Comparative study of embedding methods

C. J. Cellucci,[1,2,4,*] A. M. Albano,[2] and P. E. Rapp[3,4]
[1]*Department of Physics, Ursinus College, Collegeville, Pennsylvania 19426, USA*
[2]*Department of Physics, Bryn Mawr College, Bryn Mawr, Pennsylvania 19010, USA*
[3]*Department of Pharmacology and Physiology, Drexel University College of Medicine, Philadelphia, Pennsylvania 19129, USA*
[4]*Arthur P. Noyes Research Foundation, Norristown State Hospital, Norristown, Pennsylvania 19401, USA*
(Received 3 December 2002; published 23 June 2003)

Embedding experimental data is a common first step in many forms of dynamical analysis. The choice of appropriate embedding parameters (dimension and lag) is crucial to the success of the subsequent analysis. We argue here that the optimal embedding of a time series cannot be determined by criteria based solely on the time series itself. Therefore we base our analysis on an examination of systems that have explicit analytic representations. A comparison of analytically obtained results with those obtained by an examination of the corresponding time series provides a means of assessing the comparative success of different embedding criteria. The assessment also includes measures of robustness to noise. The limitations of this study are explicitly delineated. While bearing these limitations in mind, we conclude that for the examples considered here, the best identification of the embedding dimension was achieved with a global false nearest neighbors argument, and the best value of lag was identified by the mutual information function.

DOI: 10.1103/PhysRevE.67.066210 PACS number(s): 05.45.−a

I. INTRODUCTION

Embedding experimental data is a first step common to many forms of dynamical analysis. In this process a scalar time series $\{x_1,x_2,\ldots,x_n\}$ is used to construct vectors in $\Re^m$ of the form $X_i=(x_i,x_{i+L},x_{i+2L},\ldots,x_{i+(m-1)L})$, where m is the embedding dimension and L is the lag. For proper values of m and L a smooth dynamics F: $X_i\rightarrow X_{i+1}$ is defined which reconstructs the underlying dynamics. Measures of dynamical behavior are then based on the quantitative characterization of the m-dimensional geometry of the set $\{X_i\}$. The mathematical foundation of this procedure is the Takens-Mañé embedding theorem [1,2]. This result has been reviewed by Noakes [3] and Sauer, Yorke, and Casdagli [4]. A summary statement of the theorem is given in the Appendix.

The choice of embedding parameters m and L is crucial to the subsequent analysis. An inappropriate choice can result in the spurious indication of nonlinear structure where none is present [5,6]. Conversely, an inappropriate choice can result in the failure to resolve structures that are indeed present in the data. There is a large, growing, and somewhat conflicting literature describing candidate criteria for selecting embedding parameters [7–22].

There is no single correct answer. The optimal embedding strategy may depend on both the time series and the applied measure. That is, the embedding criterion that is optimal when studying fluid flow data may not be optimal in the analysis of a time series from an electroencephalogram. Similarly, a procedure for selecting m and L when the correlation dimension is to be estimated may not succeed when calculating Lyapunov exponents. Therefore the limitations of this investigation should be explicitly recognized. While optimistically we hope to distinguish the methods that are effective for a majority of time series and applied measures, the minimal result should be the identification of those embedding methods that are most appropriate for a specific time series and applied measure. This is, however, a better alternative than arbitrary parameter specification.

An examination of the prior literature on this subject reveals a most interesting problem. Suppose two embedding criteria are used to select embedding parameters for a time series. Let (m_1,L_1) and (m_2,L_2) denote the results. Which is the better embedding? To answer this question we need an adjudicating measure M, such that if $M_1=M(m_1,L_1)$ is greater than $M_2=M(m_2,L_2)$ we conclude that the first embedding is the better of the two. Following this reasoning, a program of comparison testing of embedding criteria consists of two elements: (i) competing embedding criteria that are used to select embedding parameters m and L, and (ii) a metric M that is used to choose between them. Unfortunately, this program has a fundamental logical flaw. The adjudicating measure M is itself an embedding criterion. By construction, the best embedding is the (m,L) pair that maximizes M. The selection of an embedding therefore becomes a constrained optimization: maximize $M(m,L)$ subject to the constraints that m and L are positive integers, but this analysis does not, and cannot, identify M. A circular logic has resulted in which embedding criteria are assessed by an adjudicating criterion which is itself an embedding criterion. The reasoning outlined above leads to the following conclusion: the optimal embedding for a time series cannot be determined by criteria based solely on the time series itself. (In this context, we wish to acknowledge the importance of Rapoport's work [23] on the analysis of paradox.) Failure to recognize this point has resulted in an embedding criterion–adjudicating measure–embedding criterion circularity that has characterized much of the literature on this subject.

In order to break this cycle, we must bring to the analysis knowledge that cannot be provided by the time series itself. We can accomplish this by basing our investigation on the analysis of time series that were generated by dynamical sys-

*Author to whom correspondence should be addressed.

tems that have explicit analytical representations as *n*-dimensional differential or functional differential equations. Because the analytical representations are available, we can apply forms of analysis that cannot be applied to the time series itself. Specifically, we can use procedures for determining the largest Lyapunov exponent that requires equations for the vector field throughout the state space constructed by Benettin *et al.* [24,25]. These values provide a gold standard for subsequent comparisons. The first phase of the investigation proceeds in five steps.

(1) Three model systems whose governing equations can be expressed analytically are identified and time series are generated from each of them.

(2) The largest Lyapunov exponents of these systems are determined using the analytical expressions of the vector field.

(3) Five criteria for selecting embedding parameters are described and applied to the time series generated by the model system.

(4) Using these embedding parameters, the largest Lyapunov exponent of each time series is calculated for the five sets of embedding parameters using a procedure published by Gao and Zheng [12] that can be applied to time series data.

(5) The Lyapunov exponents computed from the time series are compared against those determined by the more exhaustive analytically based calculations. The criterion that most consistently reproduces the reference values of the Lyapunov exponents is deemed to be the most successful.

The second phase of the investigation examines the robustness of these conclusions when sensitivity to noise is considered. This component of the analysis includes both computationally generated and experimental data.

II. SPECIFICATION OF THE EXAMPLE SYSTEMS AND THEIR LARGEST LYAPUNOV EXPONENTS

Three example systems will be considered in this study. The first is the Rössler system [26]:

$$dx/dt = -(y+z),$$

$$dy/dt = x+\alpha y,$$

$$dz/dt = \beta + z(x-\gamma),$$

$$\alpha = 0.15, \quad \beta = 0.20, \quad \gamma = 10.00, \quad \delta t = .125.$$

A 10 000-element time series was computed after the trajectory converged onto the attractor using a sixth order Runge-Kutta-Hutta algorithm [27]. The second system is the Mackey-Glass equation [28]:

$$dx/dt = \frac{ax(t-\tau)}{1+x^c(t-\tau)} - bx,$$

$$a = 0.20, \quad b = 0.10, \quad c = 10.00, \quad \tau = 17.$$

The parameter τ is a time delay. Thus, this is an infinite dimensional functional differential equation. A 10 000-point trajectory on the attractor was computed with a time interval of $\delta t = 0.10$. The third system is identical to the second except that the time delay is set equal to $\tau = 150$.

The largest Lyapunov exponent of each of these systems was calculated by a procedure published by Benettin *et al.* [24,25] that exploits the availability of analytical expressions for the vector field in the behavior space. The analysis begins by considering a small *n*-dimensional sphere of initial conditions. Over time this sphere evolves into an ellipsoid. The Lyapunov exponents determine the rate of its growth. In the Benettin *et al.* computational procedure, the trajectories of points on the surface of the sphere are approximated by the action of the linearized equations of motion. The vectors are repeatedly reorthonormalized using the Gram-Schmidt procedure. The Gram-Schmidt reorthonormalization does not affect the direction of the first vector in this system, so it tends to seek out the direction in tangent space corresponding to the most rapid growth. This provides an estimate of the largest Lyapunov exponent. The values of the Lyapunov exponents were found to be 0.129 (Rössler), 0.0071 (Mackey-Glass $\tau = 17$), and 0.0023 (Mackey-Glass $\tau = 150$).

III. EMBEDDING CRITERIA

As previously stated, an inappropriate choice of embedding dimension can result in a failure to characterize the structure of the time series. If m is too small, the embedded manifold is folded onto itself, and elements of its structure will be lost to the analysis. However, a strategy of simply using a very large embedding dimension for all cases is even less successful. The data requirements for the analysis increase with the embedding dimension. If the value of m is too great, structure is dispersed through a high dimensional space, and the time series is indistinguishable from noise. Thus we conclude that the embedding dimension must be large enough but no larger.

