6-2005

Statistical Validation of Mutual Information Calculations: Comparison of Alternative Numerical Algorithms

C. J. Cellucci

Alfonso M. Albano
Bryn Mawr College, aalbano@brynmawr.edu

P. E. Rapp

Citation

C.J. Cellucci, A.M. Albano and P.E. Rapp, *Phys. Rev. E* **71**, 66208 (2005).

Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms

C. J. Cellucci
Operational and Undersea Medicine Naval Medical Research Center, Silver Spring 20910, USA and Department of Pharmacology and Physiology Drexel University College of Medicine, Philadelphia, Pennsylvania 19102, USA

A. M. Albano*
Department of Physics, Bryn Mawr College, Bryn Mawr, Pennsylvania 19010, USA

P. E. Rapp
Department of Pharmacology and Physiology Drexel University College of Medicine, Philadelphia, Pennsylvania,19102, USA and Operational and Undersea Medicine Naval Medical Research Center, Silver Spring 20910,USA
(Received 19 March 2004; revised manuscript received 17 March 2005; published 22 June 2005)

Given two time series X and Y, their mutual information, $I(X,Y)=I(Y,X)$, is the average number of bits of X that can be predicted by measuring Y and vice versa. In the analysis of observational data, calculation of mutual information occurs in three contexts: identification of nonlinear correlation, determination of an optimal sampling interval, particularly when embedding data, and in the investigation of causal relationships with directed mutual information. In this contribution a minimum description length argument is used to determine the optimal number of elements to use when characterizing the distributions of X and Y. However, even when using partitions of the X and Y axis indicated by minimum description length, mutual information calculations performed with a uniform partition of the XY plane can give misleading results. This motivated the construction of an algorithm for calculating mutual information that uses an adaptive partition. This algorithm also incorporates an explicit test of the statistical independence of X and Y in a calculation that returns an assessment of the corresponding null hypothesis. The previously published Fraser-Swinney algorithm for calculating mutual information includes a sophisticated procedure for local adaptive control of the partitioning process. When the Fraser and Swinney algorithm and the algorithm constructed here are compared, they give very similar numerical results (less than 4% difference in a typical application). Detailed comparisons are possible when X and Y are correlated jointly Gaussian distributed because an analytic expression for $I(X,Y)$ can be derived for that case. Based on these tests, three conclusions can be drawn. First, the algorithm constructed here has an advantage over the Fraser-Swinney algorithm in providing an explicit calculation of the probability of the null hypothesis that X and Y are independent. Second, the Fraser-Swinney algorithm is marginally the more accurate of the two algorithms when large data sets are used. With smaller data sets, however, the Fraser-Swinney algorithm reports structures that disappear when more data are available. Third, the algorithm constructed here requires about 0.5% of the computation time required by the Fraser-Swinney algorithm.

DOI: 10.1103/PhysRevE.71.066208 PACS number(s): 05.45.−a

I. INTRODUCTION

Given two time series $\{X\}=\{x_1,x_2,\ldots x_{N_D}\}$ and $\{Y\}=\{y_1,y_2,\ldots y_{N_D}\}$, their mutual information, $I(X,Y)$, is the average number of bits of $\{X\}$ that can be predicted by measuring $\{Y\}$. It can be shown that this relationship is symmetrical, $I(X,Y)=I(Y,X)$. A systematic presentation of the definition of mutual information and its mathematical properties is given in Cover and Thomas [1]. In the analysis of observational data, calculation of mutual information occurs in three contexts: (i) identification of nonlinear correlation, (ii) determination of an optimal sampling interval, particularly when embedding time series data, and (iii) in the investigation of causal relationships with directed mutual information. Each of these contexts will now be briefly described.

Mutual information can be used to identify and quantitatively characterize relationships between data sets that are not detected by commonly used linear measures of correlation. Figure 1 recapitulates an example shown in Mars and Lopes da Silva [2] and displays three data set pairs. The first shows x_i when $x_i=-3$ to $+3$ in steps of 0.0006 plotted against ε_i, a random normally distributed variable with zero mean and unit variance. The second element of Fig. 1 shows x_i vs $x_i+0.2\varepsilon_i$ where ε_i is the previously used random variable. In the third example of Fig. 1, $y_i=x_i^2+0.2\varepsilon_i$. Four measures were calculated with 10 000 element data sets: (i) the Pearson linear correlation coefficient r, (ii) the Spearman rank order correlation r_S, (iii) Kendall's tau, a nonparametric measure of correlation, and (iv) the mutual information between $\{X\}$ and $\{Y\}$ using an algorithm that will be described in a subsequent section. The corresponding probabilities P_{null} of the null hypothesis of zero linear correlation for each of the four measures were also calculated.

