Delay time window and plateau onset of the correlation dimension for small data sets

H. S. Kim*
Department of Civil Engineering, Colorado State University, Fort Collins, Colorado 80523

R. Eykholt
Department of Physics, Colorado State University, Fort Collins, Colorado 80523

J. D. Salas
Department of Civil Engineering, Colorado State University, Fort Collins, Colorado 80523

(Received 9 January 1998)

The method of delays is widely used for reconstructing chaotic attractors from experimental observations. Many studies have used a fixed delay time τ_d as the embedding dimension m is increased, but this is not necessarily the best choice for obtaining good convergence of the correlation dimension. Recently, some researchers have suggested that it is better to fix the delay time window τ_w instead. Unfortunately, τ_w cannot be estimated using either the autocorrelation function or the mutual information, and no standard procedure for estimating τ_w has yet emerged. However, the recently introduced $C-C$ method can be used to estimate either τ_d or τ_w. Using this method, we show that, for small data sets, fixing τ_w, rather than τ_d, does indeed lead to a more rapid convergence of the correlation dimension as the embedding dimension m is increased.
[S1063-651X(98)05111-3]

PACS number(s): 05.45.+b, 47.52.+j

I. INTRODUCTION

Much progress has been made in understanding chaotic physical processes in science and engineering. To quantify the chaotic behavior of a time series, one often calculates the correlation dimension. The first step in this calculation is the reconstruction of the chaotic attractor from the experimental observations. The standard technique for attractor reconstruction is the method of delays developed by Packard *et al.* [1] and Takens [2]. This method embeds the finite time series $\{x_i\}$, $i=1,2,\dots,N$, into an m-dimensional space by defining the vectors

$$\vec{x}_i=(x_i,x_{i+t},x_{i+2t},\dots,x_{i+(m-1)t}), \quad \vec{x}_i\in\mathbb{R}^m, \tag{1}$$

where t is the index lag, and the number of vectors is $M=N-(m-1)t$. If the sampling time is τ_s, then the delay time is $\tau_d=t\tau_s$. One advantage of this method is that it yields the same noise level for each component of the state vector.

Since the components of the reconstructed vectors need to be independent, the quality of the reconstructed attractor depends on the choice of the delay time τ_d. If τ_d is too small, the reconstructed attractor is compressed along the identity line, and this is called redundance. If τ_d is too large, the attractor dynamics may become causally disconnected, and this is called irrelevance [3]. Most researchers have used a fixed value of τ_d, independent of the embedding dimension m, and this is usually selected using either the autocorrelation function or the mutual information. The latter approach is more reliable, but it also requires larger data sets and greater computation time than the former method. We recently introduced a method for estimating τ_d, called the $(C-C)$ method, which yields the same results as the mutual information, but which can be used with much smaller data sets, and which is more efficient computationally [4].

On the other hand, several researchers [5–9] have suggested that, rather than using a fixed delay time τ_d for various embedding dimensions m, it may be more appropriate to fix the delay time window $\tau_w=(m-1)\tau$, which is the entire time spanned by the components of each embedded vector $\vec{x}_i$ (in practice, τ_w cannot be completely fixed, since the delay time τ must be rounded off to the nearest integer multiple of the sampling time τ_s). Unfortunately, the estimation of τ_w is not fully developed, and Martinerie *et al.* [9] have shown that neither the autocorrelation function nor the mutual information can give τ_w. However, the $C-C$ method can be used to find τ_w, as well as τ_d [4]. Basically, τ_w is the optimal time for independence, while τ_d is the first locally optimal time.

Using a fixed delay time τ_d does not necessarily lead to good convergence of the correlation dimension as the embedding dimension m increases [10], and it may result in the undesirable blurring of the information from two (or more) states as the number of delay coordinates increases [7]. However, using a fixed delay time window τ_w can lead to near-minimum redundance, while keeping the irrelevance at an acceptable level [7]. Using the $C-C$ method to estimate both τ_d and τ_w, this study shows that using τ_w, rather than τ_d, does indeed lead to a more rapid convergence of the correlation dimension for small data sets. Since the estimation of τ_w using the $C-C$ method is efficient computationally, is robust to noise, and does not require large data sets [4], then the use of τ_w, rather than τ_d, should become the standard procedure.

*Present address: Department of Construction Engineering, Sun Moon University, Asan-Si, Korea.