Several methods have been developed to estimate the minimum acceptable embedding dimension [7,17,20,29]. In this paper we compare methods based on the concept of minimizing the number of false nearest neighbors. Let X_i be an embedded point in $\Re^m$, and let X_j be the point closest to it. Consider the map of X_i and X_j from $\Re^m$ to $\Re^{m+1}$. If the $(m+1)$-dimensional points are no longer nearest neighbors, then X_i and X_j in $\Re^m$ are false nearest neighbors. False nearest neighbors can result when the embedded manifold is folded onto itself in $\Re^m$. When the embedding dimension is increased, an unfolding of the embedded set can separate X_i and X_j. The argument of false nearest neighbors concludes that the minimum acceptable embedding dimension can be established by determining a measure of the frequency of false nearest neighbors as a function of embedding dimension. The optimal embedding dimension m_{opt} is the smallest dimension that results in a stable minimum of this measure.

Thus, the underlying assumption of the methods compared in this paper holds that, when $m < m_{\text{opt}}$ and m is increased from m to $m+1$, the metric that is used to reflect the frequency of false nearest neighbors will decrease. For $m \geqslant m_{\text{opt}}$, further increases in the embedding dimension will not result in a significant decrease in this metric. All of the

criteria compared in this paper are constructed on this argument. They differ, however, in the metric that is used to characterize the frequency of false nearest neighbors. The choice of this metric is by no means trivial. X_i and X_j in $\Re^m$ can be false nearest neighbors under this definition even though the data were appropriately embedded. This can happen because they were positioned on opposite sides of a separatrix or, more commonly, as the result of noise in observed data. A simple exhaustive calculation of the frequency of false nearest neighbors is not necessarily the most successful. Measures that, for example, incorporate a time history of local trajectories centered on X_i and X_j might prove to be more robust against noise. This is one of the questions examined in this investigation.

A. Method of Gao and Zheng

Gao and Zheng [11,12] use the following argument to construct a measure that reflects the incidence of false nearest neighbors. Consider two vectors X_i and X_j. If they are genuine nearest neighbors, and if the flow is uniform in this region of the state space, then X_{i+k} and X_{j+k} will also be close to each other for small k. The statistical nature of this argument is apparent when it is recognized that domains of the state space where flow separates provide exceptions to this generalization. Additionally, for bounded chaotic systems, this will cease to be true if k is large. If X_i and X_j are false nearest neighbors, they are, by definition, close to each other only because the embedded set has been folded onto itself in a neighborhood containing these points. Therefore, the flow controlling the evolution of X_i in state space is not necessarily similar to the flow controlling the evolution of X_j. Compared to genuine nearest neighbors, there is a higher probability that the trajectories corresponding to X_i and X_j will separate.

The method of Gao and Zheng is based on the following argument. Let $|X_i,X_j|$ denote the Euclidean distance between points X_i and X_j. Typically, $|X_{i+k},X_{j+k}|/|X_i,X_j|$ will be greater if X_i and X_j are false nearest neighbors. A successful embedding is one that will, on average, reduce this ratio. Therefore, they construct the following measure:

$$\Lambda(k,m,L)=\frac{1}{N_{\mathrm{ref}}}\sum_{i,j}\ \ln\left\{\frac{|X_{i+k},X_{j+k}|}{|X_i,X_j|}\right\}.$$

From this equation it is seen that four parameters must be specified, N_{ref}, k, m, and L. In our implementation of the Gao-Zheng criterion, the average is taken from N_{ref} points X_i, randomly selected from points in the embedding space. In the calculations of Fig. 1, 10 000 data points are used and $N_{\mathrm{ref}}=500$. After X_i has been chosen, an X_j is found that satisfies two criteria. First, we require $|X_i,X_j|\leqslant r$, that is, the average is taken over points that are initially close to each other. For example, in the calculations shown in Fig. 1, r is 10% of the standard deviation of the time series. Numerical experiments indicated that the results are robust against variations in r. This condition alone is insufficiently restrictive. If this were the sole criterion used to select X_j, Λ could emphasize those points that are close to X_i because the corresponding data points in X_j were sampled at approximately the same time. If an oversampled signal is being examined, this can lead to a spurious indication of structure in the state space. In order to control against this possibility, we impose a second condition on X_j, namely, a minimum elapsed time between sampled data points X_i and X_j. This is done by requiring $|i-j|$ to be greater than some minimum temporal spacing, denoted $k_{\mathrm{separation}}$, which can be expressed in terms of the autocorrelation time. This is an application of a procedure originally introduced by Theiler [44] in the specific context of calculating the correlation dimension. In the calculations shown in Fig. 1, we required $|i-j|\geqslant 25$. This is equal to the first minimum of the autocorrelation function. After X_i is selected at random, X_j is determined. X_j is specified by the value of j closest to i that satisfies $|i-j|\geqslant 25$ and $|X_i,X_j|\leqslant r$. If no value of j satisfying these criteria exists, X_i is discarded and another random selection is made.

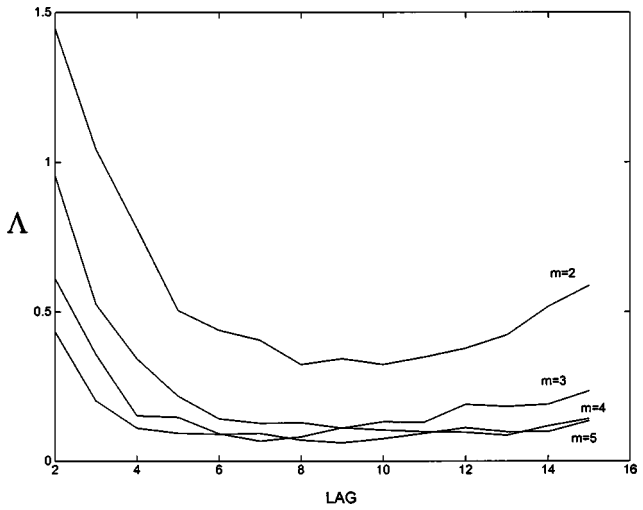

FIG. 1. Λ as a function of lag L for the Rössler attractor. The original time series contains 10 000 points at sampling interval $\delta t=0.125$. Parameter r is 10% of the standard deviation of the original data and $k=9$. Λ is calculated for $m=2,3,\ldots,5$ and $L=2,3,\ldots,15$. $N_{\mathrm{ref}}=500$. The minimum sampling separation $|i-j|\geqslant 25$.

Another parameter to be specified is the evolution time k. If k is too small, the noise in the time series could obscure the separation of trajectories corresponding to false nearest neighbors. If k is too large, the exponential separation of trajectories in chaotic systems will end and the distinction between false and genuine nearest neighbors will diminish. It is therefore necessary to fix k in terms of a natural time scale of the time series. In our calculations we set k equal to the autocorrelation time (the time required for the autocorrelation function to drop to $1/e$ of its initial value). The dependence of the method on the choice of evolution time is considered again in the presentation of the method of characteristic length.

The calculation of $\Lambda(k,m,L)$ can be reduced to the following sequential process.

(1) For a specified m,L pair, the mean distance between points in the embedding space and the standard deviation of that mean are determined. This can be done by an exhaustive

calculation of all i,j pairs or by a random sample that is large enough to achieve a stable value. The local neighborhood radius r is specified in terms of the standard deviation of the time series, for example, 10%.

(2) N_{ref} is specified. This is the number of reference points X_i that will be randomly sampled from the embedding space.

(3) $k_{\mathrm{separation}}$, the minimum temporal separation of reference point X_i and its neighbor X_j, must be specified. As discussed in the preceding text, the first minimum of the autocorrelation function of the original time series can be used.

(4) The value of k, the evolution time, must be determined. We have used the autocorrelation time (the time required for the autocorrelation function to drop to $1/e$ of its original value).

(5) The following computation is performed for each of the N_{ref} reference points X_i randomly sampled from the embedding space. A point X_j is found that satisfies the two criteria $|X_i,X_j|\leqslant r$ and $|i-j|\geqslant k_{\mathrm{separation}}$. If no point X_j satisfying these conditions can be found, then X_i is discarded and replaced with another randomly selected reference point. Using a successful X_i,X_j pair, the value of $\ln\{|X_{i+k},X_{j+k}|/|X_i,X_j|\}$ is computed.

