d_3)}) - E[F_{d_3}(Z_t^{(d_3)})|Z_t^{(d_2)}]]^2.$$

Since $F_{d_1}(Z_t^{(d_1)}) \in \mathcal{L}_2(Z_t^{(d_2)})$ and $E[F_{d_3}(Z_t^{(d_3)})|Z_t^{(d_2)}]$ is the orthogonal projection of $F_{d_3}(Z_t^{(d_3)})$ in $\mathcal{L}_2(Z_t^{(d_2)})$, therefore

(iv)
$$\Delta(d_2, d_3) \leqslant E[F_{d_3}(Z_t^{(d_3)}) - F_{d_1}(Z_t^{(d_1)})]^2 = \Delta(d_1, d_3).$$

$$\begin{aligned}\Delta(d_1, d_2) &= E[\{X_t - F_{d_1}(Z_t^{(d_1)})\} - \{X_t - F_{d_2}(Z_t^{(d_2)})\}]^2 \\ &= E[\varepsilon_t^{(d_1)}]^2 + E[\varepsilon_t^{(d_2)}]^2 - 2E[\varepsilon_t^{(d_1)}\varepsilon_t^{(d_2)}].\end{aligned}$$

However,

$$E[\varepsilon_t^{(d_1)}\varepsilon_t^{(d_2)}] = E[\varepsilon_t^{(d_2)}\{\varepsilon_t^{(d_2)} + F_{d_2}(Z_t^{(d_2)}) - F_{d_1}(Z_t^{(d_1)})\}] = E[\varepsilon_t^{(d_2)}]^2$$

and hence the result.

(v) Define
$$e_t^{(d)} = F_d(Z_t^{(d)}) - F_{d+1}(Z_t^{(d+1)}).$$

It is easy to check that for $d' \neq d$

$$E[e_t^{(d')}e_t^{(d)}] = 0.$$

We have therefore the following orthogonal decomposition

$$E[X_t|X_{t-1}] - E[X_t|X_{t-1}, X_{t-2}, \ldots] = \sum_{d=1}^{\infty} e_t^{(d)}.$$

It follows from this and the finiteness of EX_t^2 that

$$\sum_{d=1}^{\infty} E[e_t^{(d)}]^2 < \infty.$$

This implies that
$$\sum_{d=1}^{\infty} \Delta(d, d+1) < \infty.$$

(vi) If there are only finitely many d for which

$$\Delta(d, d+1) = 0(1/d),$$

then $\exists$ a d^* and a constant $C > 0$ such that $\forall d \geqslant d^*$

$$\Delta(d, d+1) \geqslant C/d.$$

This implies that

$$\sum_{d=1}^{\infty} \Delta(d, d+1) \geqslant \sum_{d \geqslant d^*} \Delta(d, d+1) \geqslant C \sum_{d \geqslant d^*} (1/d) = \infty.$$

This is a contradiction and therefore (ii) is proved.

(vii) Trivial.

■

Note that the bound in (vi) is almost sharp because, for example,

$$\sum_{d=2}^{\infty} \frac{1}{d((\ln(d)))^2} < \infty \quad \text{and} \quad \sum_{d=1}^{\infty} \frac{1}{d^{1-\epsilon}} = \infty,$$

where $\epsilon > 0$.

Note that for continuous parameters such as the bandwidth parameters in kernel smoothing, we may use the euclidean norm as an appropriate distance function for parameter (e.g. the bandwidth) choice. However, for many discrete cases, the euclidean norm is found to be unsuitable. For example, Akaike (1974) has used instead the Kullback–Leibler information to construct a suitable distance function for order determination. For our case, we have obtained an appropriate distance function, namely $\Delta^{1/2}(.,.)$ on $\boldsymbol{N} \times \boldsymbol{N}$, based on the projection of the skeleton from a low dimension to a high dimension as described in § 2. Proposition 3.1(ii) reveals the relation between $\Delta(.,.)$ and the euclidean norm.

Suppose that henceforth $\{X_t\}$ is bounded (cf. Chan & Tong 1994). Let $d_2 \geqslant d_1$. We write

$$\Delta_N(d_1, d_2) = \sigma_N^2(d_1) - \sigma_N^2(d_2), \tag{3.3}$$

where

$$\sigma_N^2(d) = (N - r + 1)^{-1} \Sigma(\varepsilon_t^{(d)})^2, \tag{3.4}$$

the summation extending over $t \geqslant r$ and r being a positive integer $\geqslant \max(d_1, d_2)$.

Trivially, $E\Delta_N(d_1, d_2) = \Delta(d_1, d_2)$.

Proposition 3.2.