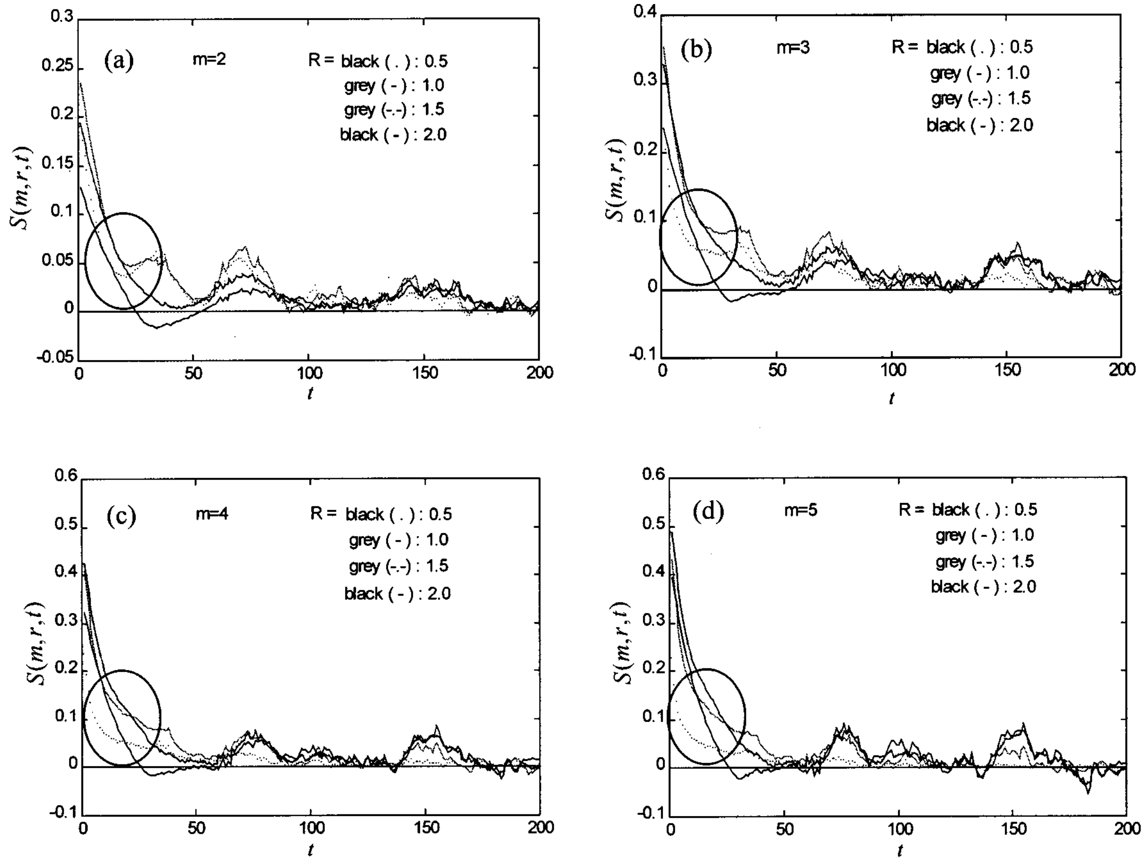

FIG. 1. $S(m,r,t)$ for the variable x from the Lorenz system of Eq. (15) with $a=10.0$, $b=28.0$, $c=8/3$, and $\tau_s=0.01$ using 3000 data points. The circles indicate the vicinity of τ_d, where the first local minimum occurs in the variation of $S(m,r,t)$ with r. Note that $R=r/\sigma$.

II. MEASURE OF NONLINEAR DEPENDENCE

A. Correlation integral and BDS statistic

The correlation dimension introduced by Grassberger and Procaccia [11] is widely used in many fields for the characterization of strange attractors. The correlation integral for the embedded time series is the following function:

$$C(m,N,r,t)=\frac{2}{M(M-1)}\sum_{1\leq i<j\leq M}\Theta(r-\|\vec{x}_i-\vec{x}_j\|),\quad r>0, \tag{2}$$

where

$$\Theta(a)=\begin{cases}0, & \text{if } a<0\\ 1, & \text{if } a\geq 0,\end{cases}$$

N is the size of the data set, t is the index lag, $M=N-(m-1)t$ is the number of embedded points in m-dimensional space, and $\|\cdot\cdot\|$ denotes the sup-norm. $C(m,N,r,t)$ measures the fraction of the pairs of points $\vec{x}_i$, $i=1,2,\ldots,M$, whose sup-norm separation is no greater than r. If the limit of $C(m,N,r,t)$ as $N\rightarrow\infty$ exists for each r, we write the fraction of all state vector points that are within r of each other as $C(m,r,t)=\lim_{N\rightarrow\infty}C(m,N,r,t)$, and the correlation dimension is defined as $D_2(m,t)=\lim_{r\rightarrow 0}[\log_{10}C(m,r,t)/\log_{10}r]$. In practice, N remains finite, and, thus, r cannot go to zero; instead, we look for a linear region of slope $D_2(m,t)$ in the plot of $\log_{10}C(m,N,r,t)$ versus $\log_{10}r$.