(6) The average value of $\ln\{|X_{i+k},X_{j+k}|/|X_i,X_j|\}$ is determined. This is the value of $\Lambda(k,m,L)$.

We used the Rössler equations to generate the results presented in Fig. 1. The original time series contained 10 000 points, and N_{ref} was set equal to 500. The local neighborhood radius r is 10% of the standard deviation of the time series. The evolution time k is 9, which is the corresponding autocorrelation time. $k_{\mathrm{separation}}$ is 25, which is the first minimum of the autocorrelation function. The initial embedding dimension m is fixed at 2 and Λ is calculated as a function of the lag L. This process is repeated for increasing values of m. As shown in this figure, the value of Λ decreases significantly as m is increased from 2 to 3. However, successive increases in m do not result in further significant decreases in Λ. Therefore it is concluded that $m=3$ is an appropriate embedding dimension. The best value of L corresponds to the L at the first minimum value of Λ in the $m=3$ case. This results in fixing $L=8$. This result is consistent with those published by Gao and Zheng [11]. The results obtained when this criterion was applied to the other time series in the test collection are reported in Sec. III.

B. Method of Schuster

The procedure for estimating an optimal embedding dimension presented by Schuster and his colleagues [29] examines the relationship between sets of nearest neighbors in successive embeddings. Let $X_i^{(m)}$ be an embedded point in $\Re^m$, where it should be recalled that the construction of $X_i^{(m)}$ includes the specification of the lag L. In this procedure, the N_n nearest neighbors of $X_i^{(m)}$ are identified. They are denoted by $X_{i,1}^{(m)},X_{i,2}^{(m)},\ldots,X_{i,N_n}^{(m)}$. They are ordered in the sense that $X_{i,1}^{(m)}$ is the closest neighbor of $X_i^{(m)}$, $X_{i,2}^{(m)}$ is the next closest, and so on. In their implementation, Liebert *et al.* set $N_n=10$ for an example problem containing 10 000 data points.

Liebert *et al.* consider the impact of increasing m to $m+1$ on the nearest neighbor set. Let $X_i^{(m+1)}$ denote the element in $\Re^{m+1}$ corresponding to $X_i^{(m)}$ in $\Re^m$. Let $X_{i,k}^{(m+1)}$ denote the kth nearest neighbor of $X_i^{(m+1)}$ in $\Re^{m+1}$, where again the nearest neighbors are ordered with $X_{i,1}^{(m+1)}$ being the closest to $X_i^{(m+1)}$. It should be stressed that points $X_{i,k}^{(m+1)}$ are defined by their proximity to $X_i^{(m+1)}$ in $\Re^{m+1}$. They are not necessarily the projections of $X_{i,k}^{(m)}$ to $\Re^{m+1}$. (We use the term projection to denote a relationship defined by embedding processes in two consecutive dimensions.)

If an embedding were ideal, then the transition from $\Re^m$ to $\Re^{m+1}$ would preserve nearest neighbor relationships, and $X_{i,k}^{(m+1)}$ would be the $(m+1)$-dimensional point corresponding to $X_{i,k}^{(m)}$ in $\Re^m$. The Liebert *et al.* metric provides a means of quantifying the degree to which this relationship fails to be true. Let $Z_{i,1}^{(m+1)}$ be the point in $\Re^{m+1}$ corresponding to $X_{i,1}^{(m)}$, that is, the projection of $X_{i,1}^{m}$ to $\Re^{m+1}$. $Z_{i,k}^{(m+1)}$ is defined analogously for $k=2,\ldots,N_n$. The relationships between these points is depicted below; $\uparrow$ denotes the projection from $\Re^m$ to $\Re^{m+1}$:

$$\begin{array}{ccccccccccc} X_{i,N_n}^{(m+1)} & \cdots & X_{i,2}^{(m+1)} & X_{i,1}^{(m+1)} & X_i^{(m+1)} & Z_{i,1}^{(m+1)} & Z_{i,2}^{(m+1)} & \cdots & Z_{i,N_n}^{(m+1)} & \Re^{m+1} \\ & & & & \uparrow & \uparrow & \uparrow & & \uparrow & \\ & & & & X_i^{(m)} & X_{i,1}^{(m)} & X_{i,2}^{(m)} & \cdots & X_{i,N_n}^{(m)} & \Re^m. \end{array}$$

In the case of an ideal embedding, $Z_{i,1}^{(m+1)}=X_{i,1}^{(m+1)}$ and the ratio

$$\frac{|X_i^{(m+1)}-Z_{i,1}^{(m+1)}|}{|X_i^{(m+1)}-X_{i,1}^{(m+1)}|}$$

is equal to 1. If $Z_{i,1}^{(m+1)}\neq X_{i,1}^{(m+1)}$, then this ratio is greater than 1. The product

$$\prod_{k=1}^{N_n}\left\{\frac{|X_i^{(m+1)}-Z_{i,k}^{(m+1)}|}{|X_i^{(m+1)}-X_{i,k}^{(m+1)}|}\right\}$$

is an empirical measure of the degree of correspondence between the sets $\{X_{i,k}^{(m+1)}\}$ and $\{Z_{i,k}^{(m+1)}\}$. A large value of this product will indicate a distortion of nearest neighbor relationships that results from an insufficient value of m.

The Liebert *et al.* analysis also considers the relationship between the nearest neighbor set of $X_i^{(m+1)}$ in $\Re^{m+1}$ and the corresponding set of points in $\Re^m$. As previously defined, $X_{i,k}^{(m+1)}$ is the kth nearest neighbor of $X_i^{(m+1)}$ in $\Re^{m+1}$. Let $V_{i,k}^m$ denote the corresponding point in $\Re^m$. In analogy with the previous diagram, the relationship between these sets is given below. In this case, $\downarrow$ indicates the projection from $\Re^{m+1}$ to $\Re^m$:

$$\begin{array}{ccccccccccc} X_{i,N_n}^{(m+1)} & \cdots & X_{i,2}^{(m+1)} & X_{i,1}^{(m+1)} & X_i^{(m+1)} & & & & & & \Re^{m+1} \\ \downarrow & & \downarrow & \downarrow & \uparrow & & & & & & \\ V_{i,N_n}^{(m)} & \cdots & V_{i,2}^{(m)} & V_{i,1}^{(m)} & X_i^{(m)} & X_{i,1}^{(m)} & X_{i,2}^{(m)} & \cdots & X_{i,N_n}^{(m)} & & \Re^m. \end{array}$$

The corresponding product is

$$\prod_{k=1}^{N_n} \left\{ \frac{|X_i^{(m)} - X_{i,k}^{(m)}|}{|X_i^{(m)} - V_{i,k}^{(m)}|} \right\}.$$

For the point $X_i^{(m)}$, Liebert *et al.* define $W_i(m,L)$ as

$$W_i(m,L) = \prod_{k=1}^{N_n} \left\{ \left(\frac{|X_i^{(m+1)} - Z_{i,k}^{(m+1)}|}{|X_i^{(m+1)} - X_{i,k}^{(m+1)}|} \right) \left(\frac{|X_i^{(m)} - X_{i,k}^{(m)}|}{|X_i^{(m)} - V_{i,k}^{(m)}|} \right) \right\}.$$

$W_i(m,L)$ is averaged over a set of N_{ref} points selected randomly in the $\Re^m$ embedding space. Liebert *et al.* sample 10% of the embedded points. $W(m,L)$ is defined as

$$W(m,L) = \ln \langle W_i(m,L) \rangle,$$

where

$$\langle W_i(m,L) \rangle = \frac{1}{N_{\text{ref}}} \sum_{i=1}^{N_{\text{ref}}} W_i(m,L).$$

As in the case of the Gao-Zheng criterion, m is fixed and $W(m,L)$ is calculated as a function of L for progressively increasing values of m.

For specified values of m and L, $W(m,L)$ is calculated by the following procedure.

(1) N_{ref}, the number of references points to be used, must be specified. Liebert *et al.* [29] use 10% of the total.

(2) N_n, the number of nearest neighbors computed for each reference point, must be specified. Liebert *et al.* [29] use $N_n = 10$.

