On delay co-ordinates in stochastic dynamical systems

Bing Cheng and Howell Tong

Abstract

We study the asymptotic distribution of the cross-validatory estimate of the delay co-ordinates in a stochastic dynamical system. By studying the tail probabilities of under-fitting and over-fitting, we obtain an estimate of the sample size requirement under realistic conditions.

1 Introduction

For the analysis of deterministic dynamical systems, the delay co-ordinates approach due to Takens (1981) is now firmly established and is one of the most frequently employed techniques in the dynamical systems literature. Although the actual mechanics of delay co-ordinates construction was pre-dated by the statistical literature (notably Yule, 1927), it is only through the celebrated embedding theorem of Takens that we understand the full impact of such a construction.

Nowadays, delay co-ordinates are so widely used in the dynamical systems literature that they are often applied even when the system noise (also called the intrinsic noise or the dynamic noise) is present. Strictly speaking, this situation is beyond the scope of Takens' theorem. The primary motivation of Takens' delay co-ordinates construction is the recruitment of a finite and minimally sufficient set of past observations with which we analyse the dynamical system (e.g. the attractors) and Takens' embedding theorem assures us of the existence of such sets which preserve all the essential features of the deterministic dynamical system under generic conditions. The recruitment process can also be likened to an information condensation process: the recruitment of redundant past observations provides no additional information. It is pertinent to discuss the purpose and the methodology of a similar delay co-ordinates construction within the wider context of stochastic dynamical systems, in which system noise (or noise for short) is present.

2 Delay Co-ordinates

Let $\{X_t\}$ be a discrete-time stationary time series with $EX_t^2 < \infty$. The conditional expectation of X_t given $(X_{t-1}, ..., X_{t-d})$ will be denoted by $E[X_t|X_{t-1}, ..., X_{t-d}]$. Define the prediction error variance by

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

Define the generalized partial autocorrelation function (PACF) by

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

Definition 2.1 $\{X_t\}$ is said to be generated by a stochastic dynamical system with d_0 delay co-ordinates, in short $SDS(d_0)$, if $\exists$ a non-negative integer $d_0 < \infty$ such that $\phi(d_0 - 1) \neq 0$ and $\phi(d) = 0$ for all $d \geq d_0$. If no such finite d_0 exists, then $\{X_t\}$ is said to be generated by a stochastic dynamical system with infinite number of delay co-ordinates, or $SDS(\infty)$.

The underlying idea is the recruitment of past observations for the purpose of one-step-ahead prediction by the least-square method. The function $\phi^2(d)$ measures the percentage reduction in the prediction error variance in adding X_{t-d-1} to the recruitment set consisting of $X_{t-1}, ..., X_{t-d}$. Clearly, an $SDS(d_0)$ may be modelled as

$$X_t = F_{d_0}(X_{t-1}, ..., X_{t-d_0}) + \epsilon_t, \tag{3}$$

where

$$E[\epsilon_t | X_{t-1}, ..., X_{t-d_0}] = 0. \tag{4}$$

A more general idea is to identify d_0 as the minimum integer such that the vector time series $\{X_t^{(d_0)}\}$ is a Markov chain on $\mathbb{R}^{d_0}$. Here, $X_t^{(d_0)} = (X_{t-1}, ..., X_{t-d_0})^T$. We shall pursue this more general idea elsewhere.

Example 2.1 Consider the stochastic logistic map

$$X_t = A_t X_{t-1}(1 - X_{t-1}) + \eta_t g(X_{t-1}), \quad X_0 \in (0, 1),$$

where $\{\eta_t\}$ is a sequence of independent and identically distributed random variables, each with zero mean, finite variance and finite support, $g(.)$ is any suitable function which ensures that $X_t \in (0, 1), \forall t \geq 1$, and η_t is independent of $X_s, s < t$. Moreover, A_t is a random variable with mean $\alpha, (0 \leq \alpha \leq 4)$, finite variance and compact support and is independent of $X_s, s < t$. Clearly,