*Author to whom correspondence should be addressed. Electronic address: aalbano@brynmawr.edu

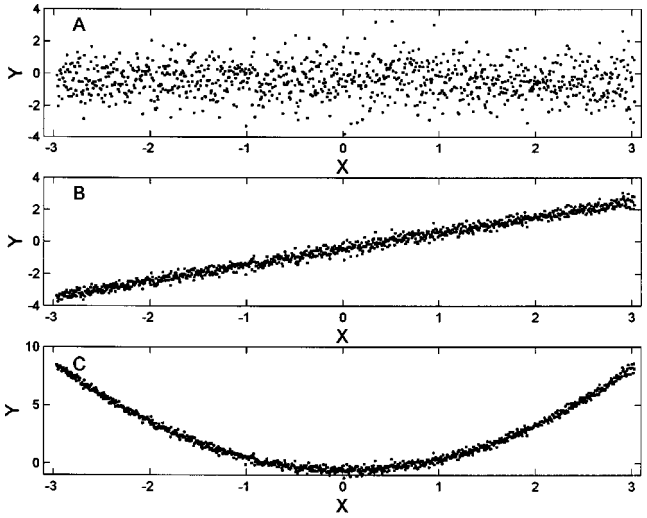

FIG. 1. Data sets used in the correlation study of Table I. In each case, x varies from −3 to +3 in steps of 0.0006. (A) $y_i = \varepsilon_i$, a normally distributed random variable with zero mean and unit variance. (B) $y_i = x_i + 0.2\varepsilon_i$. (C) $y_i = x_i^2 + 0.2\varepsilon_i$.

The results are shown in Table I. In the case of normally distributed random numbers, all four measures behave in a manner that is consistent with our qualitative understanding of the word correlation. Similarly, in the case of calculations with linearly correlated noise the results are consistent with expectations.

The results obtained in the case of parabolic correlation merit closer inspection. The first three measures r, r_S, and τ are small and the corresponding P_{null} values are high which indicates that no correlation was detected. In contrast, the value of mutual information is high, essentially equal to that obtained using linearly correlated data, and the probability of the null hypothesis of statistical independence is zero.

In the second context, mutual information estimates can be also used to determine an appropriate sampling interval T_S, which is the time between consecutive measurements of a time series. Many of the calculations presented here will be calculations directed to this question. The selection of an appropriate sampling interval is an important consideration when the quantitative methods of dynamical analysis are applied to time series data. On first consideration, one might suppose that the smallest possible T_S would be the best option. While this may be a reasonable approach during data acquisition, this strategy can fail during analysis because calculations with oversampled data can produce misleading results [3]. Historically, calculation of the autocorrelation time, the time required for the autocorrelation function to drop to $1/e$ of its initial value, has been used to establish an approximate sense of the time scale corresponding to significant changes in a time series' behavior. However, as we have seen in the preceding calculations, linear measures can give an incomplete characterization of behavior. This recognition has motivated the calculation of lagged mutual information.

Let $\{X\}$ be the original time series, and let time series $\{Y\}$ be the same time series shifted by a time lag, that is, $y_i = x_{i+lag}$. The mutual information $I(X_i, X_{i+lag})$ is then calculated as a function of lag. In order to get the most new information from a measurement, we want to take the next measurement when there is maximum uncertainty in the relationship between $\{X\}$ and $\{Y\}$. The maximum uncertainty in the relationship between $\{X\}$ and $\{Y\}$ will occur at a minimum of $I(X_i, X_{i+lag})$. Fraser and Swinney [4] argue that among the many different minima of $I(X_i, X_{i+lag})$, the sampling interval should correspond to the first minimum of $I(X_i, X_{i+lag})$.