(a) The embedding dimension and the order of autoregressive models of $\{X_t\}$ *coincide.*

(b) Let d_0 $(\geqslant 1)$ *be the embedding dimension for* $\{X_t\}$. *Then the following hold:*

(i) if $d < d_0$ *then* $\Delta_N(d, d_0) \neq 0$ *except for finitely many* N;

(ii) if $d \geqslant d_0$ *then* $\Delta_N(d_0, d) = 0$.

Proof. (*a*) This is just a simple corollary of Proposition 3.1(iv).

(*b*) For $d \geqslant d_0$,

$$F_d(Z_t^{(d)}) = F_{d_0}(Z_t^{(d_0)}).$$

Therefore,

$$\varepsilon_t^{(d)} = \varepsilon_t^{(d_0)},$$

and thus (ii) is proved.

For $d < d_0$, if for all but finitely many N

$$\Delta_N(d, d_0) = 0,$$

then, letting $N \to \infty$, we have, by any standard ergodic theorem,

$$0 = \lim_{N \to \infty} \Delta_N(d, d_0) = \sigma^2(d) - \sigma^2(d_0) = \Delta(d, d_0).$$

This contradicts the definition of embedding dimension. Hence, (i) is proved. ■

4. Loss functions for estimations of the embedding dimension

The squared distance function $\Delta(.,.)$ as defined by (3.1) naturally suggests the following loss function as a measure of goodness of fit

$$ASE(d) = N^{-1}\sum_{t=1}^{N}[F_{d,N}(Z_t^{(d)}) - F_{d_0}(Z_t^{(d_0)})]^2, \tag{4.1}$$

where d_0 is the embedding dimension and $F_{d,N}(Z_t^{(d)})$ is the Nadaraya–Watson kernel estimate of $E[X_t|Z_t^{(d)}]$ based on the observations $(X_1,\ldots,X_N)$. Loss functions based on similar motivation have been used by Akaike (1974) and Shibata (1980) for the case where F is linear. Manipulations of U-statistics as in Cheng & Tong (1992, 1993*a*) establish the following theorem. (Details of the proof are similar to the references just cited and are therefore omitted.)

Theorem 4.1. *Under the same conditions as in Theorem 1 of Cheng & Tong (1992),*

$$ASE(d) = \tilde{\Delta}_N(d,d_0) - \{2\alpha(d)\tilde{\Delta}(d,d_0) - \beta(d)\sigma_N^2(d)\}(Nh_{d,N}^d)^{-1} + \circ_p((Nh_{d,N}^d)^{-1}), \tag{4.2}$$

where

$$\tilde{\Delta}_N(d,d_0) = N^{-1}\sum_{t=1}^{N}\{F_d(Z_t^{(d)}) - F_{d_0}(Z_t^{(d_0)})\}^2,$$

$$\tilde{\Delta}(d,d_0) = E[\{F_d(Z_t^{(d)}) - F_{d_0}(Z_t^{(d_0)})\}^2/f(Z_t^{(d)})],$$

$$\alpha(d) = \{k(0)\}^d,$$

$$\beta(d) = \left\{\int k^2(u)\,\mathrm{d}u\right\}^d,$$

and where f is the probability density function of $Z_t^{(d)}$, k is the kernel (function) as defined in Cheng & Tong (1992) and $h_{d,N}$ is the bandwidth of the kernel (previously denoted as $B(N)$ in Cheng & Tong (1992)).

In a sense, the above result gives an affirmative answer to the question raised by one of us (B.C.) in his discussion of Hall & Johnstone (1992).

Theorem 4.2. *Under the same conditions as in Theorem 4.1,*

$$\lim_{N\to\infty} P\{\hat{d} = d_0\} = 1,$$

where $\hat{d}$ is the minimizer of $ASE(d)$.

Proof. We imitate the proof of Theorem 2 of Cheng & Tong (1992).

For $d < d_0$, $ASE(d)$ is dominated by $\tilde{\Delta}_N(d,d_0)$, which is non-zero. For $d > d_0$, $\tilde{\Delta}_N(d,d_0) = 0$ and $\tilde{\Delta}(d,d_0) = 0$. Therefore,

$$Nh_{d,N}^d\{ASE(d) - ASE(d_0)\} = c(d) + \circ_p(1),$$

where $c(d) > 0$. The conclusion of the theorem follows immediately. ■

5. Sample size requirement for consistent estimation of embedding dimension

$ASE(d)$ is useful for theoretical discussion but it involves the unknown d_0. To make the notion of distance practically useful for the determination of embedding dimension, we need to remove d_0 and one way to do this is by a comparative approach. Returning to the basic squared distance function $\Delta(.,.)$, we may solve the problem systematically by noting that Part (iv) of Proposition 3.1 together with Proposition 3.2 suggests an obvious consistent estimate of $\Delta(d_1, d_2)$, $d_1 \leqslant d_2$, namely

