Postprocessing methods for finding the embedding dimension of chaotic time series

Teck Por Lim and Sadasivan Puthusserypady

Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, 117576 Singapore

(Received 18 November 2004; published 24 August 2005; corrected 6 September 2007)

One problem when using the global false nearest-neighbors (GFNN) method and Cao's method to estimate embedding dimension is that their effectiveness is affected by the ratio of signal power to noise power (SNR). Simple models are proposed to explain the curves commonly obtained when using the GFNN method and Cao's method. Methods are proposed for systematically estimating the embedding dimension. Prior information is incorporated to improve the estimates.

DOI: 10.1103/PhysRevE.72.027204 PACS number(s): 05.45.Tp, 05.45.Ac

I. INTRODUCTION

Takens' delay embedding theorem [1] states that a pseudo-state space can be reconstructed from an infinite, "noiseless" time series. However, ubiquitous measurement noise means that delay embeddings are inevitably constructed from noisy data. Consider a finite time series $\{y(n)\}_{n=1}^{N^{total}}$, with N^{total} samples, where the nth measured sample

$$y(n) \triangleq \check{y}(n) + \eta(n), \tag{1}$$

where $\check{y}(n)$ is the noiseless sample, and $\eta(n)$ is the measurement noise. Each embedding vector, $\boldsymbol{v}(n) \triangleq (y(n), y(n-\tau), \ldots, y(n-(d_E-1)\tau))$, is formed from $\{y(n)\}_{n=1}^{N^{total}}$; the embedding dimension $d_E \in \mathbb{Z}^+$ and the embedding delay [2] $\tau \in \mathbb{Z}^+$ need to be determined.

II. EMBEDDING DIMENSION

Sauer *et al.* [3] had shown that d_E is related to D_0, the box-counting dimension, [2] by

$$d_E > 2D_0. \tag{2}$$

Working with dimensions larger than the minimum required leads to excessive computations [4]. As (2) only provides a sufficient condition, it may be possible to use a smaller value of d_E. Ding *et al.* [5] showed that, for computing correlation dimension [2]

$$d_E = \mathrm{int}(D_0) + 1 \tag{3}$$

suffices. From the practical point of view, d_E should be chosen to give the best results for one's application [6]. However, it can be impractical to vary d_E over a large range of values. Thus, empirical methods such as the GFNN method [4] and Cao's method [7] are usually used.

A. Global false nearest-neighbors method

GFNN are embedding vectors $\boldsymbol{v}_d(n)$ [$\boldsymbol{v}(n)$ of dimension $d \in \mathbb{Z}^+$] which should be far apart in state space, but are nearby because d is too low. Define

$$R_d^2(n,n_{NN}) \triangleq \|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|^2 = \sum_{k=0}^{d-1} (y(n-k\tau) - y(n_{NN}-k\tau))^2, \tag{4}$$

where $R_d^2(n,n_{NN})$ is the squared Euclidean distance between $\boldsymbol{v}_d(n)$ and $\boldsymbol{v}_d(n_{NN})$ (where n_{NN} is the index of the nearest neighbor $\boldsymbol{v}_d(n_{NN})$). The first criterion to determine a GFNN is that the distance is large when d is changed to $d+1$

$$\frac{\|y(n-d\tau) - y(n_{NN}-d\tau)\|}{\|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|} > R_{tol}. \tag{5}$$

Inequality (5) is obtained by substituting Eq. (4) into $R_{d+1}^2(n,n_{NN}) - R_d^2(n,n_{NN})$, where an arbitrary threshold $R_{tol} \in \mathbb{R}^+$ is fixed such that $R_{tol} \geqslant 10$ [4] or $R_{tol} \approx 15$ [2]. The second criterion is

$$\frac{R_{d+1}(n,n_{NN})}{R_A} > A_{tol}, \tag{6}$$

where R_A^2 is the sample variance of the time series and $A_{tol} \approx 2$ is an arbitrary threshold. If either criterion is true, a false nearest neighbor is declared. The dimension d_E is the value of d at which the percentage of GFNN plateaus off [4]. One problem that Ref. [8] mentioned was that, with insufficient data and large d, false neighbors are found even for deterministic systems. There is poor discrimination between neighboring points as d increases [9]; this problem is unavoidable for distance-based methods. Some issues concern-