A specific application of $I(X_i, X_{i+lag})$ calculations can occur when embedding dynamical data. In the simplest case, an analysis based on embedded data begins with a scalar time series $\{X\}$. The elements of $\{X\}$ are then used to form an m-dimensional set $\{Z\} \in \mathbb{R}^m$ with the construction

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

The analysis continues with the investigation of the geometrical properties of $\{Z\}$. A crucial operational difficulty is encountered when embedding finite observational data sets. Embedding parameters m and lag must be chosen. Inappropriate choices of m and lag can result in the spurious indication of structure in random data [3]. Conversely an inappropriate specification can, in other cases, result in the unnecessary failure to identify structures that are indeed present in the time series. Several candidate criteria for selecting m and lag have been proposed. An incomplete review of the very large embedding criterion literature is given in Cellucci, *et al.* [5]. Fraser and Swinney [4] proposed that the best value of lag to use in an embedding is given by the first minimum of the $I(X_i, X_{i+lag})$ vs lag function. This proposal is supported by Abarbanel [6]. To a limited degree the Fraser-Swinney proposal was confirmed in a recent comparative study of embedding criteria [5].

A third circumstance in which calculation of mutual information is important is in the characterization of causal relationships between two time series. By definition, a correlation measure, either linear or nonlinear, quantifies the degree of correlation between $\{X\}$ and $\{Y\}$ under their respective definitions, but correlation does not necessarily identify causal relationships in the sense of identifying which variable drives the other, if indeed such a relationship exists. Historically the most commonly employed measure of cau-

TABLE I. Correlation analysis.

	Pearson r	Pearson P_{null}	Spearman r_S	Spearman P_{null}	Kendall's tau	Kendall's P_{null}	$I(X,Y)$	$I(X,Y)P_{null}$
Normally distributed random	−0.0037	0.7112	−0.0040	0.6854	0.0027	0.6845	0.1356	0.7851
Linearly correlated	0.9934	0	0.9936	0	0.9270	0	2.9186	0
Parabolically correlated	0.0001	0.9912	$<10^{-4}$	0.9928	$<10^{-5}$	0.9989	3.0304	0

sality in economics research is Granger causality [7,8] which is based on the construction of bivariate autoregressive processes. A complementary procedure for the investigation of causal relationships can be constructed by examining delayed mutual information functions. Stated informally, if a measurement of variable x can predict the future of y more effectively than measurement of y can predict x, then, in that limited sense, in an isolated system variable x can be said to drive variable y. Xu *et al.* [9] describe $I(X_i, Y_{i+\tau})$ as the rate of information transmission from variable x to variable y at a delay of τ. Several investigators have used this technique to assess the time dependence of between channel information transfer in multichannel EEGs [9–13]. Significant limitations of causality measures based on lagged mutual information have been identified by Schreiber [14]. He argues, in our view correctly, that "time delayed mutual information fails to distinguish between information that is exchanged from shared information due to common history and inputs." He addresses these limitations with the construction of a transfer entropy.

II. CALCULATING $I(X,Y)$ WITH A UNIFORM PARTITION OF THE XY PLANE

Let $\{X\}=\{x_1,x_2,x_3\ldots x_{N_D}\}$ and $\{Y\}=\{y_1,y_2,y_3\ldots y_{N_D}\}$ be time series of equal length. Suppose that the distributions of X and Y, $P_X(i)$ and $P_Y(j)$ are approximated by histograms of N_X and N_Y elements that uniformly divide the range x_{min}–x_{max} and y_{min}–y_{max}. It is not necessary for N_X to be equal to N_Y. Let $O_{XY}(i,j)$ denote the occupancy of the (i,j)th element of the partition of the XY plane that extends from x_{min} to x_{max} on the X axis (N_X equal elements) and from y_{min} to y_{max} on the Y axis (N_Y equal elements). $P_{XY}(i,j)$ is determined by normalizing the occupancy against the number of paired observations; $P_{XY}(i,j)=O_{XY}(i,j)/N_D$. The joint probability distribution, $P_{XY}(i,j)$, has N_XN_Y values, many of which may be zero. A discrete approximation of $I(X,Y)$ is computed using the following relation [1]:

$$I(X,Y)=\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}P_{XY}(i,j)\log_2\left\{\frac{P_{XY}(i,j)}{P_X(i)P_Y(j)}\right\}, \qquad (1)$$

where there is no contribution to the sum if $P_{XY}(i,j)$ is equal to zero.