$$\hat{\Delta}(d_1, d_2) = RSS(d_1) - RSS(d_2), \tag{5.1}$$

where, as in Cheng & Tong (1992, eqn (2.5)) but setting $W(x) \equiv 1$ and assuming boundedness for the time series),

$$RSS(d) = (N - r + 1)^{-1} \sum \{X_t - \hat{F}_{d,N}(Z_t^{(d)})\}^2. \tag{5.2}$$

the sumation extending over $t \geqslant r$ and $\hat{F}_{d,N}$ is the kernel estimator of F_d.

We now appeal to some results in Cheng & Tong (1992, 1993*a*). Before doing so, we need to first of all simplify conditions (*a*)–(*o*) of Theorem 1 of their paper in 1992 for convenience. Specifically, we retain conditions (*a*)–(*i*), and replace the rest by the existence of bounded first partial derivatives of F_d for each d and

(j') $\quad \beta_j = 0(\beta^j), \quad 0 < \beta < 1$;

(k') $\quad h_{d,N} \in H_{d,N} = [aN^{-(1/(2d+1))-\xi}, \quad bN^{-(1/(2d+1))+\xi}]$,

where a and b are arbitrary real positive constants, ξ is any real positive constant strictly less than $\{2(d+1)(2d+1)\}^{-1}$. The detailed justification is available in the Technical Report no. UKC/IMS/S93/6a. We shall refer to the above simplified conditions collectively as Condition B.

From Theorem 3 of Cheng & Tong (1992), we have for each $d \geqslant 1$ and $h_{d,N} \in H_{d,N}$

$$RSS(d) = \sigma_N^2(d)\{1 - (2\alpha(d) - \beta(d))/Nh_{d,N}^d\} + \circ_p(1/Nh_{N,d}^d). \tag{5.3}$$

Choose, for explicitness,

$$h_{d,N} = N^{-1/(2d+1)},$$

and write $Nh_{d,N}^d = N^{1/\eta(d)}$, where $\eta(d) = (2d+1)/(d+1)$. Note that (5.1) and (5.3) yield, for general d_1 and d_2,

$$\hat{\Delta}(d_1, d_2) = \Delta_N(d_1, d_2) + \circ_p(N^{-1/2}).$$

By standard arguments, we have

$$\Delta_N(d_1, d_2) = \Delta(d_1, d_2) + O_p(N^{-1/2}).$$

Now, let d_0 denote the embedding dimension. Then we have from (5.3)

$$\begin{aligned}\hat{\Delta}(d_0, d_0+1) &= RSS(d_0) - RSS(d_0+1) \\ &= \sigma_N^2(d_0) - \sigma_N^2(d_0+1) - \sigma_N^2(d_0)\{2\alpha(d_0) - \beta(d_0)\}N^{-1/\eta(d_0)} \\ &\quad + \sigma_N^2(d_0+1)\{2\alpha(d_0+1) - \beta(d_0+1)\}N^{-1/\eta(d_0+1)} \\ &\quad + \circ_p(N^{-1/\eta(d_0)}) + \circ_p(N^{-1/\eta(d_0+1)}).\end{aligned} \tag{5.4}$$

However, from Proposition 3.2,

$$\sigma_N^2(d_0) - \sigma_N^2(d_0+1) = \Delta_N(d_0, d_0+1) = 0.$$

Now

$$N^{-1/\eta(d_0+1)} \div N^{-1/\eta(d_0)} \to \infty \quad \text{as} \quad N \to \infty,$$

implying that

$$N^{-1/\eta(d_0)} = \circ(N^{-1/\eta(d_0+1)}).$$

Hence, we have proved the following theorem.

Theorem 5.1. *Under Condition B and the boundedness of the time series,*

$$\hat{\Delta}(d_0, d_0+1) = \sigma_N^2(d_0+1)\{2\alpha(d_0+1) - \beta(d_0+1)\}N^{-1/\eta(d_0+1)} + \circ_p(N^{-1/\eta(d_0+1)}). \tag{5.5}$$

We now use Theorem 5.1 to throw some light on the sample size requirement for a consistent estimation of d_0. The derivation will be heuristic. The embedding dimension is characterized by $\Delta^{1/2}(d_0, d_0+1)$, which measures the distance between the projection of the skeleton F_{d_0} in R^{d_0+1} and the skeleton F_{d_0+1}. This was summarized in Lemma 2.1 and Proposition 3.2. Since $\Delta^{1/2}$ is unknown, we use its consistent estimate $\hat{\Delta}^{1/2}$. Then Proposition 3.1(v)–(vii) suggests that for each d it is reasonable to set

$$\hat{\Delta}(d, d+1) = \Delta(d, d+1) + O_p(N^{-1/2}) \approx \text{const.}/d + O_p(N^{-1/2}). \tag{5.6}$$