$$\phi(d) = 0, \quad (d \geq 1). \tag{5}$$

Therefore, we have an $SDS(1)$. Note that we have incorporated a non-additive system noise process as well as parameter uncertainty in the above map. Further, we note that $(F_1(x_1), x_1)^T$ traces a parabola in $\mathbb{R}^2$, whilst $(F_d(x_d, ..., x_1), x_d, ..., x_1)^T$ traces a parabolic *cylinder* in $\mathbb{R}^{d+1}$ for each $d \geq 2$. It is clear that the cylindrical structure will prevail even if we consider general maps (of possibly higher dimensions and with more exotic "shape".) □

The fact that redundancy is characterized by a cylindrical structure suggests that cylinder hunters will reap great rewards in the face of noisy data.

3 Distance Function

Recall that $X_t^{(d)} = (X_{t-1}, ..., X_{t-d})$ and that $E[X_t | X_t^{(d)}]$ is denoted by $F_d(X_t^{(d)})$. Let $\mathcal{L}_2(\mathbb{R}^d)$ denote the set of all square-integrable measurable functions on $\mathbb{R}^d$. Obviously, $\mathcal{L}_2(\mathbb{R}^1) \subset \mathcal{L}_2(\mathbb{R}^2) \subset ... \subset \mathcal{L}_2(\mathbb{R}^d) \subset ....$

Denote

$$\mathcal{L}_2(X_t^{(d)}) = \{F(X_t^{(d)}) | F \in \mathcal{L}_2(\mathbb{R}^d)\}. \tag{6}$$

Then

$$\mathcal{L}_2(X_t^{(1)}) \subset \mathcal{L}_2(X_t^{(2)}) \subset \ldots \subset \mathcal{L}_2(X_t^{(d)}) \subset \ldots \tag{7}$$

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

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

and

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

Our objective is to define a suitable distance function on $\mathbb{N} \times \mathbb{N}$ which enables us to determine d_0. Clearly, the Euclidean distance is not appropriate for many "discrete" problems such as ours. For example, Akaike (1974) has used instead the Kullback-Leibler information to construct a suitable distance function for linear autoregressive order determination. For our purpose, it turns out that a feasible non-Euclidean distance emerges if we consider the Euclidean distance between two cylinder sets of dimensions say d_1 and d_2 $(d_1 \leq d_2)$ in the space of square-integrable functions on $\mathbb{R}^{d_2}$. This motivates the following definition of the function $\Delta(.,.)$ on $\mathbb{N} \times \mathbb{N}$, which will serve as our choice of a non-Euclidean (squared) distance on $\mathbb{N} \times \mathbb{N}$.

$$\Delta(d_1, d_2) = E[F_{d_1}(X_t^{(d_1)}) - F_{d_2}(X_t^{(d_2)})]^2, \tag{8}$$

where the expectation is taken with respect to the distribution of $X_t^{(d_2)}$. Note that F_d is uniquely determined once d is defined. Thus, $\Delta(.,.)$ is well defined.

Definition 3.1 The time series $\{X_t\}$ is an $SDS(d_0)$, $(d_0 \geq 1)$, if and only if

(i) $\Delta(d, d_0) \neq 0$ for all $d < d_0$, and

(ii) $\Delta(d, d_0) = 0$ for all $d \geq d_0$.

Proposition 3.1

(i) $\Delta^{1/2}$ *is a properly defined distance function on* $\mathbb{N} \times \mathbb{N}$, *i.e.* $\Delta^{1/2}(d_1, d_2) = \Delta^{1/2}(d_2, d_1)$, $\Delta^{1/2}(d, d) = 0$, $\Delta^{1/2}(d_1, d_3) \leq \Delta^{1/2}(d_1, d_2) + \Delta^{1/2}(d_2, d_3)$.