Brock *et al.* [12,13] studied the BDS statistic, which is based on the correlation integral, to test the null hypothesis that the data are independently and identically distributed (*iid*). This test has been particularly useful for chaotic systems and nonlinear stochastic systems.

Under the *iid* hypothesis, the Brock-Dechert-Scheinkman (BDS) statistic for $m>1$ is defined as

$$S_{\mathrm{BDS}}(m,M,r)=\frac{\sqrt{M}}{\sigma(m,M,r)}[C(m,M,r)-C^m(1,M,r)], \tag{3}$$

and this converges to a standard normal distribution as $M\rightarrow\infty$. Note that the asymptotic variance $\sigma^2(m,M,r)$ can be estimated as

$$\sigma^2(m,M,r)=4\Bigg\{m(m-1)\hat{C}^{2(m-1)}(\hat{K}-\hat{C}^2)+\hat{K}^m-\hat{C}^{2m}+2\sum_{i=1}^{m-1}[\hat{C}^{2i}(\hat{K}^{m-i}-\hat{C}^{2(m-i)})-m\hat{C}^{2(m-i)}(\hat{K}-\hat{C}^2)]\Bigg\}, \tag{4}$$

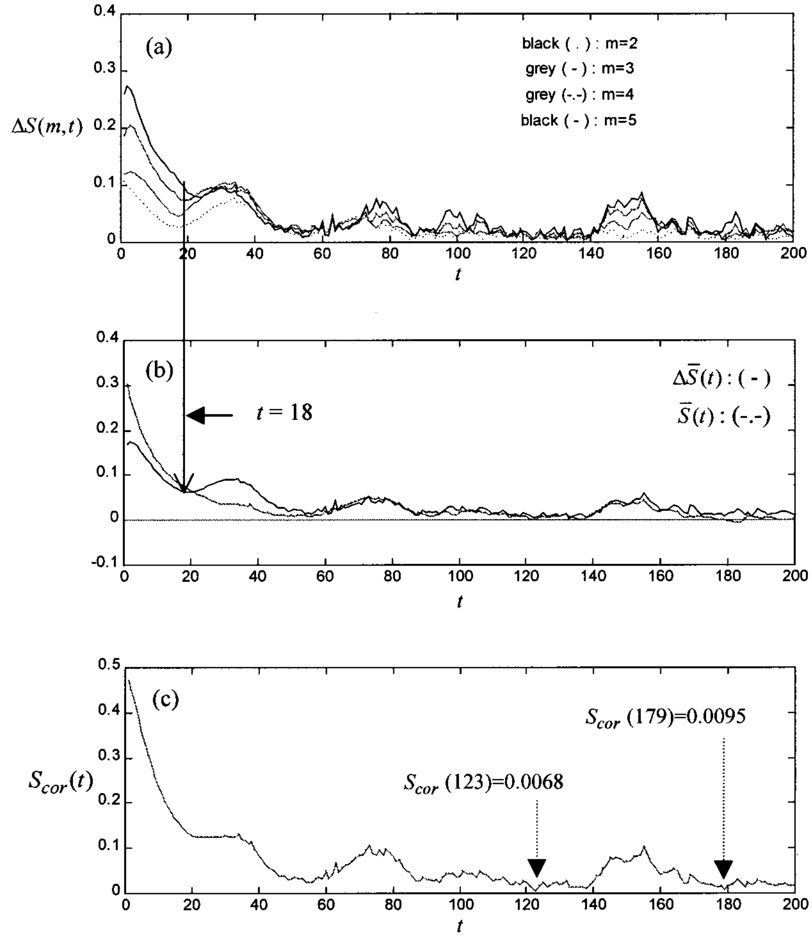

FIG. 2. $\Delta S(m,t)$, $\Delta \overline{S}(t)$, $\overline{S}(t)$, and $S_{\text{cor}}(t)$ for the variable x from the Lorenz system of Fig. 1. The solid line locates $\tau_d = 18\tau_s$, and the minimum of $S_{\text{cor}}(t)$ yields $\tau_w = 123\tau_s$.