However, Theorem 5.1 yields

$$\hat{\Delta}(d_0, d_0+1) = \text{const.} \times \rho^2(d_0)N^{-1/\eta(d_0+1)} + \circ_p(N^{-1/\eta(d_0+1)}), \tag{5.7}$$

where $\rho^2(d_0) = \sigma^2(d_0)/\text{var}(X_t)$, the normalized dynamic noise variance. Therefore, combining (5.6) and (5.7), we have up to $O_p(N^{1/2})$

$$N \approx (d_0\rho^2(d_0)/\kappa)^{\eta(d_0+1)} \leqslant \{d_0\rho^2(d_0)/\kappa\}^2. \tag{5.8}$$

We will discuss the constant κ and the term $O_p(N^{1/2})$ in the next section. Cheng & Tong (1992) have proved that by relying on a penalized form of $\tilde{\Delta}_N(d, d+1)$ and $\hat{\Delta}(d, d+1)$, namely the CV criterion (the definition will be recalled in (6.2)), a consistent estimate of d_0 may be obtained. Note that the difference between the CV criterion (or more precisely $\{CV(d) - CV(d+1)\}$) and $\hat{\Delta}(d, d+1)$ is $O_p(1/(Nh_{d,N}^d))$, which is only $\circ_p(N^{-1/2})$. Hence, as a practical guidance we may offer the advice that for useful estimation of the embedding dimensions the sample size requirement is bounded by a constant multiple of {(embedding dimension) × (normalized dynamic noise variance)}^{2}. First, note the presence of the dynamic noise variance. This is in contrast to the results for the estimation of correlation dimension obtained by Smith (1988) and Ruelle (1990), which deal with the noise-free case. Next, by far the more significant is the lifting of the curse of dimensionality! This also throws substantial light on the 'better-than-originally-expected' simulation results reported in Tong (1994) and Yao & Tong (1994). Of course, the curse stays if we replace const./d in (5.6) by (const.)$^{-d}$. However, Proposition 3.1(v)–(vii) suggests that the latter is excessively stringent. In fact, the same stringency will lead to the same curse even in the linear case, which is discussed in the next section.

By now, the case for determining the embedding dimension first before focusing on the attractors is clearly overwhelming not only on grounds of logic but also statistical soundness. Once the embedding dimension is determined, or even better once a parsimonious set of stochastic regressors is determined using, for example, the approach of Cheng & Tong (1992) and Yao & Tong (1994), we can set about searching for exotic attractors within this set. We can then face the horrendous task of map reconstruction and correlation dimension estimation, etc., with perhaps a better chance. Of course, for these, the curse of dimensionality returns now unless parametric models are used, but then the perennial problem of subjectivity returns.

We now summarize some of our simulation results with a view to suggesting some preliminary empirical guidance for sample size requirement in the estimation of embedding dimension.

6. Simulations

Notice that the requirement of sample size given by formula (5.8) is bounded by a constant multiple of {(embedding dimension) × (normalized dynamic noise variance)}^{2}, which is independent of the forms of the skeleton. As a typical illustration of our simulations with $d \geqslant 1$, we have used the following nonlinear models:

$$X_t = 0.1X_{t-1} + (-0.5 + 0.2\exp[-0.1X_{t-d}^2])X_{t-d} + \epsilon_t, \quad d = 1, 2, \ldots, \tag{6.1}$$

where $\{\epsilon_t\}$ are independent random variables with mean zero and variance 0.1.

Following Cheng & Tong (1992), for each d_0, we estimate d_0 consistently by minimizing the cross-validation criterion with respect to d:

$$CV(d) = (N - r + 1)^{-1} \sum \{X_t - \hat{F}_{d,N,-t}(Z_t^{(d)})\}^2, \tag{6.2}$$

where $\hat{F}_{d,N,-t}(Z_t^{(d)})$ is the *leave-one-out* estimator of F_d at $Z_t^{(d)}$ (see Cheng & Tong (1992) for details). In our simulations, $h \in H_{d,N}$ with $h = c \times N^{-1/(2d+1)}$. For different d and N, the constant c was adjusted between 1 and 10. In principle, a data-driven bandwidth may be preferred but the computations involved would be quite excessive relative to the computing power at our dispoal. Nevertheless our random checks suggest that the results are unlikely to be fundamentally different from those summarized in table 1. We set the true orders at 4, 8, 11 and 13. For each sample size N, $CV(d)$ was searched over d from 1 to 20 and the estimated order, $\hat{d}_{CV}$, was given by

$$\hat{d}_{CV} = \arg \min_{1 \leqslant d \leqslant 20} \{CV(d)\}.$$

We used the NAG library (G05DDF and G05CBF) to generate independent samples of size N and 100 replications were generated for each d. All computations were run in a SUN SPARC 2 workstation. The results of the simulations are summarized in the following tables.