While easy to implement, this procedure for estimating mutual information contains a serious deficiency. The calculation will be sensitive to the choice of N_X and N_Y. An example is shown in Fig. 2. $I(X_i,X_{i+lag})$ is plotted as a function of lag, for data generated by the Lorenz system,

$$dx/dt=\sigma(x-y),$$

$$dy/dt=-xz+rx-y,$$

$$dz/dt=xy-bz,$$

where $\sigma=10$, $b=8/3$, and $r=28$. Ten thousand values of the x variable of the Lorenz system were used in calculations where the number of bins in the distribution histogram is the same for both variables. $N_X=N_Y=N_{elements}$. equally sized elements partition each axis. In these calculations, a well characterized minimum of $I(X_i,X_{i+lag})$ appears at lag=18 when $N_{elements}=50$. However, as the diagram indicates, this minimum is lost if other values of $N_{elements}$ are used. Since the location of the first minimum of the $I(X_i,X_{i+lag})$ vs lag is frequently the object of a mutual information calculation, this result argues against the common practice of selecting N_X and N_Y arbitrarily.

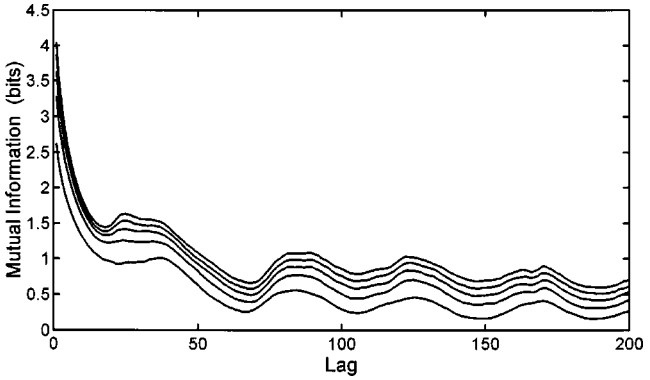

FIG. 2. $I(X_i,X_{i+lag})$ as a function of lag. Ten thousand consecutive values of the Lorenz x variable were used. In the case of the top curve, $N_{elements}=50$. The value of $N_{elements}$ decrease in steps of 10 to the lower curve where $N_{elements}=10$.

The preceding example indicates that the value of mutual information can be sensitive to the number of elements used when a uniform partition of the XY plane is implemented. We must therefore address the question what is the optimal number of elements? This is a restatement of the histogram problem in the specific context of mutual information calculations. The histogram problem is: given a scalar data set $X=\{x_1,x_2,\ldots x_n\}$, how many elements should be used to construct a histogram of X? If there are too many elements, each element has an occupancy of 0 or 1 and fails to identify the distribution of X in a meaningful way. Similarly, if there are only a small number of elements (consider the limiting case of a single element), the structure of the distribution cannot be discerned. A successful answer therefore lies at an intermediate value. The histogram problem has a long history and has been examined by several investigators [15–17].

Tukey [17] suggested that $n^{1/2}$, where n is the number of observations, is the best choice. Bendat and Piersol [15] recommended $1.87(n-1)^{0.4}$. A systematic theoretical development of the question is given by Rissanen [18]. Rissanen uses a minimum description length argument to conclude that the optimal value of the number of elements to use in a histogram is the value of m, m_{opt}, that gives a minimum value of the stochastic complexity, $F(m)$,

$$F(m)=n\log_2\left(\frac{R}{m\Delta}\right)+\log_2\binom{n}{n_1,\ldots,n_m}+\log_2\binom{n+m-1}{n}.$$

n is the number of data points in set X. R is the range of X, $R=x_{max}-x_{min}$. m is the number of elements in a uniform partition. Δ is the resolution of the measurement of x, and