$$\hat{C}(m,M,r) = \frac{2}{M(M-1)} \sum_{1 \leq i < j \leq M} \Theta(r - \|\vec{x}_i - \vec{x}_j\|), \quad (5)$$

$$\hat{K}(m,M,r) = \frac{6}{M(M-1)(M-2)} \sum_{1 \leq i < j < k \leq M} \times \Theta(r - \|\vec{x}_i - \vec{x}_j\|)\Theta(r - \|\vec{x}_j - \vec{x}_k\|). \quad (6)$$

The BDS statistic originates from the statistical properties of the correlation integral, and it measures the statistical significance of calculations of the correlation dimension. Even though the BDS statistic cannot be used to distinguish between a nonlinear deterministic system and a nonlinear stochastic system, it is a powerful tool for distinguishing random time series from the time series generated by chaotic or nonlinear stochastic processes. Its statistical properties, along with proofs, can be found in the literature [12,13].

B. $C-C$ method

The present study is concerned with the properties of the quantity $S(m,N,r,t) = C(m,N,r,t) - C^m(1,N,r,t)$. We refer to a comment by Brock *et al.* [12]: ''If a stochastic process $\{x_i\}$ is *iid*, it will be shown that $C(m,r) = C^m(1,r)$ for all m and r. That is to say, the correlation integral behaves much like the characteristic function of a serial string in that the correlation integral of a serial string of independent random variables is the product of the correlation integrals of component substrings.'' This led us to interpret the statistic $S(m,N,r,t)$ as a nonlinear analog of the serial correlation of a nonlinear time series. More precisely, it can be regarded as a dimensionless measure of nonlinear dependence, and it can be used to determine an appropriate index lag t. For fixed m, N, and r, a plot of $S(m,N,r,t)$ versus t is a nonlinear analog of the plot of the autocorrelation function versus t.

In order to study the nonlinear dependence and eliminate spurious temporal correlations, we must subdivide the time series $\{x_i\}$, $i = 1,2,\ldots,N$, into t disjoint time series of size N/t. $S(m,N,r,t)$ is then computed from the t disjoint time series as follows:

TABLE I. Summary of results for three dynamical systems.

				$C-C$ Method	
System	Parameters	Variable	τ_s	τ_d	τ_w
Lorenz	$a = 10.0$, $b = 28.0$, $c = 8/3$	x	0.01	0.18	1.23
Rabinovich-Fabrikant	$\gamma = 0.87$, $\alpha = 1.1$	x	0.01	0.52	1.28
Three-torus		x	1.00	55	101

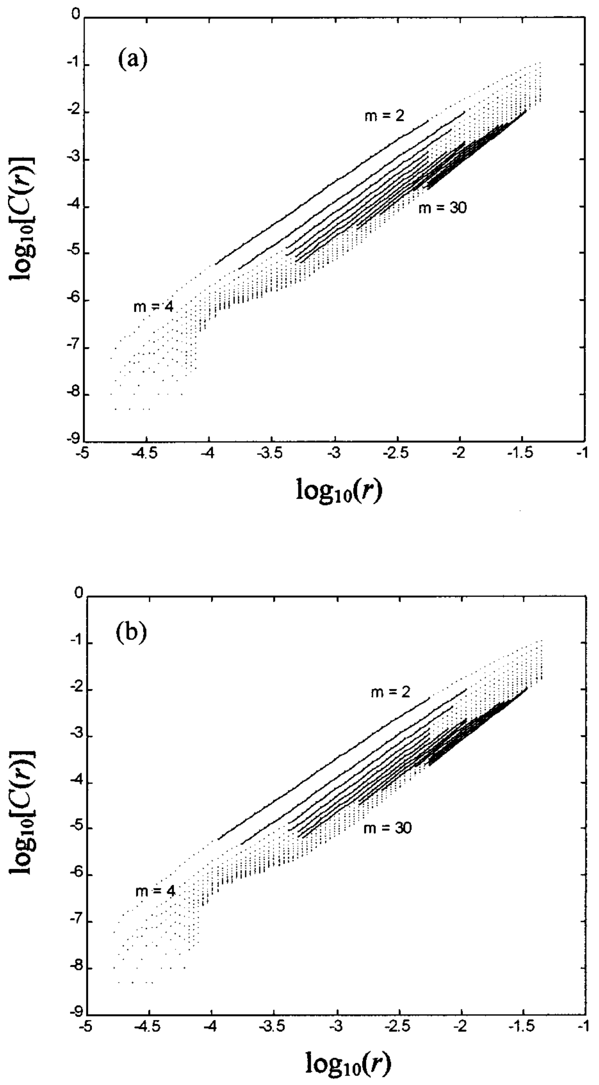

FIG. 3. Correlation integrals for 20 000 data points generated from the Lorenz system of Eq. (15) using (a) $\tau_d = 18\tau_s$ and (b) $\tau_w = 123\tau_s$.