Now, we define the frequency of 'success' for the true order d using sample size N, $\mathrm{FREQ}_N(d)$ say, by

$$\mathrm{FREQ}_N(d) = \frac{\#(\hat{d}_{CV} = d) + \#(\hat{d}_{CV} = d + 1)}{100}.$$

Table 1. *Frequencies of estimated order (true order $d = 4, 8, 11, 13$)*

	true order $d = 4$			true order $d = 8$			true order $d = 11$			true order $d = 13$		
$\hat{d}_{CV}$	$N = 100$	300	500	100	500	850	100	500	1000	550	750	1250
1	3	0	0	21	0	0	18	0	0	3	0	1
2	0	0	0	8	0	0	4	0	0	1	0	1
3	3	0	0	5	0	0	1	1	0	2	0	0
4	31	58	91	0	0	0	5	0	0	0	0	1
5	8	14	4	2	0	0	2	0	0	1	0	0
6	6	8	1	1	0	0	8	0	0	1	0	1
7	6	5	0	1	0	0	1	0	0	1	0	0
8	9	3	0	23	60	81	1	0	0	0	0	0
9	5	1	0	12	18	14	1	0	0	0	0	0
10	3	0	0	7	5	5	0	0	0	0	0	0
11	4	2	0	6	5	0	20	60	86	1	0	0
12	4	3	1	3	5	0	13	14	11	0	0	0
13	3	4	0	6	1	0	4	14	1	56	57	88
14	0	0	1	0	2	0	2	6	1	23	31	7
15	2	1	0	1	0	0	6	2	0	4	8	1
16	6	1	1	3	2	0	2	1	0	5	2	0
17	1	0	0	1	2	0	1	1	0	0	2	0
18	1	0	0	0	0	0	2	1	0	0	0	0
19	3	0	0	0	0	0	2	0	0	2	0	0
20	2	0	1	0	0	0	7	0	0	0	0	0

Then, as a typical example, from table 1, $\mathrm{FREQ}_{500}(4) = (91 + 4)/100 = 95\% = 0.95$.

If we define the sample size requirement for order d, $N_{\mathrm{req}}(d)$, by

$$\mathrm{FREQ}_{N_{\mathrm{req}}(d)}(d) \geqslant 95\%,$$

i.e. there is at least 95% 'success' when sample size $N_{\mathrm{req}}(d)$ is used, then from tables 1–4, we know that

$$N_{\mathrm{req}}(4) = 500; \qquad N_{\mathrm{req}}(8) = 850; \tag{6.3}$$

$$N_{\mathrm{req}}(11) = 1000; \qquad N_{\mathrm{req}}(13) = 1250. \tag{6.4}$$

From Proposition 3.1(i) we know that $\Delta^{1/2}$ is a distance function and if we bound $\hat{\Delta}^{1/2}(d, d+1)$ by $1 - 95\% = 5\% = 0.05$, i.e. bounded by the 'failure rate', then formula (5.8) becomes

$$N_{\mathrm{req}}(d) = N_0 + \{d\rho^2(d)/\kappa\}^{\eta}, \tag{6.5}$$

where we simply take N_0 and η as constants, note that $\rho^2(4)$, $\rho^2(8)$, $\rho^2(11)$ and $\rho^2(13)$ are all approximately 0.1 and $\kappa = (0.05)^2$. We may interpret N_0 as the 'baseline' sample size, which seems to be related to the term $O_p(N^{1/2})$ in (5.8). Since $\rho^2(d)/\kappa = 40$, we obtain

$$N_{\mathrm{req}}(d) = N_0 + \{40d\}^{\eta}, \tag{6.6}$$

Since $N_{\mathrm{req}}(4) = 500$ and $N_{\mathrm{req}}(8) = 850$, it is easy to see that N_0 is between 120