(ii) *If for each* $d \geq 1$, F_d *has bounded first partial derivatives on* $\mathbb{R}^d$, *then* $\Delta(d_2, d_1) \leq c|d_2 - d_1|$, *where* c *is a constant.*

(iii) *For* $d_1 \leq d_2 \leq d_3$, $\Delta(d_2, d_3) \leq \Delta(d_1, d_3)$. *That is for fixed* d_3, $\Delta(d, d_3)$ *is a decreasing function in* d.

(iv) For any $d_1 \leq d_2$, we have $\Delta(d_1, d_2) = \sigma^2(d_1) - \sigma^2(d_2)$.

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

(vi) *There are infinitely many* d *for which* $\Delta(d, d+1) \leq \kappa/d$, *where* κ *is a constant.*

(vii) $\forall d \leq D < \infty, \exists \kappa_D, 0 < \kappa_D < \infty$, *such that* $\Delta(d, d+1) \leq \kappa_D/d$.

The proofs are given in Cheng and Tong (1994).

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

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

where $\epsilon > 0$.

Note also that for continuous parameters such as the bandwidth parameters in kernel smoothing, we may use the Euclidean distance as an appropriate distance function for parameter (e.g. bandwidth) choice. However, as mentioned earlier, for many discrete cases, the Euclidean distance is found to be inappropriate. For our case, we have obtained an appropriate non-Euclidean distance function, namely $\Delta^{1/2}(.,.)$ on $\mathbb{N} \times \mathbb{N}$, based on the projection of the skeleton from a low dimensional space to a high dimensional space as described earlier. Proposition 3.1 (ii) reveals the relation between $\Delta(.,.)$ and the Euclidean distance.

4 Estimation

Henceforth we suppose that $\{X_t\}$ is a bounded time series (Cf. Chan and Tong, 1994). Let $\mathbb{B}_s^t(X)$ denote the sigma algebra generated by $(X_s, ..., X_t)$ and suppose that the following conditions are satisfied:

(a) $E[\epsilon_t | \mathbb{B}_{-\infty}^{t-1}(X)] = 0$, almost surely.

(b) $E[\epsilon_t^2 | \mathbb{B}_{-\infty}^{t-1}(X)] = \sigma^2$, a strictly positive constant, almost surely.

(c) For each d, $E[X_t | X_{t-1}, ..., X_{t-d}]$ is Hölder continuous.

(d) Let the probability density function of $(X_t, ..., X_{t-d})$ be strictly positive and Lipschitz continuous on a compact set in $\mathbb{R}^d$.

(e) Let k denote a probability density function with compact support on $\mathbb{R}^1$, and $\forall x, y \in \mathbb{R}^1, |k(x) - k(y)| \leq c_3|x - y|$.

(f) For every $t, s, \tau, t', s', \tau' \in \mathbb{N}$, the joint probability density function of $(X_t, X_s, X_\tau, X_{t'}, X_{s'}, X_{\tau'})$ is bounded.

(g) Let $1/p + 1/q = 1$. For some $p > 2$ and $\delta > 0$ such that $\delta < 2/q - 1$, $E|\epsilon_s|^{2p(1+\delta)} < \infty$ and $E|F(X_1)|^{2p(1+\delta)} < \infty$.

(h) For each d, F_d has bounded first partial derivative.

(i)

$$\sup_{i\in\mathbb{N}}(E[\sup_{A\in\mathbb{B}_{i+j}^{\infty}(X)}\{|P(A|\mathbb{B}_1^i(X)) - P(A)|\}]) = \bigcirc(\beta^j), 0 < \beta < 1.$$

