

非線形時系列データに対する 埋め込み次元・遅延時間の推定法について

米本 孝二 (九大・数理)¹

1 Introduction

Fueda-Yanagawa (2001) は dynamic noise を考慮に入れた非線形自己回帰モデル

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

where $F: \mathbf{R}^d \rightarrow \mathbf{R}$ 未知の非線形写像
 $d, \tau \in \mathbf{Z}_{>0}$
 ε_t : dynamic noise

に従う時系列 $\{X_t\}$ に対し、Nadaraya-Watson のカーネル推定量と交差検証法を用いた埋め込み次元・遅延時間の推定法を提案し、その推定量が一致推定量であることを示した。しかし彼らは、応用の際重要となる交差検証法の値の最小化の手順を明示していないので、そのままでは実データに対し応用することが出来ない。また単純に交差検証法の値が小さいほうが最適であるとして彼らの手法を人工的に発生したデータに当てはめてみたが、誤った結果を得ることがしばしばあった。そこで、私は交差検証法の値の最小化に新しい評価基準を導入した推定手順を提案し、数値実験でその推定値の挙動を確かめた。

2 推定手順

次のように埋め込み次元・遅延時間を定義する。

Def. (Yonemoto and Yanagawa(1998))

時系列 $\{X_t\}$ の埋め込み次元が d_0 、遅延時間が τ_0 である。

$\Leftrightarrow$

$\exists d_0, \tau_0 \in \mathbf{Z}_{>0}$ such that $d_0 < \infty, \tau_0 < \infty$,

$$E[X_t | X_{t-\tau}, X_{t-2\tau}, \dots, X_{t-d\tau}] \neq E[X_t | X_{t-\tau_0}, X_{t-2\tau_0}, \dots, X_{t-d_0\tau_0}] \text{ a.e.}$$

for $\forall d < d_0$, and $\forall \tau > 0$, and

$$E[X_t | X_{t-\tau}, X_{t-2\tau}, \dots, X_{t-d\tau}] = E[X_t | X_{t-\tau_0}, X_{t-2\tau_0}, \dots, X_{t-d_0\tau_0}] \text{ a.e.}$$

for $\forall (d, \tau) \in \mathbf{B}(d_0, \tau_0)$,

where $\mathbf{B}(d_0, \tau_0) = \{(d, \tau) | \{\tau_0, 2\tau_0, \dots, d_0\tau_0\} \subset \{\tau, 2\tau, \dots, d\tau\}\}$.

この定義の下で私は次のような埋め込み次元・遅延時間の推定手順を提案する。

¹The author thank the Institute of Statistical Mathematics for the facilities and the use of SGI2800.

$\{X_1, X_2, \dots, X_N\} \subset \mathbf{R}$ をモデル (1) に従う観測時系列データとする。

このとき正整数 $d, \tau, L \geq d\tau$ に対して

$$CV(d, \tau) = \frac{1}{N - L + 1} \sum_{t=L}^N (X_t - \hat{F}_{\setminus t(d, \tau)}(X_{t-\tau}, \dots, X_{t-d\tau}))^2,$$

where

$$\left\{ \begin{array}{l} \hat{F}_{\setminus t(d, \tau)}(z) = \frac{\sum_{s=L, s \neq t}^N X_s K_{d, h}(z - (X_{s-\tau}, \dots, X_{s-d\tau}))}{\sum_{s=L, s \neq t}^N K_{d, h}(z - (X_{s-\tau}, \dots, X_{s-d\tau}))} \\ K_{d, h}(z) = \frac{1}{h^d} \prod_{i=1}^d \phi\left(\frac{z_i}{h}\right), \quad (z = (z_1, z_2, \dots, z_d)) \\ \phi: \text{標準正規分布の密度関数} \\ h = c \times N^{-\frac{1}{2d+1}} \\ c: \text{定数} \end{array} \right.$$

とおく。このとき、次のように埋め込み次元・遅延時間を推定する。

1. 未知の埋め込み次元 d_0 、遅延時間 τ_0 に対して、十分大きな正整数 $D > d_0, T > \tau_0$ と $L = DT$ を与える。
2. 各 $\tau \in \{1, 2, \dots, T\}$ と $d \in \{1, 2, \dots, D\}$ に対し、 $\hat{CV}(d, \tau) := \min_c CV(d, \tau)|_{h=h(c)}$ を求める。
3. 各 $\tau \in \{1, 2, \dots, T\}$ に対して、 $CV^*(\tau) := \min_d \hat{CV}(d, \tau)$ を求める。
4. 十分小さな $\varepsilon > 0$ を与え、各 $\tau \in \{1, 2, \dots, T\}$ に対して $\hat{d}(\tau) = \min\{d : |\hat{CV}(d, \tau) - CV^*(\tau)| < \varepsilon\}$ を求める。
5. $CV^* = \min\{\hat{CV}(\hat{d}(\tau), \tau) | \tau \in \{1, 2, \dots, T\}\}$ とおくと、埋め込み次元の推定量 $\hat{d}_0$ と遅延時間の推定量 $\hat{\tau}_0$ は $\hat{d}_0 = \min\{\hat{d}(\tau) : |\hat{CV}(\hat{d}(\tau), \tau) - CV^*| < \varepsilon\}$ と $\hat{\tau}_0 = \operatorname{argmin}\{\hat{d}(\tau) : |\hat{CV}(\hat{d}(\tau), \tau) - CV^*| < \varepsilon\}$ で与えられる。