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

and 240 (we allow ±50 oscillation here) and $\eta \approx 1.12$. This leads to the formula

$$N_{\text{req}}(d) = [120, 240] + \{40d\}^{1.12}, \tag{6.7}$$

where $[120, 240]$ means some integer between 120 and 240. Now we use this formula to predict the sample size requirements for $d = 11$ and $d = 13$. We get

$$N_{\text{req}}(11) = [120, 240] + \{40 \times 11\}^{1.12} = [1033, 1153],$$

$$N_{\text{req}}(13) = [120, 240] + \{40 \times 13\}^{1.12} = [1221, 1341].$$

These rough-and-ready arguments seem to give quite encouraging results by reference to those in (6.4). Generally we would propose the empirical formula

$$N_{\text{req}}(d) = N_0 + \{\gamma d\}^{\eta}, \tag{6.8}$$

where $\eta \leqslant 2$ and

$$\gamma = \rho^2(d)/[\text{'failure rate'}]^2.$$

Inverting (6.8) gives us an analogue of Ruelle's formula (1.1):

$$\text{embedding dimension} \leqslant \frac{\sqrt{N} \times (\text{'failure rate'})^2}{\text{normalized dynamic noise variance}}. \tag{6.9}$$

We suggest that the denominator be estimated (albeit roughly) by fitting a linear model of low order first or by the so-called 'noise floor' of a principal component analysis (Broomhead & King 1986). Other methods have also been suggested (see, for example, Szpiro 1993). It is interesting to note that for the annual sunspot numbers, if we are prepared to tolerate a failure rate of $20 - 25\%$ and accept the normalized dynamic noise variance at 15% (many reported parametric models for the data set have a lower value), then formula (6.9) gives an upper bound of about 6. The cross-validatory choice of order 4 by Cheng & Tong (1992) has a normalized dynamic noise variance of 15% , which is consistent with a failure rate of about 20%.

Using eqn (7) of Hannan & Quinn (1979), we may deduce that for the *linear* autoregressive model,

$$\hat{\Delta}(d_0, d_0 + 1) \approx \rho^2(d_0) \ln \ln(N)/N. \tag{6.10}$$

Following the same derivation as (5.7), we get

$$\rho^2(d_0) \ln \ln(N)/N \approx \kappa_1/d_0. \tag{6.11}$$

Repeating the same arguments as before we may then obtain an empirical formula analogous to (6.3) but with $\eta = 1$, leading to the formula:

$$\text{order of linear autoregression} \leqslant \frac{N \times (\text{'failure rate'})^2}{\text{normalized dynamic noise variance}}. \tag{6.12}$$

7. Discussion

Our approach to the detection of chaos in real time series in the absence of any substantive theory is quite different from those often found in the literature of the physical sciences. Let us summarize our position briefly. As indicated in Tong (1990), Cheng & Tong (1992) and Chan & Tong (1994), we use the class of

nonlinear autoregressive models as our basic framework and start by estimating the order of the autoregressive model for the data at hand by the cross-validation approach of leaving-k-out. We often set $k = 1$ but for over-sampled data we might prefer to set k much greater than 1 (cf. Cheng & Tong 1993*b*). We are now convinced that the sample size requirement is normally not excessive and an empirical formula is available for its evaluation. Indeed, a parsimonious set of stochastic regressors may be identified by using an optimal subset selection based on the cross-validation approach as confirmed in Yao & Tong (1994). We may then perform tests for linearity with respect to the selected regressors using an assortment of those described in Tong (1992, ch. 5) for example. If linearity is rejected, then the next job is the fitting of an appropriate nonlinear function of the selected regressors. For this a number of techniques are avaiable and they fall roughly into two groups: local function approximation and global function approximation. The former includes the threshold models of Tong (1990) and their younger relatives due to Casdagli *et al.* (1991) and Lewis & Stevens (1991), and others; the latter includes the polynomial autoregressive models (see Cox 1977; Chan & Tong 1994) and others. Casdagli *et al.* (1991) gives some details. It is probably fair to say that map reconstruction using noisy data is still an open problem. At present, which function approximation/map reconstruction we use for noisy data is often a trade-off between the curse of dimensionality and subjectivity.

Once a model has been fitted, we may examine its skeleton (= signal) with the aim of detecting chaos by evaluating its correlation dimension, Lyapunov spectrum, etc. Fortunately the curse of dimensionality does not necessarily apply here. In short, we prefer to place chaos detection within the context of signal extraction, the signal being the skeleton which may be a limit point, a limit cycle or a strange attractor.

This research was partly supported by an SERC grant under the Complex Stochastic Systems Initiative.

References

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Autom. Control* **AC-19**, 716–723.

Broomhead, D. S. & King, G. P. 1986 Extracting qualitative dynamics from experimental data. *Physica* D **20**, 217–236.