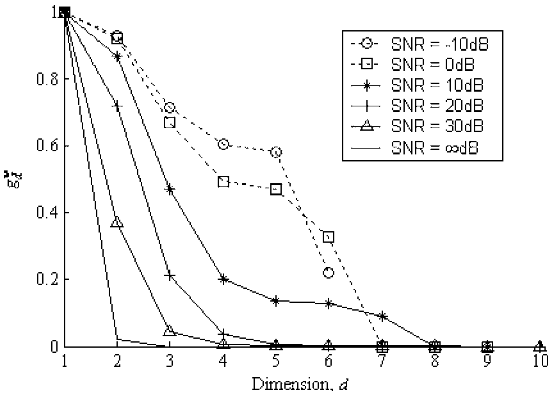

FIG. 1. Plots of $g_d^{\boldsymbol{v}}$ (Lorenz data) for various values of SNR (AWGN).

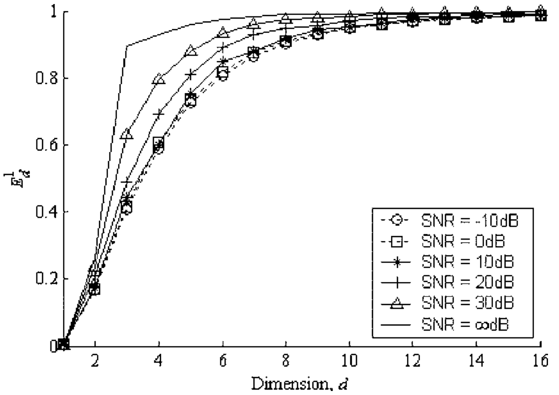

FIG. 2. Plots of E_d^1 (Lorenz data) for various values of SNR (AWGN).

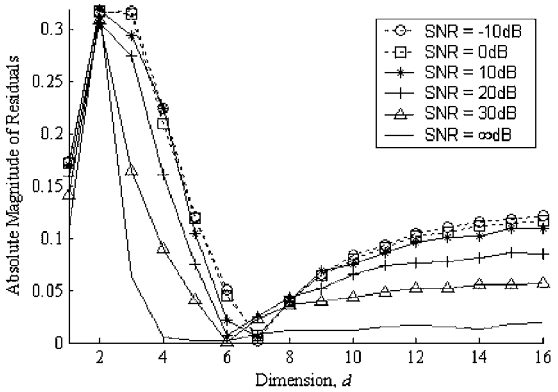

FIG. 3. Absolute residuals from fitting values in Fig. 2 to model given by Eq. (15).

ing GFNN method are discussed in Refs. [8,10].

B. Cao's method

Define

$$a_d(n) \triangleq \frac{\|\boldsymbol{v}_{d+1}(n) - \boldsymbol{v}_{d+1}(n_{NN})\|}{\|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|}, \tag{7}$$

where $n \in [1, N-d\tau]$. The mean of $a_d(n)$ is defined as

$$\langle a_d \rangle \triangleq \frac{1}{N-d\tau} \sum_{n=1}^{N-d\tau} a_d(n).$$

Define

$$E_d^1 \triangleq \frac{\langle a_{d+1} \rangle}{\langle a_d \rangle}; \tag{8}$$

E_d^1 stops changing when $d \geqslant d_E$. Unlike the GFNN method, the embedding dimension for Cao's method is decided not by the presence of a plateau, but by the dimension whereby the curve of E_d^1 vs d saturates (location of knee).

III. PROPOSED METHODS FOR ESTIMATING d_E IN THE PRESENCE OF NOISE

Time series of Lorenz and Rössler systems are generated using the MATLAB© ODE solver `ode45`. Unless otherwise specified, Lorenz data refer to the x component of the Lorenz system (N^{total}=30 000, step size 0.01); Rössler data refers to the x component of the Rössler system (N^{total}=30 000, step size 0.1). Considering Eqs. (2) and (3), the likely range of values of d_E should be $3 \leqslant d_E \leqslant 5$ for both systems. Additive white Gaussian noise (AWGN) and Additive white uniform noise (AWUN) are added to the time series to form signals with different SNR. The implementation of the GFNN method in the TISEAN toolbox [8,11] (`false_nearest.exe`) is used to compute $g_d^{\boldsymbol{v}}$, the proportion of GFNN from $\{\boldsymbol{v}(n)\}_{n=N^{begin}}^{N^{total}}$ [where $N^{begin} \triangleq (d_E - 1)\tau + 1$], and the implementation of Cao's method (`cao.dll`) in the TSTOOL toolbox [12] is used to compute E_d^1.