3 Simulation

次の4つのモデルから生成されるデータに対しシミュレーションを行った。

$$\text{Model 1: } X_t = 1 - 1.4X_{t-1}^2 + 0.3X_{t-2} + \varepsilon_t \quad \text{with } d_0 = 2 \text{ and } \tau_0 = 1$$

$$\text{Model 2: } X_t = -0.1X_{t-1} + (-1 + 2 \exp(-X_{t-1}^2))X_{t-3} + \varepsilon_t \quad \text{with } d_0 = 3 \text{ and } \tau_0 = 1$$

$$\text{Model 3: } X_t = -0.1X_{t-2} + (-1 + 2 \exp(-X_{t-2}^2))X_{t-6} + \varepsilon_t \quad \text{with } d_0 = 3 \text{ and } \tau_0 = 2$$

$$\text{Model 4: } X_t = 0.5X_{t-1} + (-0.5 + 2 \exp(-X_{t-1}^2))X_{t-7} + \varepsilon_t \quad \text{with } d_0 = 7 \text{ and } \tau_0 = 1$$

但し、 $\varepsilon_t \sim N(0, \sigma^2)$ とする。まず適当な初期値を与え $N + 1000$ 個のデータを発生し、初期値の影響を避けるため最初の 1000 点を除き、 N 点のデータを生成した。そしてこの N 点のデータを標準化し、標準化したデータに対し交差検証法の値が小さいほうが良いとした Fueda-Yanagawa の方法 (F-Y method) と提案した手法を適用した。この手順を、各モデル・データ数・dynamic noise の分散 σ^2 の組み合わせに対して 100 回ずつ繰り返し、推定値の挙動を見た。また、提案する手法で用いる ε の値は $\frac{1}{10}, \frac{1}{20}$ の2つの値を用いた。

3.1 Model 1

$N = 1000, 2000, 4000$, $\sigma^2 = 0, 0.005, 0.01$ の各組み合わせに対して次のようにデータを生成した。ここでは $D = 5, T = 5$ として手法を適用した。

Model 1 に従う時系列データは、 $\sigma^2 = 0$ のときであっても初期値の選び方によってはデータが発散することが知られている。そこで初期値 (x_2, x_1) を四辺形 $ABCD$ の中からランダムに与える。但し、 $A = (-1.33, 1.4)$, $B = (1.32, 0.433)$, $C = (1.245, -0.466)$, $D = (-1.06, -1.666)$ 。この四辺形 $ABCD$ は M.Henon (1976) が収束領域として紹介しているものである。ただ、初期値をこの四辺形の中から選択しても $\sigma^2 \neq 0$ のときにはまだ発散する可能性がある。そこでデータを次のように生成した。

1. 初期値 (x_2, x_1) を四辺形 $ABCD$ の中からランダムに選択する。
2. 今、 i 番目のデータ x_i を生成したとする。
 - $i \leq 12$ かつ $|x_i - x_{i-1}| > 3$ のときには $x_3, x_4, \dots, x_i$ を破棄し、初期値 x_1, x_2 を用いて x_3 から再び生成する。
 - $i > 12$ かつ $|x_i - x_{i-1}| > 3$ のときには $x_{i-10}, x_{i-9}, \dots, x_i$ を破棄し、 x_{i-10} から再び生成する。

推定結果は TABLE にまとめる。 $\sigma^2 = 0$ のときは、F-Y method と提案した手法はともにすべての場合において正しい推定値を得ることが出来た。 σ^2 の値が $0.005, 0.01$ となるに従って F-Y method は推定結果が悪くなり、 $\sigma^2 = 0.01$ のときは $N = 4000$ であっても正しい推定値を 100 回中 15 回しか推定することが出来なかった。一方、提案した手法は $\sigma^2 = 0.005, 0.01$ の場合でもすべての場合において正しい推定値を得ることが出来た。