Without loss of generality, let $d_2 \geq d_1$. Let M be a pre-specified maximum lag in the delay co-ordinate construction. Equation (9) suggests that a natural estimate of $\Delta(d_1, d_2)$ is

$$\hat{\Delta}(d_1, d_2) = RSS(d_1) - RSS(d_2), \tag{9}$$

where

$$RSS(d) = (N - M + 1)^{-1} \sum_{t=M}^{N} \{X_t - \hat{F}_{d,N}(X_t^{(d)})\}^2, \tag{10}$$

with $\hat{F}_{d,N}$ being the Nadaraya-Watson kernel estimate of F_d based on the observations $X_1, ..., X_N$, namely

$$\hat{F}_{d,N}(x_1, , ..., x_d) = \frac{\sum_{t=M}^{N-1} X_{t+1} k(\frac{x_1 - X_t}{h}) k(\frac{x_2 - X_{t-1}}{h}) ... k(\frac{x_d - X_{t-d+1}}{h})}{\sum_{t=M}^{N-1} k(\frac{x_1 - X_t}{h}) k(\frac{x_2 - X_{t-1}}{h}) ... k(\frac{x_d - X_{t-d+1}}{h})}.$$

Here, $h \equiv h_{d,N} \in [aN^{-(1/(2d+1))-\xi}, bN^{-(1/(2d+1))+\xi}]$, with a and b being arbitrary real positive constants and ξ any real positive constant strictly less than $\{2(d+1)(2d+1)\}^{-1}$. Cheng and Tong (1992) have proved the following theorem.

Theorem 4.1 *Under the above conditions,*

$$RSS(d) = \sigma_N^2(d)\{1 - (2\alpha - \beta)/(Nh^d) + o_p(1/(Nh^d))\},$$

where

$$\sigma_N^2(d) = (N - M + 1)^{-1} \sum_{t=M}^{N} \{X_t - F_d(X_t^{(d)})\}^2, \tag{11}$$

$$\alpha(d) = \{k(0)\}^d \text{ and } \beta(d) = \{\int k^2(u)du\}^d.$$

Now, using this theorem, we may easily deduce that for each $d \geq d_0$ and $h_{d,N} = N^{-1/(2d+1)}$

$$\begin{aligned} \hat{\Delta}(d, d+1) &= RSS(d) - RSS(d+1) && (12)\\ &= \sigma_N^2(d_0)\{2\alpha(d+1) - \beta(d+1)\}N^{-(d+2)/(2d+3)} && (13)\\ &\quad + o_p(N^{-(d+2)/(2d+3)}). && (14) \end{aligned}$$

This analysis shows that if we use $\hat{\Delta}(d, d+1)$ to obtain an estimate of d_0, we have to decide where the former cuts off. Recall that $\Delta(d, d+1) = 0$ for $d \geq d_0$. (Cf. equation

(9).) Thus, we are facing a statistical problem of the same type as described in Akaike (1974). A conventional statistical approach prior to Akaike's innovation would be along the line of testing a class of null hypotheses: $\Delta(d, d+1) = 0, d \in \{1, 2, ..., M\}$. In the present setting of a nonparametric autoregression, Robinson (1989) has adopted the conventional approach by considering the problem of testing the null hypothesis that d takes a specified value say $\tilde{d}$ versus the alternatives $d > \tilde{d}$. Presumably, one would then have to "scan" $\tilde{d}$ over the set say $\{1, 2, ..., M\}$ in a suitable manner, which has to be specified. Recently, Cheng and Tong (1992, 1994) have adopted an approach in the spirit of Akaike (1974). Specifically, they have proposed a cross-validatory method: replace $\hat{F}_{d,N}(X_t^{(d)})$ in equation (11) by an estimate which is obtained from the observed sample but with X_t deleted. Let $\hat{F}_{d,N,\backslash t}(X_t^{(d)})$ denote this delete-one estimate and

$$CV(d) = (N - M + 1)^{-1} \sum_{t=M}^{N} \{X_t - \hat{F}_{d,N,\backslash t}(X_t^{(d)})\}^2. \tag{15}$$

Effectively the "delete-one" device penalizes model complexity and Cheng and Tong (1992) have shown that $argmin_{1 \le d \le M} CV(d)$, or $\hat{d}_{CV}$ for short, yields a *consistent* estimate of d_0 provided $d_0 \le M$, i.e. $Pr\{\hat{d}_{CV} = d_0\} \to 1$ as $N \to \infty$. Briefly, from Cheng and Tong (*op. cit.*) we can easily deduce that for bounded time series, (*i.e.* X_t is bounded.)