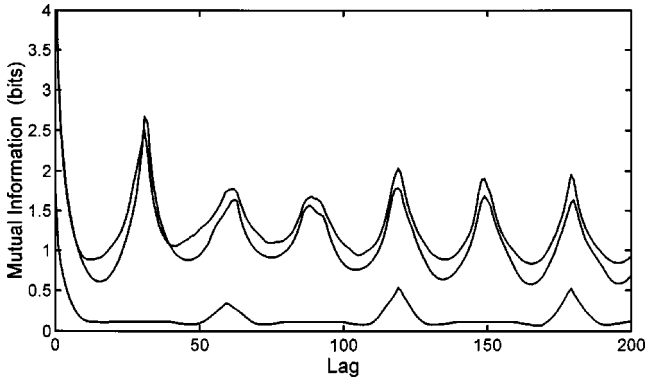

FIG. 3. Mutual information $I(X_i, X_{i+lag})$ as a function of lag for Rössler data. A uniform partition of the XY plane was constructed using 40 elements on each axis. 100 000 data points were used. The top curve was obtained with variable x. The curve immediately below it was constructed with variable y data. The lower curve was calculated with variable z data.

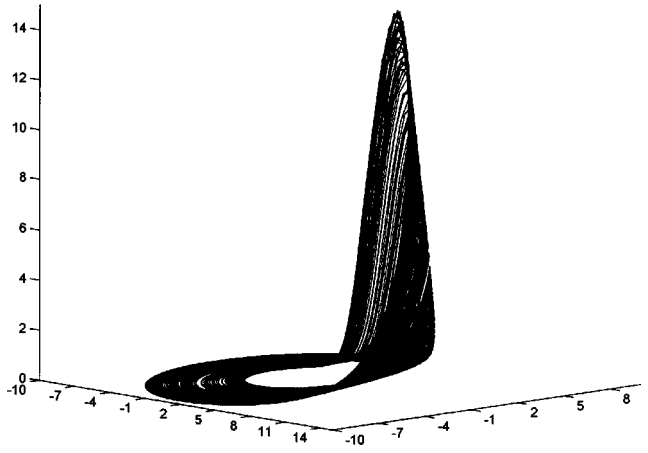

FIG. 4. Three-dimensional construction of the Rössler attractor using 10 000 point x, y, and z vectors generated using the differential equation and parameter values specified in the text.

n_1, n_2, $\ldots n_m$ are the occupancies of each element in the partition. The multinomial coefficient is

$$\binom{n}{n_1,\ldots,n_m} = \frac{n!}{n_1!\, n_2!\, \ldots n_m!}$$

and the binomial coefficient is

$$\binom{n+m-1}{n} = \frac{(n+m-1)!}{n!\,(m-1)!}.$$

The value of Δ only shifts the function by an additive constant. It will not affect the value of m_{opt}. If the only object of the calculation is to determine m_{opt}, Δ can be set equal to 1. Base two logarithms are used throughout the development in Rissanen, but again if the sole object is a determination of m_{opt}, the choice of base is immaterial.

$F(M)$ was calculated using the Lorenz data used to construct Fig. 2 A minimum was obtained at $M_{opt}=32$. Using this value for the number of elements in the uniform partition of the X and Y axes in a calculation of $I(X_i, X_{i+lag})$ gives a mutual information versus lag function with a well characterized first minimum at lag=21. This analysis would seem therefore to provide a rational procedure for calculating $I(X,Y)$. Application to the Rössler equations, however, raises additional questions. The Rössler equations used in the next calculations were

$$dx/dt = -y - z,$$

$$dy/dt = x + 0.2y,$$

$$dz/dt = 0.4 + xz - 5.7z.$$

Using x-axis data generated by this system, a calculation of the Rissanen $F(M)$ gives a minimum at $M=40$. A 40-element partition of each axis was used in the subsequent calculations of mutual information as a function of lag for x-, y-, and z-variable data. The resulting mutual information versus lag functions are shown in Fig. 3. It is seen that while x-axis and y-axis data give functions with first minima that are roughly coincident, the function obtained with z-axis data is very different.

The cause of the differences in the z-variable mutual information function in Fig. 3 can be identified by examining a three-dimensional construction of the trajectory using all three variables (Fig. 4). The activity of the Rössler system is confined predominantly to the $z \approx 0$ plane. At irregular, chaotic intervals there is an abrupt excursion into the $z > 0$ domain. An examination of the histograms formed with x, y, and z data (Fig. 5) shows that while the x and y values are approximately uniformly distributed, most of the activity of the z variable is confined to [0,0.375] even though the maximum value of z is approximately 15.