(3) A reference point $X_i^{(m)}$ is randomly selected from the embedded set in $\Re^m$. For each $X_i^{(m)}$, the following calculations are performed. (a) The N_n nearest neighbors of $X_i^{(m)}$ are determined. They are denoted by $X_{i,1}^{(m)}, X_{i,2}^{(m)}, \ldots, X_{i,N_n}^{(m)}$. (b) The projections of these nearest neighbors into $\Re^{m+1}$ are determined. They are denoted by $Z_{i,1}^{(m+1)}$, $Z_{i,2}^{(m+1)}, \ldots, Z_{i,N_n}^{(m+1)}$. (c) $X_i^{(m+1)}$ is the projection of $X_i^{(m)}$ into $\Re^{m+1}$. The N_n nearest neighbors of $X_i^{(m+1)}$ are determined. They are denoted by $X_{i,1}^{(m+1)}, X_{i,2}^{(m+1)}, \ldots, X_{i,N_n}^{(m+1)}$. (d) The projections of $X_{i,j}^{(m+1)}$ to $\Re^m$ are determined. They are denoted by $V_{i,1}^{(m)}, V_{i,2}^{(m)}, \ldots, V_{i,N_n}^{(m)}$. (e) The product $W_i(m,L)$ is calculated:

$$W_i(m,L) = \prod_{k=1}^{N_n} \left\{ \left(\frac{|X_i^{(m+1)} - Z_{i,k}^{(m+1)}|}{|X_i^{(m+1)} - X_{i,k}^{(m+1)}|} \right) \left(\frac{|X_i^{(m)} - X_{i,k}^{(m)}|}{|X_i^{(m)} - V_{i,k}^{(m)}|} \right) \right\}.$$

(4) $W(m,L)$ is the logarithm of the average value of $W_i(m,L)$:

$$W(m,L) = \ln \left\{ \frac{1}{N_{\text{ref}}} \sum_{i=1}^{N_{\text{ref}}} W_i(m,L) \right\}.$$

Figure 2 shows plots of $W(m,L)$ versus L using data from the previously defined implementation of the Rössler equations. The best choice of embedding corresponds to the smallest value of m that produces the limiting behavior of $W(m,L)$. In this example, this is seen to correspond to $m = 3$. The best choice of L corresponds to the lag at the first minimum value of $W(m,L)$ in the $m = 3$ case. This results in $L = 8$. As an additional test, a time series was generated using

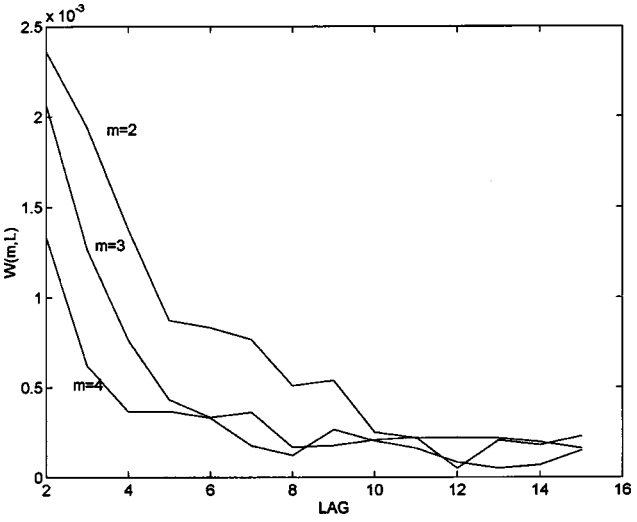

FIG. 2. $W(m,L)$ versus lag for the Rössler data set. In these calculations 10 000 points were used. W is calculated for $m = 2$, 3, and 4; $L = 2,3, \ldots, 15$. Number of reference points $N_{\text{ref}} = 300$. Number of nearest neighbors $N_n = 25$.

the Lorenz equations $dx/dt=\alpha(y-x)$, $dy/dt=x(R-z)$, $dz/dt=xy-bz$, $\alpha=16.0$, $R=45.92$, $b=4$, and $\delta t=0.125$. The Liebert *et al.* procedure was applied to this time series and produced embedding parameters in agreement with those found using a procedure published by Wolf *et al.* [30].

The most computationally demanding element of this procedure is the identification of the N_n nearest neighbors of each X_i. (Similarly, the search for the single nearest neighbor $N_n=1$ which is implemented in the method of global false nearest neighbors, is the most computationally expensive element of that method.) There is a large literature describing procedures that can be modified to produce methods that will accelerate nearest neighbor searches in $\Re^m$ [spanning trees [31], *KD* trees [32], *K* trees [33–36] (structures for optimizing orthogonal range searches)]. In our recent calculations, we used our implementation of Schreiber's linked-list search procedure [37].

C. Method of characteristic length

As previously described, the Gao-Zheng method is based on the rate of separation of points that are initially close to each other. It is therefore closely related to the estimation of the largest Lyapunov exponent. This relationship is developed explicitly in the next section. There are operational difficulties associated with the Gao-Zheng method. They turn on the choice of the evolution time parameter *k*, which specifies the time over which the divergence of trajectories is observed. The evolution time before two nearby points become uncorrelated is a function of both the largest Lyapunov exponent and the initial separation of these points. However, without some knowledge of the spatial extent of the system's attractor, it is difficult to estimate when the evolution time is too large. The method of characteristic length addresses this point by estimating the size of the attractor and using this length in an assessment of the separation time of trajectories that are close initially. For a given scalar time series, the characteristic length $J(m,L)$ is a function of m and L and is defined as

$$J(m,L)=\langle |X_i,X_j| \rangle,$$

where $\langle \cdots \rangle$ denotes the average Euclidean distance taken over randomly selected pairs of points in the embedding space. $J(m,L)$ provides an imperfect measure of the size of the attractor. In our calculations, the number of pairs of points used to calculate $J(m,L)$ was 15% of the number of embedded points. It should be noted that in the case of $J(m,L)$ calculations, the choice of i and j is random and is not subject to the restrictions on i,j pairs employed in the calculation of $\Lambda(k,m,L)$.

The argument for indirectly assessing the frequency of false nearest neighbors with the method of characteristic length follows a development analogous to that used to construct the Gao-Zheng criterion. Suppose that X_i and X_j, points that are initially close in phase space, are true nearest neighbors. The time required for them to separate to some fraction of $J(m,L)$ will depend on the Lyapunov exponent. We denote this separation time as T_J. If, in contrast, X_i and X_j are false nearest neighbors, they are close to each other because the embedded set is folded onto itself in a neighborhood containing these points. Under these circumstances, the time evolution of X_i and X_j could display very different dynamical behavior. This would typically result in a faster separation of their trajectories.

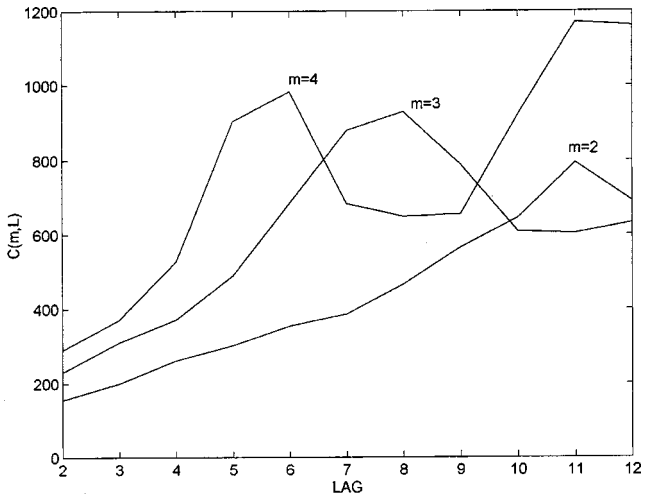

FIG. 3. $C(m,L)$ versus lag for the Rössler data set. In these calculations 10 000 points were used. $r=10\%$ of the standard deviation of the data set. C is calculated for $m=2$, 3, and 4; $L=2,3,\ldots,12$. $N=500$ and $|i-j|\geqslant 25$.

On average, therefore, we expect the separation time T_J for false nearest neighbors to be shorter than the average separation time for true nearest neighbors. An average separation time is calculated for $m=2$ as a function of L. As m is increased the frequency of false nearest neighbors is reduced and the average separation time increases. The embedding dimension m is increased until a further increase in m does not have an impact on the average separation time.

The procedure can be operationalized by the following sequence of calculations. For a given m,L pair, $C(m,L)$ is calculated in the following steps.