One problem with the GFNN method is that the location of the plateau becomes difficult to discern when SNR is low. For example, in Fig. 1 it is unclear if d_E=8 when SNR is 10 dB, as d_E=4 may be more appropriate. Also, there will be false neighbors even for high values of d, as noise can be regarded as a high-dimensional signal [13]

Figure 2 is the plot of E_d^1 vs d for Lorenz data for various values of SNR. One problem with Cao's method is that the location of the knee becomes difficult to discern when AWGN is added; it is only possible to estimate $d_E \geqslant 8$ when SNR is 10 dB.

A. Method 1 (based on GFNN method)

Let

$$\boldsymbol{v}(n) \triangleq \breve{\boldsymbol{v}}(n) + \boldsymbol{\eta}(n), \tag{9}$$

where the noiseless embedding vector $\breve{\boldsymbol{v}}(n) \triangleq (\breve{y}(n), \breve{y}(n-\tau), \ldots, \breve{y}(n-(d_E-1)\tau))$ and the noise vector $\boldsymbol{\eta}(n) \triangleq (\eta(n), \eta(n-\tau), \ldots, \eta(n-(d_E-1)\tau))$. Assume

TABLE I. Embedding dimensions (Lorenz data) using proposed methods.

	AWGN					AWUN				
SNR (dB)	τ	Method 1	Method 2	Method 3	Method 4	τ	Method 1	Method 2	Method 3	Method 4
−10	1	3	3	3	3	2	6	4	2	
0	1	7	4	2		1	4	4	2	4
10	15	4	3	2		14	4	3	2	
20	12	3	4	2		12	3	4	2	
30	11	2	3	2	2	11	3	3	2	3
∞	11	2	3	2	2	11	2	3	2	2

TABLE II. Embedding dimensions (Rössler data) using proposed methods.

	AWGN					AWUN				
SNR (dB)	τ	Method 1	Method 2	Method 3	Method 4	τ	Method 1	Method 2	Method 3	Method 4
−10	1	4	3	3	3	2	3	3	2	3
0	15	4	4	3	4	14	3	4	3	3
10	14	4	4	2	4	14	4	5	2	
20	14	3	4	2		14	3	4	2	
30	13	2	4	2	2	13	2	3	2	2
∞	12	2	3	2	2	12	2	3	2	2

$$g_d^{\boldsymbol{v}} = g_d^{\check{\boldsymbol{v}}} + g_d^{\boldsymbol{\eta}}, \tag{10}$$

where $g_d^{\check{\boldsymbol{v}}}$ is the proportion of GFNN from $\{\check{\boldsymbol{v}}(n)\}_{n=N^{begin}}^{N^{total}}$, and $g_d^{\boldsymbol{\eta}}$ is the proportion of GFNN from $\{\boldsymbol{\eta}(n)\}_{n=N^{begin}}^{N^{total}}$. Note that Eq. (10) ignores two possibilities:

(i) both $\check{\boldsymbol{v}}(n)$ and $\boldsymbol{\eta}(n)$ may contribute to the same nearest neighbor ($g_d^{\boldsymbol{\eta}}$ is overestimated); and

(ii) neither $\check{\boldsymbol{v}}(n)$ nor $\boldsymbol{\eta}(n)$ may contribute to a false neighbor, but the combination $\boldsymbol{v}(n)$ may contribute one ($g_d^{\boldsymbol{\eta}}$ is underestimated).