For $t=1$, we have the single time series $\{x_1, x_2, \ldots, x_N\}$, and

$$S(m,N,r,1) = C(m,N,r,1) - C^m(1,N,r,1). \quad (7)$$

For $t=2$, we have the two disjoint time series $\{x_1, x_3, \ldots, x_{N-1}\}$ and $\{x_2, x_4, \ldots, x_N\}$, each of length $N/2$, and we average the values of $S(m,N/2,r,1)$ for these two series:

$$S(m,N,r,2) = \tfrac{1}{2}\{[C_1(m,N/2,r,2) - C_1^m(1,N/2,r,2)] + [C_2(m,N/2,r,2) - C_2^m(1,N/2,r,2)]\}. \quad (8)$$

For general t, this becomes

$$S(m,N,r,t) = \frac{1}{t}\sum_{s=1}^{t} [C_s(m,N/t,r,t) - C_s^m(1,N/t,r,t)]. \quad (9)$$

Finally, as $N \rightarrow \infty$, we can write

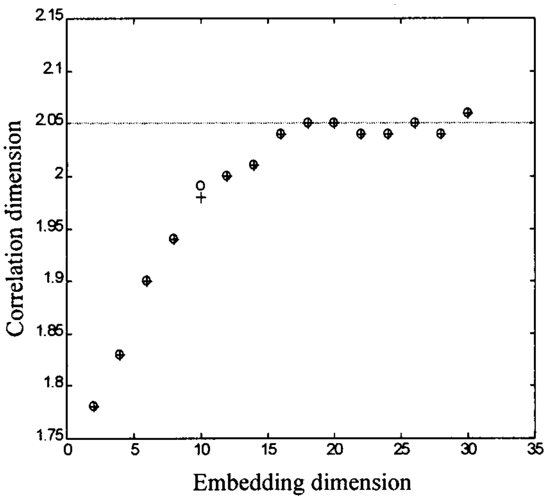

FIG. 4. Plateau onset of the correlation dimension for 20 000 data points generated from the Lorenz system using $\tau_d = 18\tau_s$ (circles) and $\tau_w = 123\tau_s$ (crosses). A horizontal line is drawn at the true correlation dimension of $D_2 = 2.05$.

$$S(m,r,t) = \frac{1}{t}\sum_{s=1}^{t} [C_s(m,r,t) - C_s^m(1,r,t)], \quad m = 2,3, \ldots . \quad (10)$$

For fixed m and t, $S(m,r,t)$ will be identically equal to zero for all r if the data is *iid* and $N \rightarrow \infty$. However, real data sets are finite, and the data may be serially correlated, so, in general, we will have $S(m,r,t) \neq 0$. Thus, the locally optimal times for independence of the data may be either the zero crossings of $S(m,r,t)$ or the times at which $S(m,r,t)$ shows the least variation with r, since this indicates a nearly uniform distribution of points (since a uniform distribution is length-scale invariant). Hence, we select several representative values r_j, and we define the quantity

$$\Delta S(m,t) = \max\{S(m,r_j,t)\} - \min\{S(m,r_j,t)\}, \quad (11)$$

which is a measure of the variation of $S(m,r,t)$ with r. The locally optimal times t are then the zero crossings of $S(m,r,t)$ and the minima of $\Delta S(m,t)$. In the first case, the zero crossings should be nearly the same for all m and r, and, in the second case, the minima should be nearly the same for all m (otherwise, the time is not locally optimal). The delay time τ_d will correspond to the first of these locally optimal times.

In determining the nonlinear dependence of a finite time series by using the statistic $S(m,N,r,t)$, one must have criteria for selecting the values of m and r. In addition, one must know the role of the sample size N. For a fixed value of N, as m becomes large, the data become very sparse, so that $C(m,N,r,t)$ becomes vanishingly small. Also, if r exceeds the size of the attractor, then $C(m,N,r,t)$ saturates, since all pairs of points are within the distance r. Thus, neither m nor r should be too large.