(1) The characteristic length $J(m,L)$ is calculated by the average $J(m,L)=\langle |X_i,X_j| \rangle$, where i,j are selected randomly. The number of pairs used to form the average is equal to 15% of the number of points in the embedding space.

(2) N_{ref}, the number of reference points used in the separation time calculations, is specified. In the calculations shown in Fig. 3, where 10,000 points are in the time series, N_{ref} is set equal to 500.

(3) A value of r is specified. The specification used in our implementation of the Gao-Zheng method is also used in the Fig. 3 calculations. Specifically, r is set equal to 10% of the standard deviation of the original time series.

(4) The embedded point X_i is chosen at random. X_j is defined as the value of j closest to i that satisfies the conditions that $|i-j|$ is greater than the signal's autocorrelation time and $|X_i,X_j|\leqslant r$. If no value of j satisfying these two conditions exists, X_i is discarded and another point is selected.

(5) $T_J(X_i,X_j)$ is determined. This is the minimum integer k required for $|X_{i+k},X_{j+k}|$ to exceed $0.4J(m,L)$. If these points do not separate to this distance, X_i is discarded and another point is chosen.

(6) This process is repeated until N_{ref} values of $T_J(X_i,X_j)$ have been obtained. $C(m,L)$ is their average:

$$C(m,L)=\frac{1}{N_{\text{ref}}}\sum_{i,j} T_J(X_i,X_j).$$

As shown in Fig. 3, m is first set equal to 2 and $C(m,L)$ is calculated as a function of L. The embedding dimension is then increased and $C(m,L)$ is again calculated. The increase in $C(m,L)$ that was anticipated by the preceding argument is observed. Further increases in m do not, however, result in further increases $C(m,L)$; therefore it is concluded that $m=3$ is an effective choice. The indicated value of lag corresponds to the first maximum of $C(m,L)$ when $m=3$. This results in $L=8$. The procedure was also applied to the Lorenz time series, and again results consistent with those of Wolf *et al.* [30] were obtained.

D. Global false nearest neighbors and the autocorrelation function

The three methods presented thus far determine the embedding dimension and lag simultaneously. In this section we combine a method for choosing a proper embedding dimension, the method of global false nearest neighbors, with a separate method for determining the lag based on the autocorrelation function. This criterion for specifying lag sets it equal to the value of delay corresponding to the first zero of the autocorrelation function. The autocorrelation function $C(k)$ for a time series x_i, $i=1,2,\ldots,N$ is given by

$$C(K)=\frac{\sum_{i=1}^{N-k}(x_{i+k}-\overline{x})(x_i-\overline{x})}{\sum_{i=1}^{N-k}(x_i-\overline{x})^2} \quad \text{where } \overline{x}=\frac{1}{N}\sum_{i=1}^{N} x_i.$$

The determination of the embedding dimension using a global false nearest neighbors argument begins with an embedding in $\Re^m$ which uses the lag established using the autocorrelation function. Let X_i denote an element in this embedding, and let $X_i^{\text{NN}}=(x_i^{\text{NN}},x_{i+L}^{\text{NN}},\ldots,x_{i+(m-1)L}^{\text{NN}})$ denote its nearest neighbor. The Euclidean distance between these two points in $\Re^m$ is denoted by $|X_i-X_i^{\text{NN}}|_m$:

$$|X_i-X_i^{\text{NN}}|_m^2=\sum_{k=0}^{m-1}(x_{i+kL}-x_{i+kL}^{\text{NN}})^2.$$

The Euclidean distance between the projection of these two points into $\Re^{m+1}$ is given by

$$|X_i-X_i^{\text{NN}}|_{m+1}^2=|X_i-X_i^{\text{NN}}|_m^2+(x_{i+mL}-x_{i+ml}^{\text{NN}})^2.$$

Abarbanel [38] defines R, a measure of the distance between X_i and X_i^{NN} in $\Re^{m+1}$ normalized against their distance in $\Re^m$, as

$$R=\left\{\frac{|X_i-X_i^{\text{NN}}|_{m+1}^2-|X_i-X_i^{\text{NN}}|_m^2}{|X_i-X_i^{\text{NN}}|_m^2}\right\}^{1/2},$$

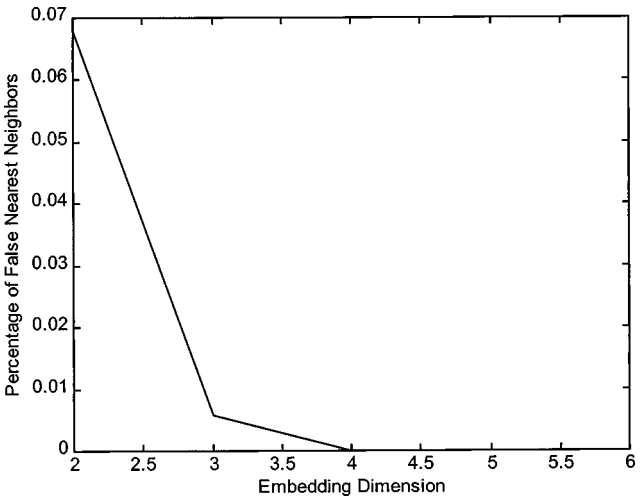

FIG. 4. Percentage of false nearest neighbors versus embedding dimension for the Rössler data set. In these calculations 10 000 points were used. $m=2,3,\ldots,6$; $L=9$. The threshold is equal to 15.

$$R=\frac{|x_{i+mL}-x_{i+mL}^{\text{NN}}|}{|X_i-X_i^{\text{NN}}|}.$$

X_i^{NN} is deemed to be a false nearest neighbor of X_i in $\Re^m$ if R exceeds the constant R_{tol}. The choice of R_{tol} was discussed by Abarbanel [38]. We follow his recommendation here and set $R_{\text{tol}}=15$. The use of global false nearest neighbors to determine the embedding dimension is implemented by the following procedure.

(1) L is set equal to the first zero of the autocorrelation.

(2) R_{tol} is set equal to a fixed value.

(3) The percentage of false nearest neighbors is calculated as a function of m using the following procedure. (a) For every point $X_i\in\Re^m$, the nearest neighbor X_i^{NN} is determined. (b) The corresponding value of R is calculated. (c) If $R>R_{\text{tol}}$, then X_i^{NN} is deemed to be a false nearest neighbor of X_i.

(4) The value of m is increased until false nearest neighbors are no longer observed or until the frequency of false nearest neighbors is below an acceptable value.

Figure 4 shows the results obtained with the Rössler data. The value of the lag determined from the autocorrelation function was 9. Using this value of the lag, the procedure identified $m=4$ as the optimal embedding dimension.

E. Global false nearest neighbors and mutual information

This procedure differs from the immediately preceding method in the criterion used to determine the lag. The same procedure, global false nearest neighbors, is used to determine the embedding dimension. Choosing the lag L to be the first zero crossing of $C(k)$ means that, on average, the observations x_i and x_{i+L} will be linearly independent. This is the optimal linear choice, from the point of view of predictability in a least squares sense of x_{i+L} from a knowledge of x_i. Although historically it has been widely used to determine the time delay, some authors now question its use when the underlying process is nonlinear [38]. Abarbanel [38] and

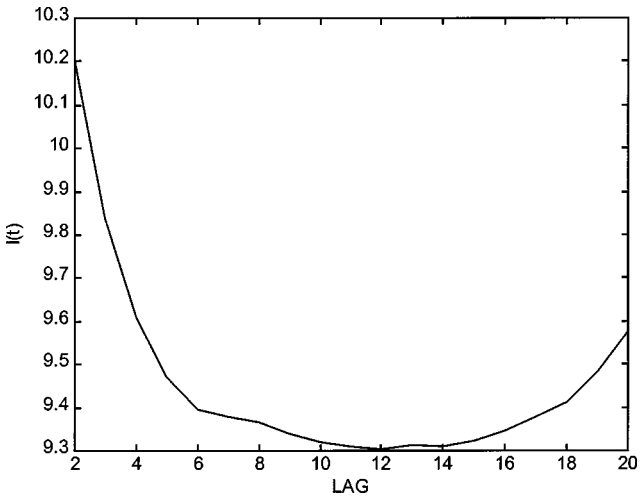

FIG. 5. Mutual information versus lag for the Rössler data set. In these calculations 10 000 points were used.