The function $g_d^{\check{\boldsymbol{v}}}$ vs d should decrease monotonically, as the proportion of GFNN should decrease as d increases. If $N^{total} < \infty$, the proportion of GFNN will drop to 0, as the average value of $R_d^2(n, n_\eta)$ will increase with d. The function $g_d^{\boldsymbol{\eta}}$ vs d could be approximated by a constant, as the proportion of GFNN is independent of d. Thus, $g_d^{\check{\boldsymbol{v}}} + g_d^{\boldsymbol{\eta}}$ results in a plateau. Since $g_d^{\boldsymbol{\eta}}$ cannot exceed 1, Eq. (10) can be changed to

$$g_d^{\boldsymbol{v}} = \max((g_d^{\check{\boldsymbol{v}}} + g_d^{\boldsymbol{\eta}}), 1). \tag{11}$$

A systematic way to locate the plateau is to find the maximum of the second-order central difference.

B. Method 2 (based on Cao's method)

Observe that the E_d^1 plots in Fig. 2 converge to 1. Using the triangle inequality on (7)

$$\frac{\|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|}{\|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|} \leqslant \frac{\|\boldsymbol{v}_{d+1}(n) - \boldsymbol{v}_{d+1}(n_{NN})\|}{\|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|} \leqslant 1 + \frac{\|y(n - d\tau) - y(n_{NN} - d\tau)\|}{\|\boldsymbol{v}_d(n) - \boldsymbol{v}_d(n_{NN})\|}. \tag{12}$$

Using the squeeze theorem

$$\lim_{d\to\infty} a_d(n) = 1. \tag{13}$$

Thus, $E_d^1 \approx 1$ for large d, suggesting that E_d^1 will converge to 1. A systematic way to identify the location of the knee for Cao's method is to find the minimum of the second-order central difference.

C. Method 3 (based on Cao's method)

For the case where the data are pure noise, E_d^1 should not change much as d is varied. This can be expressed by

$$E_d^1 + \frac{\delta E_d^1}{\delta d}\Delta d \approx b, \tag{14}$$

where $b \in \mathbb{R}^+$ is a constant, and the change in dimension is $\Delta d = 1$, as $d \in \mathbb{Z}^+$. From (14)

$$\frac{\delta E_d^1}{b - E_d^1} \approx \delta d$$

$$E_d^1 = b - e^{-d-c}, \tag{15}$$

where $c \in \mathbb{R}$ is a constant. The values of b and c are estimated using nonlinear least-squares (Gauss-Newton method) with the initial values $b=1$ [from Eq. (13)] and $c=0$, to obtain a close fit to E_d^1 plots under low SNR. Note that $b \neq 1$ as $d \nrightarrow \infty$. The value of d_E corresponds to the value of d at which the E_d^1 plots deviate most from the model of Eq. (15); see Fig. 3 (the residual is the difference between E_d^1 and $b - e^{-d-c}$ vs d).

D. Method 4

As method 1 seems to underestimate d_E, while method 2 seems to overestimate d_E (Tables I and II), good estimates of

TABLE III. Embedding dimensions (Lorenz data) using proposed methods and prior information.

	AWGN					AWUN				
SNR (dB)	τ	Method 1	Method 2	Method 3	Method 4	τ	Method 1	Method 2	Method 3	Method 4
−10	1	3	3	3	3	2	6	4	3	
0	1	7	4	3		1	4	4	3	4
10	15	4	3	3	3	14	4	3	3	3
20	12	3	4	3	3	12	3	4	3	3
30	11	3	3	3	3	11	3	3	3	3
∞	11	3	3	3	3	11	3	3	3	3

TABLE IV. Embedding dimensions (Rössler data) using proposed methods and prior information.

	AWGN					AWUN				
SNR (dB)	τ	Method 1	Method 2	Method 3	Method 4	τ	Method 1	Method 2	Method 3	Method 4
−10	1	4	3	3	3	2	3	3	3	3
0	15	4	4	3	4	14	3	4	3	3
10	14	4	4	3	4	14	4	5	3	
20	14	3	4	3	3	14	3	4	3	3
30	13	3	4	3	3	13	3	3	3	3
∞	12	3	3	3	3	12	3	3	3	3

d_E should occur when methods 1 and 2 agree (similar idea to sandwich theorem). However, such a criterion would be too strict, and it may be practical to check for agreement with method 3 as well. Thus, a voting scheme is suggested: d_E is the value selected most frequently by methods 1–3. No answer is given in the event of a tie.