Brock *et al.* [12] investigated the BDS statistic for time series generated from six distributions in order to determine what values of m and r are appropriate. Time series with three sample sizes, $N = 100$, 500, and 1000, were generated by Monte Carlo simulation from six distributions: a standard

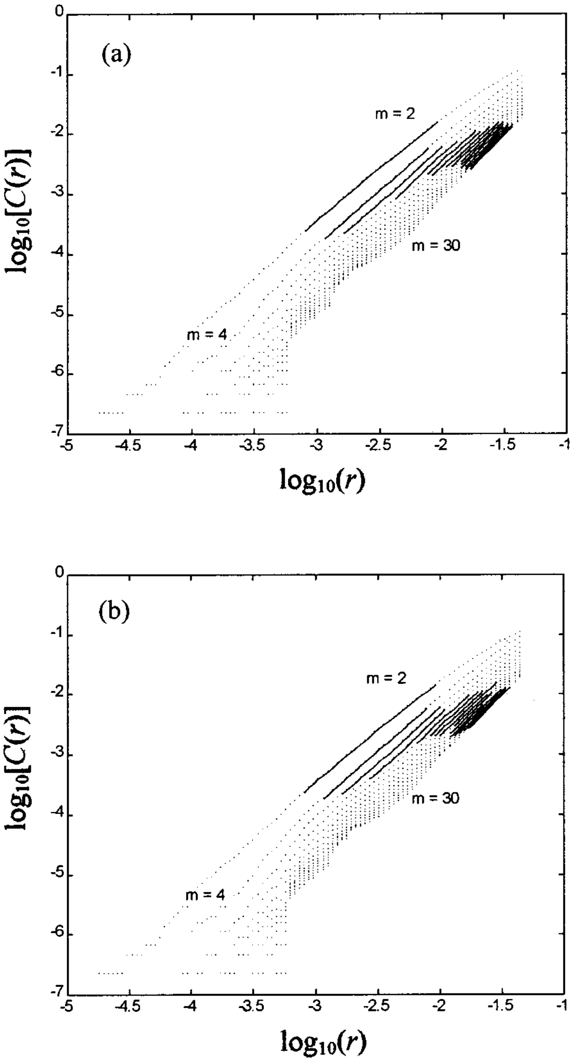

FIG. 5. Correlation integrals for 3000 data points generated from the Lorenz system using (a) $\tau_d=18\tau_s$ and (b) $\tau_w=123\tau_s$.

normal distribution, a student-t distribution with 3 degrees of freedom, a double exponential distribution, a chi-square distribution with 4 degrees of freedom, a uniform distribution, and a bimodal mixture of normal distributions. These studies led to the conclusion that m should be between 2 and 5, and r should be between $\sigma/2$ and 2σ. In addition, the assumed distributions were well approximated by finite time series when $N \geqslant 500$. Note that examining the statistic $S(m,r,t)$ only for $2 \leqslant m \leqslant 5$ does not restrict its use to systems for which the correlation dimension lies in this range.

Thus, we select four values of r in the range $\sigma/2 \leqslant r \leqslant 2\sigma$, $r_1=(0.5)\sigma$, $r_2=(1.0)\sigma$, $r_3=(1.5)\sigma$, and $r_4=(2.0)\sigma$, as representative values. Rather than examining $S(m,r,t)$ and $\Delta S(m,t)$ for all of these values of m and r, we instead examine the averages

$$\overline{S}(t)=\frac{1}{16}\sum_{m=2}^{5}\sum_{j=1}^{4} S(m,r_j,t), \tag{12}$$

$$\Delta\overline{S}(t)=\frac{1}{4}\sum_{m=2}^{5}\Delta S(m,t). \tag{13}$$

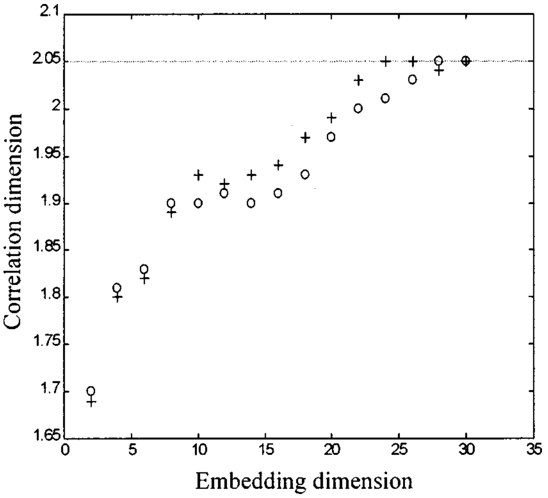

FIG. 6. Plateau onset of the correlation dimension for 3000 data points generated from the Lorenz system using $\tau_d=18\tau_s$ (circles) and $\tau_w=123\tau_s$ (crosses). A horizontal line is drawn at the true correlation dimension of $D_2=2.05$.