Casdagli, M., Jardins, D. D., Eubank, S., Farmer, J. D., Gibson, J., Hunter, N. & Theiler, J. 1991 Nonlinear modeling of chaotic time series: theory and applications. *Tech. Rep.* LA-UR-91-1637. Los Alamos National Laboratory, U.S.A.

Chan, K. S. & Tong, H. 1994 A note on noisy chaos. *Jl R. statist. Soc.* B **56**, 301–312.

Cheng, B. & Tong, H. 1992 On consistent nonparametric order determination and chaos. *Jl R. statist. Soc.* B **54**, 427–450.

Cheng, B. & Tong, H. 1993*a* On residual sums of squares in nonparametric autoregression. *Stoch. Proc. Applic.* **21**.

Cheng, B. & Tong, H. 1993*b* Nonparametric function estimation in noisy chaos. *Developments in time series analysis* (ed. T. Subba Rao). London: Chapman and Hall.

Cox, D. R. 1977 Discussion of paper by Campbell and Walker, Tong and Morris. *Jl R. statist. Soc.* A **140**, 453–454.

Drazin, P. G. & King, G. P. 1992 *Interpretation of time series from nonlinear systems.* In *Proc. IUTAM Symp. and NATO Advanced Research Workshop, Warwick, U.K.* Amsterdam: North Holland.

Eckmann, J.-P. & Ruelle, D. 1992 Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. *Physica* D **56**, 185.

Hall, P. & Johnstone, I. 1992 Empirical functionals and efficient smoothing parameter selection (with discussion). *Jl R. statist. Soc.* B **54**, 475–530.

Hannan, E. J. & Quinn, B. G. 1979 The determination of the order of an autoregression. *Jl R. statist. Soc.* B **41**, 190–195.

Jensen, J. L. 1992 Chaotic dynamical systems with a view towards statistics – a review. *Res. Rep.* no. 245, March 1992. Department of Theoretical Statistics, University of Aarhus, Denmark.

Lewis, P. A. W. & Stevens, J. G. 1991 Nonlinear modeling of time series using multivariate adaptive regression splines (MARs). *JASA* **86**, 864–877.

Ruelle, D. 1990 Deterministic chaos: the science and the fiction. *Proc. R. Soc. Lond.* A **427**, 241–248.

Shibata, R. 1980 Asymptotically efficient selction of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147–164.

Smith, L. A. 1988 Intrinsic limits on dimension calculations. *Phys. Lett.* A **133**, 283–288.

Szpiro, G. G. 1993 Measuring dynamical noise in dynamical systems. *Physica* D **65**, 289–299.

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

Tong, H. 1994 Akaike's approach can yield consistent order determination. In *Frontiers of statistical modelling: an informational approach* (ed. H. Bozdogan), pp. 93–103. Kleuwer Academic.

Tong, H. & Smith, R. L. 1992 Royal Statistical Society Meeting on Chaos. *Jl R. statist. Soc.* B **54**, 301–474.

Yao, Q. & Tong, H. 1994 On subset selection in nonparametric stochastic regression. *Statistica Sinica* **4**, 51–70.

Discussion

R. J. Bhansali (*Department of Statistics and Computational Mathematics, University of Liverpool, U.K.*). Professor Tong has given a bound for the sample size for obtaining a reliable estimator of the embedding dimension. It seems to me, however, that one should perhaps attempt also to improve the performance of the estimation procedure so as to ensure that it works better for small samples of the size often encountered in practice. He has described a cross-validation approach which leaves one observation out at a time. In the context of model selection, Stone (1977) has examined relationship between Akaike's information criterion, AIC, and this version of the cross-validation approach and established their asymptotic equivalence. There has recently been some work on extending the standard cross-validation approach in which $k > 1$ observations are deleted; thus Zhang (1993) has considered this approach to model selection and examined its relationsip with the FPE and AIC criteria of Bhansali & Downham (1977) and Akaike (1979). I appreciate that the authors are considering a different and more complex problem, but feel that one should perhaps draw on the developments occurring in cognate areas so as to improve the current estimation procedure. Perhaps the authors could comment on the relevance of the multifold cross-validation approach for the particular problem they consider.

H. Tong. Sample size is a fundamental consideration of chaos study of real data; this is the problem on which our paper has focused. However, Dr Bhansali's comments seem to address issues beyond the scope of our present paper, although the general answer to many of which may be found in our papers listed in the references. As is the case with most statistical techniques, there is always room

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

for improvement, but we are unclear as to which aspects he has in mind. (We are, of course, aware of his references, especially his numerous contributions to the problem of order determination in the case of linear models.) For example, we have elsewhere (Cheng & Tong 1992) proved that, under general conditions, minimizing a sensibly penalized residual sum of squares (RSS) yields a *consistent* estimator of d_0 in the context of non-parametric nonlinear autogression of bounded time series. This was a surprising result! By contrast, a similarly penalized RSS, such as the FPE or the AIC referred to by Dr Bhansali, generally leads to an inconsistent estimator in the context of linear autoregression. The 'secret' lies with the kernel, which helps us in a similar way as the windows in spectral estimation (see Tong 1994). In this respect (but not necessarily others), his cited references will probably not induce improvement by positive examples.