3.2 Model 2

$N = 1000, 2000, 4000$, $\sigma^2 = 0, 0.01$ の各組み合わせに対し初期値 x_1, x_2, x_3 を区間 $(-1, 1)$ からランダムに選択しデータを生成した。ここでは、 $D = 6, T = 5$ として手法を適用した。結果を TABLE にまとめる。 $\sigma^2 = 0$ のとき F-Y method では高々74回しか正しい結果を得ることが出来なかったのに対し、提案した手法は $N \geq 2000$ ですべての場合において正しい推定値を得ることが出来た。また $\sigma^2 = 0.01$ のときは、F-Y method は一度も正しい推定値を得ることが出来なかった。一方、提案した手法は $\varepsilon = \frac{1}{20}$, $N = 4000$ のとき 95 回正しい推定値を得ることが出来た。

3.3 Model 3

$N = 1000, 2000, 4000$, $\sigma^2 = 0, 0.01$ の各組み合わせに対し初期値 $x_1, x_2, \dots, x_6$ を区間 $(-1, 1)$ からランダムに選択しデータを生成した。ここでは、 $D = 6, T = 5$ として手法を適用した。結果を TABLE にまとめる。 $\sigma^2 = 0$ のとき F-Y method では高々51回しか正しい結果を得ることが出来なかったのに対し、提案した手法は少なくとも 91 回正しい推定値

を得ることが出来た。また $\sigma^2 = 0.01$ のときは、F-Y method は高々2回しか正しい推定値を得ることが出来なかった。一方、提案した手法は $\varepsilon = \frac{1}{20}$, $N \geq 2000$ のとき少なくとも 98 回正しい推定値を得ることが出来た。

3.4 Model 4

$N = 3000, 6000$, $\sigma^2 = 0, 0.01$ の各組み合わせに対し初期値 $x_1, x_2, \dots, x_7$ を区間 $(-1, 1)$ からランダムに選択しデータを生成した。ここでは、 $D = 12, T = 5$ として手法を適用した。結果を TABLE にまとめる。 $\sigma^2 = 0$ のとき F-Y method、提案した手法ともに一度も正しい推定値を得ることが出来なかった。 $\sigma^2 = 0.01$ のときは、F-Y method は高々46回しか正しい推定値を得ることが出来なかったのに対し、提案した手法は少なくとも 99 回正しい推定値を得ることが出来た。

4 References

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. Vol.31, No.1, 27-38

Henon, M. (1976) A Two-dimensional Mapping with a Strange Attractor. Communications in Mathematical Physics, 69-77

Yonemoto, K. and Yanagawa, T. (1998) Graphical representation of observed chaotic time series and its application. In proceedings of chaos and non-linear time series symposium in Hakata

TABLE I Frequency distributions of $(\hat{d}_0, \hat{\tau}_0)$ in 100 trials for Model 1, where the bold face represents the correct estimates.

$\sigma^2 = 0$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$
$N = 2000$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$
$N = 4000$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$

$\sigma^2 = 0.005$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$	$(1, 5) \times 97$ $(\mathbf{2}, \mathbf{1}) \times \mathbf{3}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$
$N = 2000$	$(1, 5) \times 79$ $(\mathbf{2}, \mathbf{1}) \times \mathbf{21}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$
$N = 4000$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{92}$ $(1, 5) \times 8$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$

$\sigma^2 = 0.01$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$	$(1, 5) \times 100$ $(\mathbf{2}, \mathbf{1}) \times \mathbf{0}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$
$N = 2000$	$(1, 5) \times 99$ $(\mathbf{2}, \mathbf{1}) \times \mathbf{1}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$
$N = 4000$	$(1, 5) \times 85$ $(\mathbf{2}, \mathbf{1}) \times \mathbf{15}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$	$(\mathbf{2}, \mathbf{1}) \times \mathbf{100}$

TABLE II Frequency distributions of $(\hat{d}_0, \hat{\tau}_0)$ in 100 trials for Model 2, where the bold face represents the correct estimates.

$\sigma^2 = 0$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$			(3, 1) $\times$ 43
	(3, 1) $\times$ 46	(3, 1) $\times$ 65	(4, 4) $\times$ 38
	(3, 2) $\times$ 34	(4, 4) $\times$ 34	(5, 5) $\times$ 15
	(4, 4) $\times$ 20	(3, 2) $\times$ 1	(4, 1) $\times$ 3
			(3, 2) $\times$ 1
$N = 2000$	(3, 1) $\times$ 74		
	(4, 4) $\times$ 21	(3, 1) $\times$ 100	(3, 1) $\times$ 100
	(3, 2) $\times$ 5		
$N = 4000$	(3, 1) $\times$ 51		
	(3, 2) $\times$ 25	(3, 1) $\times$ 100	(3, 1) $\times$ 100
	(4, 4) $\times$ 24		
$\sigma^2 = 0.01$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$	(1, 3) $\times$ 99		
	(2, 3) $\times$ 1	(1, 3) $\times$ 94	(3, 1) $\times$ 51
	(3, 1) $\times$ 0	(3, 1) $\times$ 6	(1, 3) $\times$ 49
$N = 2000$	(1, 3) $\times$ 99		
	(2, 3) $\times$ 1	(1, 3) $\times$ 91	(3, 1) $\times$ 65
	(3, 1) $\times$ 0	(3, 1) $\times$ 9	(1, 3) $\times$ 35
$N = 4000$	(1, 3) $\times$ 99		
	(2, 3) $\times$ 1	(1, 3) $\times$ 87	(3, 1) $\times$ 95
	(3, 1) $\times$ 0	(3, 1) $\times$ 13	(1, 3) $\times$ 5