We look for the first zero crossing of $\overline{S}(t)$ or the first local minimum of $\Delta\overline{S}(t)$ for finding the first locally optimal time for independence of the data, and this gives the delay time $\tau_d=t\tau_s$. The optimal time is the index lag t for which $\overline{S}(t)$ and $\Delta\overline{S}(t)$ are both closest to zero. If we assign equal importance to these two quantities, then we may simply look for the minimum of the quantity

$$S_{\text{cor}}(t)=\Delta\overline{S}(t)+|\overline{S}(t)|, \tag{14}$$

and this optimal time gives the delay time window $\tau_w=t\tau_s$.

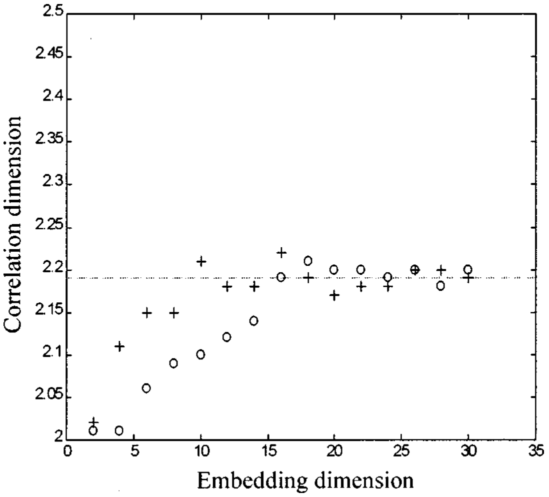

FIG. 7. Plateau onset of the correlation dimension for 3000 data points generated from the Rabinovich-Fabrikant system of Eq. (16) for $\gamma=0.87$ and $\alpha=1.1$ using $\tau_d=52\tau_s$ (circles) and $\tau_w=128\tau_s$ (crosses). A horizontal line is drawn at the true correlation dimension of $D_2=2.19$.

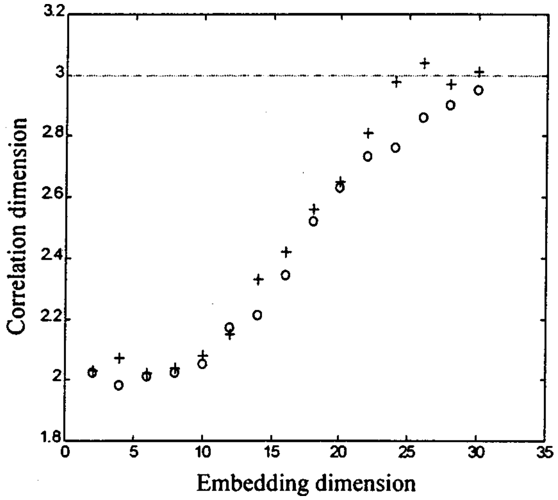

FIG. 8. Plateau onset of the correlation dimension for 3000 data points generated from the three-torus of Eq. (17) using $\tau_d = 55$ (circles) and $\tau_w = 101$ (crosses). A horizontal line is drawn at the true correlation dimension of $D_2 = 3.0$.

III. PLATEAU ONSET OF THE CORRELATION DIMENSION

In this study, we consider the following three systems: the Lorenz system [11]

$$\dot{x} = -a(x-y),$$

$$\dot{y} = -xz + cx - y, \tag{15}$$

$$\dot{z} = xy - bz,$$

the Rabinovich-Fabrikant system [14]

$$\dot{x} = y(z - 1 + x^2) + \gamma x,$$

$$\dot{y} = x(3z + 1 - x^2) + \gamma y, \tag{16}$$

$$\dot{z} = -2z(\alpha + xy),$$

and the three-torus [9]

$$x_i = \sin\left[\frac{3i}{500}\right] + \sin\left[\frac{3\sqrt{2}i}{250}\right] + \sin\left[\frac{9\sqrt{3}i}{500}\right]. \tag{17}$$

For the Lorenz system, we solve the system of equations for $a = 10.0$, $b = 28.0$, and $c = 8/3$ to generate a time series of the variable x with $\tau_s = 0.01$. We then compute $S(m,r,t)$ from Eq. (10), and the results are shown in Fig. 1. The circles in Fig. 1 indicate the index lag t where the variation of $S(m,r,t)$ with r is at its first local minimum, and Fig. 2(a) shows this first local minimum of $\Delta S(m,t)$ more clearly. We choose the delay time at this point, which gives $\tau_d = 18\tau_s = 0.18$ [see Fig. 2(b)]. This agrees with the delay time $\tau_d = 0.17$ found by Martinerie *et al.* [9] using the first local minimum of the mutual information. Also, from the minimum of $S_{\rm cor}(t)$ in Fig. 2(c), we choose the delay time window $\tau_w = 123\tau_s = 1.23$. Similar analyses are performed for the other two systems, and the delay times and the delay time windows obtained by the $C-C$ method for the three systems are summarized in Table I. These results are very robust to the addition of noise, as shown in Ref. [4].