others (notably Fraser [10]) have therefore argued that the first minimum of the average mutual information function is a more appropriate choice of the lag, because mutual information can be regarded as a nonlinear analog of the autocorrelation function. The general case of the definition of mutual information begins with two sets $A=\{a_i\}$ and $B=\{b_j\}$. The mutual information is the amount learned by the measurement of a_i about the value of b_j. In bits, it is given by

$$\log_2\left[\frac{P_{AB}(a_i,b_j)}{P_A(a_i)P_B(b_j)}\right],$$

where P_{AB} is the joint probability distribution, and P_A and P_B are the individual probability distributions. We note that if a measurement of a_i is completely independent of b_j, then the amount of information gained about b_j by measuring a_i, which is the mutual information, is zero. The average mutual information is defined as the average over all measurements of this statistic between sets A and B [38]:

$$I_{AB}=\sum_{a_i,b_j} P_{AB}(a_i,b_j)\log_2\left[\frac{P_{AB}(a_i,b_j)}{P_A(a_i)P_B(b_j)}\right].$$

The specific application to a time series follows immediately from this definition. As before, let x_i, $i=1,2,\ldots,N$, denote an observed time series. Define the set $A=\{a_i\}$ as the value of x at time i, x_i, and the set B as the value of x at time $i+\tau$, $x_{i+\tau}$. The mutual information becomes a function of the time shift variable τ,

$$I(\tau)=\sum_{x_i,x_{i+\tau}} P(x_i,x_{i+\tau})\log_2\left[\frac{P(x_i,x_{i+\tau})}{P(x_i)P(x_{i+\tau})}\right].$$

This measure tells us the average amount of information learned about $x_{i+\tau}$ by measuring x_i. Figure 5 shows the results using the Rössler equations. We conclude that $L=12$ is the indicated choice.

IV. CALCULATING THE LARGEST LYAPUNOV EXPONENT FROM A TIME SERIES

As outlined in the Introduction, these five methods for determining embedding parameters were applied to the three test cases. The results are displayed in Table I. In that table, GFNN-A identifies the embedding parameters determined by the autocorrelation function combined with the method of global false nearest neighbors and GFNN-MI identifies the results obtained when the lag was determined by calculating the mutual information.

The comparative success of these embedding parameters was assessed by using them in calculations of the largest Lyapunov exponent. For the purposes of this test, the embedding criterion that produces an embedding which in turn produces a value for the largest Lyapunov exponent that is clos-

TABLE I. Embedding parameters and Lyapunov exponents calculated by different methods.

Method	Rössler	Mackey-Glass $\tau=17$	Mackey-Glass $\tau=150$
	Embedding parameters		
	m,*L*	*m*,*L*	*m*,*L*
Gao-Zheng	3,8	3,14	6,26
Schuster	3,9	3,10	3,32
Characteristic length	3,8	4,10	5,17
GFNN-A	4,9	4,18	6,82
GFNN-MI	4,12	4,23	6,82
	Lyapunov exponents		
Benettin	0.129	0.0071	0.0023
Gao-Zheng	0.128	0.0106	0.0014
Schuster	0.135	0.0092	0.0011
Characteristic length	0.128	0.0073	0.0015
GFNN-A	0.124	0.0089	0.0020
GFNN-MI	0.125	0.0085	0.0020

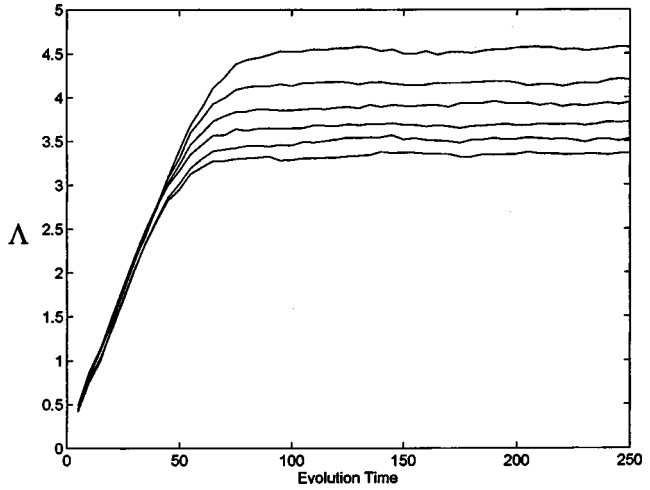

FIG. 6. Λ versus evolution time k for the Rössler data set. In these calculations 10 000 points were embedded using the embedding parameters $m=3$ and $L=8$. Neighborhood size $r=1\%,2\%,\ldots,6\%$ of the time series' standard deviation. $N_{\mathrm{ref}}=500$. The top line corresponds to $r=1\%$, and the bottom corresponds to $r=6\%$. $|i-j|\geqslant 40$.

est to the Benettin *et al.* reference value is deemed to be the most successful. Of the many candidate methods for calculating Lyapunov exponents from a time series, we chose the procedure published by Gao and Zheng [12], which is closely related to their procedure for identifying appropriate embedding parameters. The largest Lyapunov exponent λ is a quantitative characterization of the rate at which two initially close points diverge in phase space under the assumption that this separation is exponential,

$$|X_{i+k},X_{j+k}|=|X_i,X_j|e^{\lambda\,\delta t},$$

where δt is the sampling interval. As in the case of estimating embedding parameters with the Gao-Zheng method, the choice of X_i,X_j pairs cannot be arbitrary. First, the points must be close initially. Therefore, as before, we require $|X_i,X_j|\leqslant r$ where r is expressed in terms of the standard deviation of the original time series. Second, the points must have a minimum initial temporal separation; that is, we require $|i-j|\geqslant k_{\mathrm{separation}}$ where $k_{\mathrm{separation}}$ is expressed in terms of the autocorrelation function. If these conditions are met, and if the separation of X_i and X_j is exponential, then the average value of $\ln\{|X_{i+k},X_{j+k}|/|X_i,X_j|\}$ when plotted as a function of time will be linear and have the slope λ. An example using the Rössler time series is shown in Fig. 6. The function

$$\frac{1}{N_{\mathrm{ref}}}\sum_{i,j}\ \ln\left\{\frac{|X_{i+k},X_{j+k}|}{|X_i,X_j|}\right\}$$

is plotted as a function of time for six values of r (1%, 2%, . . . ,6% of the standard deviation of the time series). This function exhibits a linear region with a slope that is independent of r, followed by a region where the slope tends to zero. The slope is approximately 0.07, which is in agreement with previously published estimates [30]. The results obtained when this procedure for estimating λ was applied to the test systems are given in Table I.

Table I shows the embedding parameters and Lyapunov exponents generated by each method. Calculations using the Rössler time series produced similar embedding parameters, and in all cases the Lyapunov exponents were close to the Benettin reference value. In the trials using the Mackey-Glass system with $\tau=17$, some differences in embedding parameters and performance were observed. The characteristic length, GFNN-A, and GFNN-MI methods give a somewhat better performance. It is only in the group of calculations that examine the Mackey-Glass system with $\tau=150$ that we begin to see a notable difference in performance. In this case, only the GFNN-A and GFNN-MI methods resulted in an estimated exponent that was close to the reference value. While one might argue that the characteristic length was better for the Rössler system and the $\tau=17$ Mackey-Glass system, only the two global false nearest neighbor methods performed reasonably well in all three trials.

V. EXPERIMENTAL DATA AND SENSITIVITY TO NOISE

A long and melancholy history demonstrates that procedures that are successful in the examination of computationally generated noise-free data can fail when applied to noisy time series. This concern motivated the next phase of the investigation in which the robustness of the embedding criteria to noise is investigated.

The three model systems used in the earlier investigation (Rössler, Mackey-Glass $\tau=17$, and Mackey-Glass $\tau=150$) were used. Two experimental time series were also added to the test collection. The first is an electroencephalographic time series recorded during a clinically induced generalized seizure. Details of the recording protocol are given by Cellucci *et al.* [39]. The second experimental time series is a resting, eyes-closed electroencephalogram (EEG) recorded from a healthy control subject. Watanabe *et al.* [40] described the recording procedure. The incorporation of experimental data into the study raises a procedural dilemma. In the case of the computational systems, the Benettin *et al.* [24,25] procedure could be used to obtain high quality reference values for the Lyapunov exponents. In the case of the experimental data, this is not an option. We must therefore identify an alternative procedure for assessing an embedding criterion's robustness to noise. We operationally define a criterion as robust if the computational addition of noise to the original time series has a minimal impact on the cumulative distribution of interpoint distances in the embedding space. This is implemented in the following five-step procedure.