TABLE III Frequency distributions of $(\hat{d}_0, \hat{\tau}_0)$ in 100 trials for Model 3, where the bold face represents the correct estimates.

$\sigma^2 = 0$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$	(3, 2) × 51		
	(3, 4) × 29		(3, 2) × 93
	(6, 4) × 9	(3, 2) × 96	(3, 4) × 4
	(5, 3) × 6	(4, 2) × 2	(6, 2) × 1
	(4, 2) × 2	(3, 4) × 1	(4, 4) × 1
	(6, 2) × 1	(4, 4) × 1	(6, 4) × 1
	(5, 4) × 1		
	(5, 5) × 1		
$N = 2000$	(3, 2) × 36		
	(3, 4) × 26		(3, 2) × 91
	(6, 4) × 13		(2, 2) × 3
	(5, 3) × 12	(3, 2) × 94	(6, 1) × 2
	(5, 5) × 6	(2, 2) × 4	(4, 2) × 2
	(4, 2) × 3	(4, 2) × 2	(4, 4) × 1
	(5, 2) × 2		(3, 4) × 1
	(2, 4) × 2		
$N = 4000$	(3, 2) × 36		
	(3, 4) × 30		
	(5, 3) × 16		
	(4, 3) × 9	(3, 2) × 100	(3, 2) × 100
	(6, 4) × 8		
	(4, 2) × 1		
$\sigma^2 = 0.01$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 1000$	(2, 3) × 98	(2, 3) × 68	(3, 2) × 83
	(3, 2) × 2	(3, 2) × 32	(2, 3) × 15
$N = 2000$	(2, 3) × 100	(2, 3) × 67	(3, 2) × 98
	(3, 2) × 0	(3, 2) × 33	(2, 3) × 2
$N = 4000$	(2, 3) × 98	(2, 3) × 58	
	(3, 2) × 2	(3, 2) × 42	(3, 2) × 100

TABLE IV Frequency distributions of $(\hat{d}_0, \hat{\tau}_0)$ in 100 trials for Model 4, where the bold face represents the correct estimates.

$\sigma^2 = 0$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 3000$	$(9, 1) \times 63$		
	$(9, 2) \times 17$		
	$(6, 3) \times 9$	$(8, 2) \times 80$	$(9, 1) \times 80$
	$(7, 3) \times 5$	$(8, 1) \times 20$	$(8, 1) \times 20$
	$(7, 2) \times 4$	$(7, 1) \times 0$	$(7, 1) \times 0$
	$(5, 3) \times 1$		
	$(5, 5) \times 1$		
	$(7, 1) \times 0$		
$N = 6000$	$(9, 1) \times 80$		
	$(7, 3) \times 8$	$(9, 1) \times 80$	$(9, 1) \times 80$
	$(6, 3) \times 8$	$(8, 1) \times 20$	$(8, 1) \times 20$
	$(7, 2) \times 3$	$(7, 1) \times 0$	$(7, 1) \times 0$
	$(5, 5) \times 1$		
	$(7, 1) \times 0$		
$\sigma^2 = 0.01$	F-Y method	proposed method	
		$\varepsilon = CV^*/10$	$\varepsilon = CV^*/20$
$N = 3000$	$(4, 5) \times 34$		
	$(7, 1) \times 31$		
	$(4, 4) \times 21$		
	$(3, 5) \times 4$		
	$(3, 4) \times 2$	$(7, 1) \times 99$	$(7, 1) \times 99$
	$(5, 5) \times 2$	$(3, 4) \times 1$	$(3, 4) \times 1$
	$(7, 2) \times 2$		
	$(7, 3) \times 1$		
	$(5, 4) \times 1$		
	$(6, 5) \times 1$		
	$(1, 5) \times 1$		
$N = 6000$	$(7, 1) \times 46$		
	$(4, 5) \times 39$		
	$(4, 4) \times 12$	$(7, 1) \times 100$	$(7, 1) \times 100$
	$(5, 4) \times 2$		
	$(6, 5) \times 1$		