E. Prior knowledge

For Lorenz data, $D_0 \approx 2.06$ [14], which is close to 2; thus, relatively few GFNN are generated when $d=2$. Thus, method 1 will estimate $d_E=2$, although it is theoretically infeasible [15], as $g_d^{\boldsymbol{v}}$ will experience a gentle slope between $d=2$ and $d=3$. Thus, the maximum second-order central difference which results in $d_E>2$ should be selected. Prior knowledge can be incorporated for methods 1–3: the values of d_E obtained are constrained such that $d_E>2$ (Tables III and IV). However, prior knowledge may be of limited use for high-dimensional systems (hyperchaos).

IV. RESULTS AND DISCUSSIONS

Tables I and II show that the value of d_E selected by method 1 tends to increase as SNR decreases (for both AWGN and AWUN). Method 2 tends to favor higher values of d_E, compared to method 3. Method 3 (Tables III and IV) seems to be most reliable, but this observation may be a little naive, as $d_E \leqslant 3$ even for negative SNR. Thus, method 3 should be double-checked by methods 1 and 2, as they are more sensitive to changes in SNR. This suggests method 4, which selects $2 \leqslant d_E \leqslant 4$ for SNR$\geqslant 0$ dB (Tables I and II). However, in many cases, method 4 fails to obtain an answer because there is no majority. When prior information is incorporated, method 4 works much better, and usually selects $d_E=3$ (except $d_E=4$ for SNR of 0 dB or 10 dB).

A few words of caution are in order. Since finding derivatives is not a robust procedure, there is no guarantee that methods 1 and 2 work. Note that $g_d^{\boldsymbol{v}}$ varies with N^{total}[4]; similarly, E_d^1 varies with N^{total} [7]. Furthermore, the plots of $g_d^{\boldsymbol{v}}$ vs d vary with SNR, and there is no simple model for the $g_d^{\boldsymbol{v}}$ plots, unlike Eq. (15) for Cao's method.

For SNR$\leqslant 0$ dB, the values of τ tend to become very low (Table I–IV). Thus, inappropriate embeddings may be constructed, even though the "correct" value of d_E is estimated. In any case, results at low SNR should be taken with a grain of salt.

V. CONCLUSION

A simple model [Eq. (15)] is proposed to explain the curves commonly obtained when using Cao's method to estimate d_E. Simple rules are suggested to extract embedding dimension (methods 1 and 2). A voting scheme (method 4) is suggested, which gives $d_E=3$ even with low SNR, for both Lorenz and Rössler systems (systems with 3 degrees of freedom). Prior information that the systems are dissipative is incorporated to improve the estimates of d_E.

[1] F. Takens, in *Dynamical Systems and Turbulence (Warwick 1980)*, edited by D. A. Rand and L. S. Young (Springer-Verlag, Berlin, 1980), Vol. 898, p. 366.

[2] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996).

[3] T. Sauer *et al.*, J. Stat. Phys. **65**, 579 (1991).

[4] M. B. Kennel *et al.*, Phys. Rev. A **45**, 3403 (1992).

[5] M. Ding *et al.*, Physica D **69**, 404 (1993).

[6] L. Cao (private communication).

[7] L. Cao, Physica D **110**, 43 (1997).

[8] R. Hegger and H. Kantz, Phys. Rev. E **60**, 4970 (1999).

[9] C. C. Aggarwal *et al.*, Database Theory—ICDT 2001. 8th International Conference Proceedings, Lecture Notes in Computer Science, Vol. 1973, 420 (2001).

[10] D. R. Fredkin and J. A. Rice, Phys. Rev. E **51**, 2950 (1995).

[11] R. Hegger *et al.*, Chaos **9**, 413 (1999).

[12] C. Merkwirth *et al.*, in *Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, edited by J. A. Suykens and J. Vandewalle (Katholieke Universiteit, Leuven, Belgium, 1998), p. 144.

[13] H.-H. Kuo, *White Noise Distribution Theory* (CRC Press, Boca Raton, FL, 1996).

[14] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[15] H. Haken, Phys. Lett. **94**A, 71 (1983).