The value of optimal lag produced by the mutual information functions of Fig. 3 are lag=13, 16, and 48 for x, y, and z, respectively. Should we expect the values of optimal embedding lag to be the same for all three variables? While it can be argued that there is no *a prori* reason to suppose that they should be equal, there is a specific context in which a disparity of optimal lag values is problematic. Thus far we have considered embeddings based on a scalar variable where $Z_j=(x_j, x_{j+lag}, \ldots x_{j+(m-1)lag})$. However, in applications with experimental data where multichannel recordings are obtained, a multichannel embedding can be utilized [19,20]. In the specific case where variables x, y, and z are recorded, Z_j becomes $Z_j=(x_{1+(j-1)lag}, y_{1+(j-1)lag}, z_{1+(m-1)lag})$. In applications of this type, a common value of lag is required. The question then becomes, which value should be used?

A resolution of this difficulty, at least for the Rössler data used here, can be found by re-examining the mutual information versus lag calculations displayed in Fig. 3. A calculation of $F(M)$ using data obtained from variable x gave a value of $M_{opt}=40$. This value was used to specify the number of elements in a uniform partition calculation of mutual information. The same number of elements was used in calculations with y and z data. This is inappropriate. When $F(M)$ is calculated with data from the other variables, a value of $M_{opt}=54$ is obtained with y data, and a value of $M_{opt}=852$ is obtained with z data. The high z value of M_{opt} can be understood by examining the histogram in Fig. 5. (Note that the

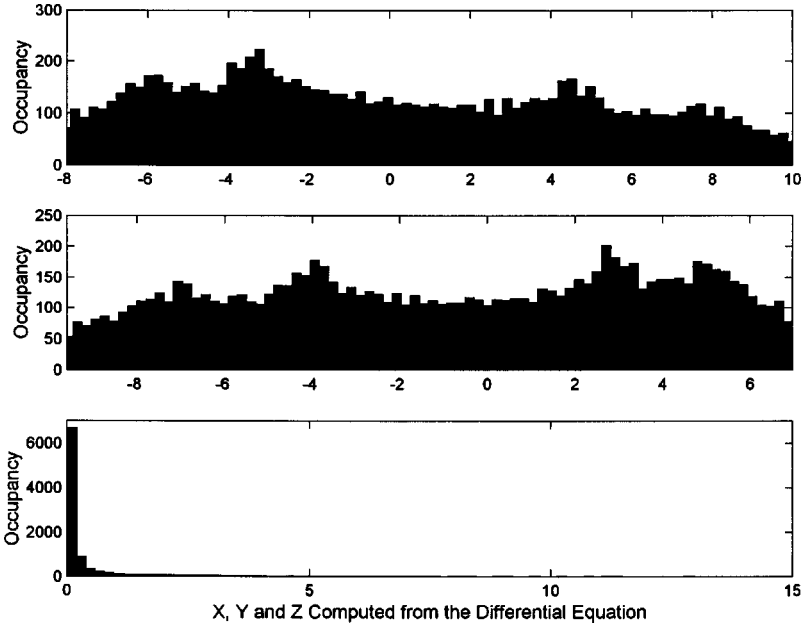

FIG. 5. Histograms constructed with Rössler data. The histograms were formed with the 10 000 points used to construct the three-dimensional attractor of Fig. 4. *X* data were used to construct the top histogram. *Y* data were used to construct the middle histogram, and the bottom histogram displays *Z* data. Note that the ranges of the vertical axes are different.

range of the vertical axes of the *x* and *z* histograms differ by a factor of 20.) The distribution of the *x* and *y* variables between their respective maximum and minimum values is approximately uniform. As previously observed, most activity of the *z* variable is confined to [0, 0.375] even though the maximum value of *z* is approximately 15. Because the *z* distribution is so strongly nonuniform, a much higher number of partition elements are needed to recover the fine structure of that variable's distribution.