$$CV(d) - CV(d_0) = \sigma_N^2(d) - \sigma_N^2(d_0) + \sigma_N^2(d^*)\beta(d^*)N^{-(d^*+1)/(2d^*+1)} \tag{16}$$

$$+ o_p(N^{-(d^*+1)/(2d^*+1)}), \tag{17}$$

where $d^* = \max\{d, d_0\}$. Note that $\sigma_N^2(d) = \sigma_N^2(d_0)$ for $d \ge d_0$ and that for $1 \le d \le M$, $\sigma_N^2(d) \to \sigma^2(d)$ almost surely as $N \to \infty$. Consistency then follows.

5 Tail Probabilities

It would be pertinent to investigate the limiting distribution of $\hat{d}_{CV}$ further. First we notice that

$$\begin{aligned} P(\{\hat{d}_{CV} = d_0\}) &= P(\{CV(d_0) \le CV(d), 1 \le d \le M\}) \\ &= P(\{CV(d_0) \le CV(d), 1 \le d < d_0\}) + P(\{CV(d_0) \le CV(d), d_0 \le d \le M\}) \\ &\quad - P(\{CV(d_0) \le CV(d), 1 \le d < d_0\} \bigcap \{CV(d_0) \le CV(d), d_0 \le d \le M\}). \end{aligned}$$

Now, let $\theta(d) = (2d+1)/(d+1)$.

Case 1: $(d < d_0)$
We have

$$CV(d) - CV(d_0) \quad = \quad \hat{\Delta}(d, d_0) - RSS(d_0) \times 2\alpha N^{-\theta^{-1}(d_0)} + o_p(N^{-\theta^{-1}(d_0)})$$

$$= \hat{\Delta}(d,\ d_0) - \Delta(d,\ d_0) + \Delta(d,\ d_0) - RSS(d_0) \times 2\alpha N^{-\theta^{-1}(d_0)} + o_p(N^{-\theta^{-1}(d_0)}).$$

Since

$$\theta^{-1}(d_0) = \frac{d_0+1}{2d_0+1} = \frac{1}{2-1/(d_0+1)} > \frac{1}{2},$$

the above is equal to

$$\Delta(d,\ d_0) + o_p(N^{-\frac{1}{2}}).$$

This implies that

$$\sqrt{N}(CV(d) - CV(d_0)) \sim_{asym} \mathcal{N}(\sqrt{N}\Delta(d,\ d_0), \Sigma).$$

So, for $1 \leq d < d_0$,

$$\begin{aligned} P(\{CV(d_0) \leq CV(d)\}) &= P(\{\sqrt{N}(CV(d_0) - CV(d)) \leq 0\}) \\ &=_{asym} P(\xi_d \leq 0), \end{aligned}$$

where $\xi_d \sim_{asym} \mathcal{N}(-\sqrt{N}\Delta(d, d_0),\ \Sigma)$. Hence, we have the formula $P(\xi_d \leq 0) = 1 - t_u$ for $d < d_0$, where t_u is the tail probability of underfitting.

Case 2: $(d > d_0)$
We have $\epsilon_t^{(d)} = \epsilon_t^{(d)}$, *a.s.*, Using formula (16), we have for $d > d_0$

$$CV(d) - CV(d_0) = \sigma_N^2(d_0)\beta N^{-\theta^{-1}(d)} + o_p(N^{-\theta^{-1}(d)}),$$

where $\sigma_N^2(d_0) = \frac{1}{N}\sum_{i=1}^{N}[\epsilon_t^{(d)}]^2$.