It is clear to us that we are into a different ball (or rather cylinder) game. Finally, Dr Bhansali might also like to refer to Cheng & Tong (1993*b*) for an example of the leave-more-than-one-out-approach.

P. M. ROBINSON (*London School of Economics, U.K.*). Have the authors established any results on the limiting distribution theory of their estimates? This would be of value in assessing the variability of the estimates in practice.

I think the authors' approach of estimating d is a good one. My paper (Robinson 1989) presents an alternative approach. There I introduced, and gave theoretical justification for, a class of statistical tests which, in the present setting of nonlinear autoregressions, can be used to test the null hypothesis that d takes a specified valud d_0, versus alternatives $d > d_0$. It is assumed that the process is β-mixing with suitable additional restrictions. The statistic avoids the stochastic denominator involved in the Nadaraya–Watson kernel regression estimate.

H. TONG. To Professor Robinson's second point, we would mention that Yao & Tong (1994) have suggested a way of avoiding the 'zero-problem' of the stochastic denominator involved in the Naderaya–Watson estimator, as well as in its locally linear extension. His question is quite pertinent. First, we should point out that when considering the (limiting) distributional theory for $\hat{d}_{CV}$, we should abandon the use of the euclidean distance in view of the integer nature of d_0 and $\hat{d}_{CV}$, and hence the conventional forms of the central limit theorem and the large deviation results. In place of the euclidean distance, we suggest that $\Delta(.,.)$ is a sensible metric. In an unpublished manuscript, we have shown asymptotically

$$P[\hat{d}_{CV} = d_0] \leqslant \max(1 - t_u, 1 - t_0), \tag{D 1}$$

where t_u is the 'tail probability (of under-fitting)' corresponding to

$$\mathcal{N}(-\sqrt{N}\Delta(d, d_0), \Sigma)$$

and t_0 is the 'tail probability (of over-fitting)' corresponding to $\mathcal{N}(-\sqrt{N}\beta\sigma^2(d_0) + \nabla, \Sigma)$ with Δ and Σ being constants depending on d_0 only and N denoting the sample size. Note the role played by $\Delta(.,.)$ in the above. Some may argue that the sample size requirement for embedding dimension estimation is a function of the 'tail probabilities', namely t_u and t_0. The above result (D 1) suggests that (i) this reduces to a function of $\Delta(.,.)$ for which we have suggested what seems to us the very reasonable bound of $1/d$; (ii) since t_u and t_0 are in general nonlinear

functions of $\Delta(.,.)$ (involving Δ^{-2}), it is much more difficult to 'control' t_u and t_0 than $\Delta(.,.)$.

A. KLECZKOWSKI (*Department of Plant Sciences, University of Cambridge, U.K.*). Embedding dimension for time series has been studied by Sauer *et al.* (1991). The required dimension is usually lower than the Takens estimate of $2d + 1$. The latter is needed to correctly represent details of some parts of the attractor, where the trajectory is highly entangled. Are their results relevant to the paper? Is it possible, by using similar arguments, to further reduce the sample size requirement, if we limit our interest to gross patterns of the attractor?

H. TONG. Our answer to his first question is 'yes' because the embedding dimension, as we have defined it, is bounded below by twice the 'fractal' dimension and our kernel estimate of the skeleton may be viewed as a low-pass filter. Our answer to his second question is 'unlikely' because our approach is already one of seeking out the gross patterns of the attractor via the geometric concept of the cylinders.

Additional references

Akaike, H. 1979 A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* **66**, 237–242.

Bhansali, R. J. & Downham, D. Y. 1977 Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547–551.

Robinson, P. M. 1989 Hypothesis testing in semi-parametric and non-parametric and non-parametric models for econometric time series. *Rev. Economic Stud.* **56**, 511–534.

Sauer, Yorke & Casdagli 1991 *J. statist. Phys.* **65**, 579–616.

Stone, M. 1977 An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Jl R. statist. Soc.* B **39**, 44–47.

Yao, Q. & Tong, H. 1994 Quantifying the influence of initial values on nonlinear prediction. *Jl R. statist. Soc.* B **56**, 301–325.

Zhang, P. 1993 Model selection via multifold cross-validation *Ann. Statist.* **21**, 299–313.