Mutual information versus lag calculations were again performed with a uniform partition algorithm. In contrast with the calculations shown in Fig. 3, the results displayed in Fig. 6 were obtained in calculations in which the number of elements in each partition were determined by a minimum description length argument, the minimum of $F(M)$, that is specific to each variable. When 852 elements are used to partition each axis in the calculation with *z* data, the resulting mutual information function is qualitatively similar to functions obtained with *x* and *y* data. The optimal lags, the first minimum of the mutual information versus lag function, for *x*, *y*, and *z* are 13, 16, and 17, respectively.

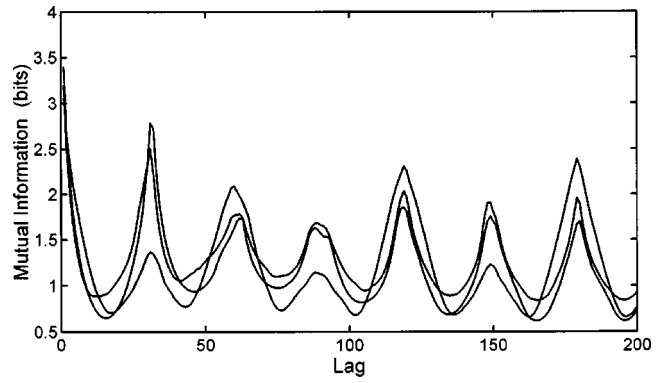

FIG. 6. Mutual information $I(X_i, X_{i+\mathrm{lag}})$ as a function of lag for the Rössler data. In the case of variable *x* data, a uniform partition of the *XY* plane was constructed using 40 elements on each axis. For the variable *y* data, 54 elements were used on each axis, and for the variable *z* data 852 elements were used on each axis. 100 000 data points were used in each calculation. Identifying at lag = 35, the top curve corresponds to variable *y*, the second curve corresponds to variable *x*, and the lowest curve to variable *z*.

The sensitivity of mutual information estimates to computational parameters identifies a compelling need for the systematic statistical validation of these calculations. This requirement motivated the construction of the algorithm described in Sec. III and IV.

III. STATISTICAL ASSESSMENT OF $I(X,Y)$ CALCULATIONS

The results with Rössler data suggest that the calculation of mutual information using a uniform partition can produce misleading conclusions. An alternative to uniform partitioning should therefore be sought. An additional and arguably more important issue should also be addressed. The calculations of mutual information should be constructed on a sound statistical foundation. When computing $I(X,Y)$ we should incorporate a statistical test of the confidence of our rejection of the null hypothesis that *X* and *Y* are statistically independent. $I(X,Y)=0$ if *X* and *Y* are statistically independent. In practice, we wish to know if a computed nonzero value of $I(X,Y)$ is statistically significant. Therefore, given time series *X* and *Y*, our object is to assess the null hypothesis that *X* and *Y* are statistically independent.

The null hypothesis of statistical independence can be addressed in the following manner. Suppose that the distributions of variables *X* and *Y* are approximated by histograms of N_X and N_Y elements. In most applications $N_X = N_Y$, but this is not required. $O_X(i)$ is the observed occupation number of the *i*th bin of the variable *X* histogram. $O_Y(j)$ is assigned analogously. $O_{XY}(i,j)$ is the observed occupation number of element i,j of the *XY* partition. $E_{XY}(i,j)$ is the expected occupancy of element i,j of the *XY* partition given the

assumption that X and Y are statistically independent

$$E_{XY}(i,j) = N_D P_X(i) P_Y(j) = N_D \left\{ \frac{O_X(i)}{N_D} \right\} \left\{ \frac{O_Y(j)}{N_D} \right\} = \frac{O_X(i) O_Y(j)}{N_D},$$

where N_D is the number of x,y pairs.

Following conventional statistical practice [21,22], we require $E_{XY}(i,j) \geqslant 1$ for all elements of the partition and $E_{XY}(i,j) \geqslant 5$ for at least 80% of these elements (the "Cochran criterion"). The value of χ^2 is

$$\chi^2 = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \frac{\{O_{XY}(i,j) - E_{XY}(i,j)\}^2}{E_{XY}(i,j)}.$$

The condition $E_{XY}(i,j) \geqslant 1$ for all values of i,j ensures that χ^2 is well behaved. In addition to χ^2, ν, the number of degrees of freedom, is also computed,

$$\nu = (N_X - 1)(N_Y - 1).$$

Using χ^2 and ν, the probability of the statistical independence null hypothesis is computed,

$$P_{\text{null}} = \text{probability of the null hypothesis} = Q\left(\frac{\nu}{2}, \frac{\chi^2}{2} \right).$$