By the standard Central Limit Theorem, we have

$$\sqrt{N}(\sigma_N^2(d_0) - \sigma^2(d_0)) \sim \mathcal{N}(0,\ \Sigma)$$

and by a high-order expansion, we may obtain

$$\sqrt{N}N^{\theta^{-1}(d)}(CV(d) - CV(d_0)) = \sqrt{N}\sigma_N^2(d_0)\beta + \bigtriangledown + o_p(1),$$

where $\bigtriangledown$ is a constant. Therefore,

$$P(\{CV(d_0) \leq CV(d)\}) = P(\{\sqrt{N}N^{\theta^{-1}(d)}(CV(d_0) - CV(d)) \leq 0\}) \sim_{aymp} P(\eta_d \leq 0),$$

where $\eta_d \sim_{asym} \mathcal{N}(-\sqrt{N}\sigma^2(d)\beta + \bigtriangledown,\ \tilde{\Sigma})$.

Putting the two results together, we have that

$$\begin{aligned} P(\hat{d}_{CV} = d_0) &\leq \max\{(1 - \text{tail prob of underfitting}), \\ &\qquad (1 - \text{tail prob of overfitting})\} \\ &= \max\{1 - t_u,\ 1 - t_o\}, \end{aligned}$$

where $t_u = P(\xi_d \geq 0)$ and $t_o = P(\eta_d \geq 0)$. Now, we show that by using the above tail probabilities, we obtain a similar formula for the sample requirement as that reported in Cheng and Tong (1994). First, we need a simple lemma.

Lemma *Let Z be a normal random variable with mean $-M$ $(M > 0)$ and variance σ^2. Then*

$$P(Z \geq 0) \;=\; \int_M^{+\infty} e^{-\frac{x^2}{2\sigma^2}} dx \;=\; O(Me^{-\frac{1}{2}M^2}).$$

Now, for the tail probability of underfitting, t_u, $M = \sqrt{N}\Delta(d,\ d_0)$ and for the tail probability of overfitting, t_o, $M = \sqrt{N}\beta\gamma\sigma^2(d_0)$. To control the tail probabilities at level $\epsilon > 0$, we need to have

$$Me^{-\frac{1}{2}M^2} \;\leq\; \epsilon \text{ asymptotically.}$$

Since $\beta = \beta(d_0)$ has a complicated form, the tail probability of overfitting is not so helpful. However, for the tail probability of underfitting, since $M = \sqrt{N}\Delta(d,\ d_0)$, it is easy to see that

$$N \;\geq\; \Delta^{-2}(d,\ d_0) log(\frac{1}{\epsilon}).$$

In particular, choosing $d = d_0 - 1$, we readily have $\Delta(d_0 - 1,\ d_0) \;=\; O(\sigma^2(d_0)/d_0)$ as in Cheng and Tong (1994). Therefore, we obtain

$$N \;=\; N(d_0) \;\geq\; \frac{d_0^2 log(1/\epsilon)}{\sigma^4(d_0)}.$$

6 Conclusion

Using an argument based on controlling the tail probabilities, we have arrived at the same sample size requirement under realistic conditions as that obtained in Cheng and Tong (1994) for the construction of delay co-ordinates in a stochastic dynamical system.

References

[1] Akaike, H., A new look at the statistical model identification. *I.E.E.E. Trans. Autom. Cont.*, **AC-19** (1974), 716–23.

[2] Chan, K.S. and H. Tong, A note on noisy chaos. *J. Roy. Statist. Soc.*, **B 56** (1994), 301–11.

[3] Cheng, B. and H. Tong, Consistent nonparametric order determination and chaos–with discussion. *J. Roy. Statist. Soc.*, **B 54** (1992), 427–49 and 451–74.

[4] Cheng, B. and H. Tong, Orthogonal projection, embedding dimension and sample size in chaotic time series from a statistical perspective. *Philos. Trans. Roy. Soc.*, **A 348** (1994), 325–41.

[5] Takens, F., Detecting strange attractors in turbulence. *Dynamical systems and turbulence,* ed. D.A. Rand and L.-S. Young, Springer Lecture Notes in Mathematics, **898** (1981), 366–81.

[6] Yule, G.U., On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philos. Trans. Roy. Soc.*, **A 226** (1927), 267–98.

Institute of Mathematics and Statistics
University of Kent at Canterbury
Kent CT2 7NF UK