(1) Let S denote the original time series. The embedding criterion is applied to S to produce embedding parameters m and L.

(2) The time series S is embedded using these parameters and the cumulative distribution of interpoint distances in the embedding space is calculated as a function of scale variable r. If there are N data points in S, then there are $K=N-(m-1)L$ points in the embedding space. Let N_P denote the

TABLE II. Kolmogorov-Smirnov P_{null}.

	Gao-Zheng	Schuster	Characteristic length	GFNN-A	GFNN-MI
			Rössler		
10 dB	0.914	0.999	0.999	0.999	0.999
5 dB	0.513	0.989	0.999	0.999	0.999
0 dB	0.002	0.014	0.179	0.084	0.152
			Mackey-Glass, $\tau=17$		
10 dB	0.999	0.295	0.927	0.999	0.999
5 dB	0.999	0.999	0.999	0.999	0.999
0 dB	0.124	0.401	no result	0.013	0.013
			Mackey-Glass, $\tau=150$		
10 dB	0.999	0.362	0.999	0.999	0.999
5 dB	0.999	0.999	0.999	0.999	0.942
0 dB	no result	0.999	no result	0.213	0.055
			EEG seizure		
10 dB	no result	0.999	0.999	0.999	0.999
5 dB	no result	0.999	no result	0.845	0.999
0 dB	no result	0.484	no result	0.065	0.972
			EEG rest		
10 dB	0.999	0.999	0.999	0.999	0.999
5 dB	0.999	0.557	0.999	0.998	0.999
0 dB	0.596	0.999	0.999	0.186	0.999

number of distinct pairs of points. The cumulative distribution $C_S(r)$ is given by

$$C_S(r)=\frac{1}{N_P}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\Theta(r-|X_i-X_j|)$$

where Θ is the Heaviside function.

(3) Gaussian distributed noise is added to the time series S. The amplitude of noise is determined by a previously specified signal to noise ratio. The resulting time series is denoted S^*. The same embedding criterion is applied to S^* to produce embedding parameters m^* and L^*.

(4) Using m^* and L^*, the cumulative distribution of S^*, $C_{S*}(r)$, is computed.

(5) The two cumulative distributions are compared using the Kolmogorov-Smirnov statistic [41,42]. The Kolmogorov-Smirnov D is the maximum value of the absolute difference between two cumulative distributions:

$$D=\max_{-\infty<x<\infty}|C_S(r)-C_{S*}(r)|.$$

The null hypothesis holds that the two data sets are drawn from the same parent distribution. The probability of the null hypothesis is given by

$$P_{\text{null}}=Q_{\text{KS}}\left\{\left[\sqrt{N_E}+0.12+\frac{0.11}{\sqrt{N_E}}\right]D\right\},$$

$$Q_{\text{KS}}(\lambda)=2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2},$$

$$N_E=\frac{N_1N_2}{N_1+N_2},$$

where N_1 and N_2 are the number of points in the S and S^* embedding spaces. Since S^* is constructed by adding noise to S, N_1 and N_2 are equal.

Operationally, an embedding criterion is deemed to be robust to noise if noise has a minimal impact on the cumulative distribution of interpoint distances in the embedding space. This is indicated by a high value of P_{null}. The results are presented in Table II. A value of ''no result'' is entered in this table if the embedding criterion in question failed to converge on values of m and L. Three noise levels corresponding to signal-to-noise ratios of 10, 5, and 0 dB were computed.

Once again there is little criterion-dependent difference in the results obtained with the Rössler data. All of the methods with the exception of the Gao-Zheng method are robust to a signal-to-noise ratio (SNR) of 5 dB (that is, a noise variance that is approximately 32% of the signal variance). They all fail uniformly at 0 dB, where the noise variance and the signal variance are equal. In the trials using the Mackey-Glass equation, we see a somewhat larger difference in performance among the methods. The Gao-Zheng, characteristic length, GFNN-A, and GFNN-MI methods all perform well down to a SNR of 5 dB. Strangely, Schuster's method performed better at the lower SNR of 0 dB than it did at 10 dB. Repeated trials produced similar results, and we can offer no reasonable explanation for this particular outcome.

In the trials using experimental data, we note even larger differences in performance among the five methods. In addition to GFNN-MI outperforming the other four methods, we

also note the failure of the Gao-Zheng and characteristic length methods to specify embedding parameters for these trials. Specifically, in the trials using seizure data, the characteristic length method failed for SNR's of 5 dB and lower. Additionally, the Gao-Zheng method failed for the original as well as the noise corrupted data sets for the case of the seizure data. These time series are apparently too noiselike to produce interpretable results when the Gao-Zheng and characteristic length procedures are applied.

VI. CONCLUSIONS

We conclude that in these trials the global false nearest neighbors method outperformed the other three procedures for determining the embedding dimension. Additionally, when used in combination with GFNN, the first minimum of the mutual information function gave a more successful value of the lag than the first zero of the autocorrelation function. However, before generalizing these results inappropriately, other factors should be considered. One must ask, is a given method consistent in its interpretation? That is, could different researchers interpret the results in the same way? In this regard, GFNN-A has advantages over the other methods. A disadvantage that those procedures share is the need to estimate where a maximum or minimum of some function has occurred. While in principle this is simple, time series that are very complex or noise corrupted can make this a difficult task. One sometimes has to choose between what could be a sharp but specious minimum caused by noise and what appears to be a more general trend. These complications of interpretation can lead to conflicting results. This is a problem that we have considered in our earlier work on estimating lag using the minimum of mutual information [43]. In that contribution, we suggested that the minimum might be estimated by first filtering the mutual information function.

Another disadvantage of the methods of Gao and Zheng, Schuster, and characteristic length is that, in addition to locating an extremum, one needs to decide if a significant change has occurred as the embedding dimension is increased. Potential difficulties in this regard can be seen in the diagrams of Sec. III. As originally published, these methods require subjective assessments that could cause different conclusions to be drawn from the same calculations. Global false nearest neighbors has an advantage over these methods because the indicated choice of embedding dimension is the minimum dimension for which the number of false nearest neighbors is zero or consistently below some explicitly specifiable threshold. There is no uncertainty in the interpretation of the results. Also, if an efficient $N \log N$ procedure is used to locate nearest neighbors, the method of global false nearest neighbors is significantly faster than the others.

We conclude by reiterating a limitation of this investigation that was made in the Introduction. These comparative computations have identified global false nearest neighbors combined with the first minimum of the mutual information function as the best procedure for identifying embedding parameters for these data. Strictly, these results are valid only for these data and these specific tests. While it is hoped that these results provide generally useful guidelines, this generalization has not been demonstrated mathematically.

ACKNOWLEDGMENTS

We would like to acknowledge support from the U.S. Department of Education Award No. H235J000001 to the Krasnow Institute and from Grant No. 601135N.4508.518.A0247 from the Office of Naval Research and the Navy Bureau of Medicine to the Naval Medical Research Center. We are grateful to Tanya Schmah, Mathematical Institute, Warwick University for her assistance, especially in the discussions of the underlying mathematics and her essential observations concerning the circularity of the embedding criterion literature.

APPENDIX: EMBEDDING OBSERVED DATA

Let the set $\{x_1, x_2, x_3, \ldots\}$, $x_j \in \Re$, [1] denote the sequential measurements of an observed signal. They can be voltage values recorded from an EEG or a sequence of heart interbeat intervals. These values are used to create a set of embedded points $\{X_j\} \in \Re^m$, where

$$X_j = (x_j, x_{j+1}, x_{j+2}, \ldots, x_{j+m-1})$$

(the case of a nonunitary value of the lag will be considered presently). The parameter m is the embedding dimension. The criterion for selecting m and generalizations of the embedding procedure will be discussed presently. The time-dependent behavior of $\{X_j\}$ is the trajectory in an m-dimensional space specified by $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \cdots$. The analysis of the original time series $\{x_j\}$ proceeds as an examination of the geometry of the m-dimensional set $\{X_j\}$. This is motivated by the Takens-Mañé embedding theorem [1,2], which shows that the dynamical properties of the system that generated the observed signal are reflected in $\{X_j\}$. A simplified statement of the theorem follows.