The correlation integrals for the Lorenz system, using the fixed value of τ_d and the fixed value of τ_w are computed for $N = 20\,000$ data points, and the results are shown in Fig. 3. From the linear regions of these correlation integrals (which have been darkened in Fig. 3), the correlation dimensions are calculated, and these results are shown in Fig. 4, together with the value $D_2 = 2.05$ obtained by Grassberger and Procaccia [11]. The two sets of results are virtually identical, and the plateau onset occurs at about $m = 16$.

Next, we perform a similar analysis for a small data set with only $N = 3000$ data points. The correlation integrals based on τ_d and τ_w are drawn in Fig. 5, and the correlation dimensions are shown in Fig. 6. For the fixed value of τ_w, the plateau onset occurs at about $m = 24$, but, for the fixed value of τ_d, the plateau onset does not occur until about $m = 28$. This difference in embedding dimensions represents a substantial difference in the amount of computer time required to obtain the correlation dimension.

We solve the Rabinovich-Fabrikant system of Eq. (16) for $\gamma = 0.87$ and $\alpha = 1.1$, and we generate a time series of 3000 data points for the variable x with $\tau_s = 0.01$. The correlation dimensions based on the values of τ_d and τ_w given in Table I are shown in Fig. 7, along with the value $D_2 = 2.19$ found in Ref. [11]. The plateau onset for the correlation dimension obtained using τ_w occurs at about $m = 10$, but the plateau onset obtained using τ_d does not occur until about $m = 16$.

Figure 8 shows similar results for the three-torus of Eq. (17) using the values of τ_d and τ_w given in Table I. Using the fixed value of τ_w causes the correlation dimension to saturate at the correct value of $D_2 = 3$ at about $m = 24$, but, when the fixed value of τ_d is used, saturation has still not occurred for $m = 30$.

IV. CLOSING REMARKS

In this study, we have shown that, for small data sets, using a fixed delay time window τ_w, rather than a fixed delay time τ_d, leads to a more rapid convergence of the correlation dimension as the embedding dimension m is increased. Although no standard technique for estimating τ_w has yet emerged, we have shown that the $C-C$ method is well suited to this task. Furthermore, this method is efficient computationally, it is robust to noise, and it may be used for small data sets. As a result, the use of a fixed value of τ_w, rather than a fixed value of τ_d, should become standard practice. This is particularly important in fields such as hydrology and atmospheric science, where small noisy data sets are common.

ACKNOWLEDGMENT

This research has been partially supported by NSF Grant No. CMS-9625685 on ‘‘Uncertainty and Risk Analysis Under Extreme Hydrologic Events.’’

[1] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).

[2] F. Takens, in *Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young, Lecture Notes in Mathematics, Vol. 898 (Springer, Berlin, 1981), p. 336.

[3] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, Physica D **51**, 52 (1991).

[4] H. S. Kim, R. Eykholt, and J. D. Salas, Physica D (to be published).

[5] D. S. Broomhead and G. P. King, Physica D **20**, 217 (1986).

[6] A. M. Albano, J. Muench, C. Schwartz, A. I. Mees, and P. E. Rapp, Phys. Rev. A **38**, 3017 (1988).

[7] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, Physica D **73**, 82 (1994).

[8] D. Kugiumtzis, Physica D **95**, 13 (1996).

[9] J. M. Martinerie, A. M. Albano, A. I. Mees, and P. E. Rapp, Phys. Rev. A **45**, 7058 (1992).

[10] Z. B. Wu, Physica D **85**, 485 (1996).

[11] P. Grassberger and I. Procaccia, Physica D **7**, 153 (1983).

[12] W. A. Brock, D. A. Hsieh, and B. Lebaron, *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence* (MIT Press, Cambridge, 1991).

[13] W. A. Brock, W. D. Dechert, J. A. Scheinkman, and B. LeBaron, Econ. Rev. **15**, 197 (1996).

[14] M. I. Rabinovich and A. L. Fabrikant, Zh. Eksp. Teor. Fiz. **77**, 617 (1979) [Sov. Phys. JETP **50**, 311 (1979)].