It is assumed that the observed signal is generated by a dynamical system composed of ω real variables. For complex systems, ω will be very large, and not all ω variables will be directly observable. As a function of time the dynamical system moves on a compact behavior space P which is a subset of $\Re^\omega$. The compactness (bounded and closed) of the behavior space is an assumption. However, we could never contradict it with real data. P is also called the state space or the phase space. In abstract terms the dynamical system is a continuous map Ψ acting on the behavior space, $\Psi: P \rightarrow P$. For any given initial point y, $y \in P \subseteq \Re^\omega$, the state of the system at time t is given by $\Psi^t(y)$. The object of signal analysis is to infer properties of Ψ from $\{x_j\}$, in this case by an examination of $\{X_j\}$.

Let $y_j \in P$ denote the position of the true system at the ith sample time. The value $x_j \in \Re^1$ is the value of the observed scalar variable at that time. It is assumed that x_j is related to y_j by a smooth map c, $c: P \rightarrow \Re^1$, such that $c(y_j) = x_j$ for all j. Additionally, it is assumed that the set of y_j's corresponding to the observed x_j's forms a dense subset of P. Φ is defined as follows. For any integer m, $m > 2\omega$, define $\Phi: P \subseteq \Re^\omega \rightarrow \Re^m$ by

$$\Phi(y)=(c(y),c(\Psi(y))),c(\Psi^2(c)),\ldots,c(\Psi^{m-1}(y)).$$

Since $\Psi(y_j)=y_{j+1}$ and $c(y_j)=x_j$

$$\Phi(y_j)=(x_j,x_{j+1},x_{j+2},\ldots,x_{j+m-1}).$$

Theorem. (1) For almost any Ψ and c, Φ is an embedding. That is, P is diffeomorphic to its image under Φ. (2) The continuous extension map $X_j \rightarrow X_{j+1}$ corresponds, under the diffeomorphism, to the original map Ψ. Therefore, the observed trajectory $X_j \rightarrow X_{j+1}$ is intimately related to the true, high dimensional system Ψ. Specifically, the relationship is a diffeomorphism (a differentiable function with a differentiable inverse). Properties of $X_j \rightarrow X_{j+1}$ as established by observed data will, up to a diffeomorphism, also be true of Ψ. Thus if the conditions of the theorem are met, we can make meaningful inferences about Ψ from $\{X_j\}$.

This is a remarkable result. It states, subject to the conditions of the theorem, that we can perform an analysis of an ω-dimensional dynamical system based on observations of a single variable. However, in the real world the conditions of the theorem are never met. The crucial assumption is that the set of y_j's corresponding to the observed x_j's forms a dense subset of behavior space P. This is clearly impossible given a finite data set $\{x_j\}$. Nonetheless, as an approximation, $X_j \rightarrow X_{j+1}$ can provide valuable insights into Ψ. Since $\{x_j\}$ is finite, a number of practical issues arise. Recall the definition of X_j:

$$X_j=(x_j,x_{j+1},x_{j+2},\ldots,x_{j+m-1}).$$

A revision of this definition that incorporates a lag L, $L \in I^+$, can help space the observed data through the approximate behavior space and thus better approximate the density requirement of the theorem:

$$X_j=(x_j,x_{j+L},x_{j+2L},\ldots,x_{j+(m-1)L}).$$

This can be addressed in the preceding analysis by incorporating a dependence on L into the definition of Ψ.

Limitations imposed by the finite size of $\{x_j\}$ can be addressed in part by observing more than one dynamical variable. The embedding procedure can be generalized to incorporate multichannel data [4]. Suppose data are recorded from K observed variables. Let $\{x_j^i\}$ denote the time series of the ith channel:

$$\{x_j^i\}=(x_1^i,x_2^i,x_3^i,\ldots).$$

The easiest procedure is to construct the embedded set in $\Re^K$ by

$$X_j=(x_j^i,x_j^2,\ldots,x_j^K).$$

For example, if three variables w, x, and y are recorded, $\{X_j\}$ can be formed in $\Re_3$ by

$$X_j=(w_j,x_j,y_j).$$

This procedure can fail if K, the number of observed variables, is less than the effective dimension of the generating dynamical system. In that case, the procedure for embedding scalar data to an arbitrary dimension can be generalized:

$$X_j=(x_j^1,x_j^2,\ldots,x_j^K,x_{j+1}^1,x_{j+1}^2,\ldots,x_{j+1}^K,\ldots).$$

[1] F. Takens, in *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics Vol. 898, edited by D. A. Rand and L. S. Young (Springer-Verlag, New York, 1980).
[2] R. Mañé, in *Dynamical Systems and Turbulence* (Ref. [1].
[3] L. Noakes, Int. J. Bifurcation Chaos Appl. Sci. Eng. **1**, 867 (1991).
[4] T. Sauer, J. A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[5] P. E. Rapp, Biologist (London) **40**, 89 (1993).
[6] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[7] H. D. I. Abarbanel and M. B. Kennel, Phys. Rev. E **47**, 3057 (1993).
[8] A. M. Albano, A. Passamante, and M. E. Farrell, Physica D **54**, 85 (1991).
[9] Z. Aleksic, Physica D **52**, 362 (1991).
[10] A. M. Fraser, Physica D **34**, 391 (1989).
[11] J. Gao and Z. Zheng, Phys. Lett. A **181**, 153 (1993).
[12] J. Gao and Z. Zheng, Europhys. Lett. **25**, 485 (1994).
[13] J. F. Gibson, J. D. Farmer, M. Casdagli, and S. Eubank, Physica D **57**, 1 (1992).
[14] H. Kantz, Stoch. Dyn. **1**, 85 (2001).
[15] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997).
[16] G. Kember and A. C. Fowler, Phys. Lett. A **179**, 72 (1993).
[17] M. D. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).
[18] D. Kugiumtzis, Physica D **95**, 13 (1996).
[19] W. Liebert and H. G. Schuster, Phys. Lett. A **142**, 107 (1988).
[20] H. Schuster, in *Measures of Complexity and Chaos*, edited by N. B. Abraham, A. M. Albano, A. Passamante, and P. E. Rapp (Plenum, New York, 1989).
[21] R. Wayland, D. Bromley, D. Pickett, and A. Passamante, Phys. Rev. Lett. **70**, 580 (1993).
[22] R. Wayland, D. Bromley, D. Pickett, and A. Passamante, Physica D **79**, 320 (1994).
[23] A. Rapoport, Sci. Am. **217**, 50 (1967).
[24] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, Meccanica **15**, 9 (1980).
[25] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, Meccanica **15**, 21 (1980).
[26] O. E. Rössler, Phys. Lett. A **57**, 397 (1976).
[27] J. D. Lambert, *Computational Methods in Ordinary Differential Equations* (Wiley, New York, 1973).
[28] M. C. Mackey and L. Glass, Science (Washington, DC, U.S.) **197**, 287 (1977).

[29] W. Liebert, K. Pawelzik, and H. G. Schuster, Europhys. Lett. **14**, 521 (1991).
[30] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, Physica D **16**, 285 (1985).
[31] J. H. Bentley, IEEE Trans. Software Eng. **SE-5**, 333 (1979).
[32] W. Cunto, G. Lau, and P. Flajolet, Lect. Notes Comput. Sci. **382**, 24 (1989).
[33] A. V. Aho, *Data Structures and Algorithms* (Addison-Wesley, Reading, MA, 1983).
[34] H. W. DeVenema, Pattern Recogn. Lett. **12**, 445 (1991).
[35] D. Ibaroudene and R. Aiharya, Inf. Sci. (N.Y.) **68**, 123 (1993).
[36] E. Vidal, Pattern Recogn. Lett. **15**, 1 (1994).
[37] T. Schreiber, Int. J. Bifurcation Chaos Appl. Sci. Eng. **5**, 349 (1995).
[38] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer, New York, 1996).
[39] C. J. Cellucci, A. M. Albano, P. E. Rapp, and A. D. Krystal, Int. J. Bifurcation Chaos Appl. Sci. Eng. (to be published).
[40] T. A. A. Watanabe, C. J. Cellucci, E. Kohegyi, T. R. Bashore, R. C. Josiassen, N. N. Greenbaun, and P. E. Rapp, Psychophysiology **40**, 1 (2003).
[41] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1986).
[42] A. M. Albano, P. E. Rapp, and A. Passamante, Phys. Rev. E **52**, 196 (1995).
[43] J. Martinerié, A. M. Albano, A. I. Mees, and P. E. Rapp, Phys. Rev. A **45**, 7058 (1992).
[44] J. Theiler, Phys. Rev. A **34**, 2427 (1986).