Q is the incomplete gamma function,

$$Q(x,y) = 1 - \frac{1}{\Gamma(x)} \int_0^y e^{-t} t^{x-1} dt = \frac{1}{\Gamma(x)} \int_y^\infty e^{-t} t^{x-1} dt \quad \Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt.$$

IV. CALCULATION OF $I(X,Y)$ USING AN ADAPTIVE XY PARTITION

As previously outlined, we propose that calculation of mutual information should be statistically validated by application of a χ^2 test of the null hypothesis of statistical independence. Additionally, the partition of the XY plane, which is used to calculate the joint probability distribution P_{XY}, should satisfy the Cochran criterion on the expectancies E_{XY}. In the following algorithm, we use the expectation criterion to construct a nonuniform XY partition. This procedure has two advantages over the use of a naïve uniform partition. First, it reduces sensitivity to outlying values of X and Y. Second, it provides an approximation of the highest partition resolution consistent with the expectation criterion.

Let N_D denote the number of X, Y pairs. N_X is the number of elements used in the partition of the x axis. N_Y is the number of elements used to partition the y axis. For this implementation of the algorithm, N_X and N_Y are equal and denoted by the number of elements N_E. N_E is determined by the following procedure: after determining $x_{\min}$ and $x_{\max}$, the x axis is partitioned into N_E elements so that there is an equal occupancy in each element. This partition is nonuniform in the sense that the widths of each element are adjusted individually in order to meet the requirement of uniform occupancy. Let $P_X(i)$ denote the probability of X's membership in the ith element of the x axis partition. We have

$$P_X(i) = 1/N_E.$$

Similarly, after determining $y_{\min}$ and $y_{\max}$, the y axis is partitioned into N_E elements so that there is an equal number of occupants in each y axis element,

$$P_Y(j) = 1/N_E.$$

Under the null hypothesis of statistical independence, the expected occupancy of the (i,j)th element of the partition of the XY plane is

$$E_{XY}(i,j) = N_D P_X(i) P_Y(j) = \frac{N_D}{N_E^2}.$$

N_E is determined by finding the largest possible value that gives $E_{XY}(i,j) \geqslant 5$ for all elements of the XY partition. This criterion is therefore more conservative than the Cochran [21] criterion that requires E_{XY} to be greater than five in at least 80% of the elements. N_E is the greatest integer such that

$$N_E \leqslant \left(\frac{N_D}{5} \right)^{1/2}.$$

$P_{XY}(i,j)$ is calculated using this partition. Mutual information is calculated with Eq. (1). χ^2 and P_{null} are calculated as previously described. If N_D is exactly divisible by N_E, then the formula for mutual information simplifies and becomes

$$I(X,Y) = \sum_{i=1}^{N_E} \sum_{j=1}^{N_E} P_{XY}(i,j) \ln\{N_E^2 P_{XY}(i,j)\}.$$

However, when N_D is not a multiple of N_E, elements of the x axis and y axis partitions do not have exactly identical probabilities equal to $1/N_E$, and the preceding formula should be used. If the Cochran expectation criterion is satisfied (and by construction it will be) and the null hypothesis is not rejected, then, to the extent that can be determined by calculations with this algorithm, the two data sets are statistically independent. Under these conditions, reporting a nonzero value of mutual information cannot be justified. Therefore, in cases where the null hypothesis is not rejected, the algorithm returns $I(X,Y) = 0$ rather than the numerical value produced by the formula. This practice incorporates a conservative understanding of statistical significance. As an alternative, the numerical value of mutual information obtained from the algorithm and its uncertainty can be reported.

The application of this procedure to the Rössler data is shown in Fig. 7. In contrast with the results of Fig. 3, which were obtained with a uniform partition, it is seen that the first minimum of the mutual information versus lag functions obtained with x-, y-, and z-variable data approximately coincide when the adaptive partition is used. The probability of the null hypothesis was calculated for each value of lag. With these data, P_{null} was found to be numerically indistinguishable from zero for each value of lag. Since the data set Y used in these calculations of $I(X,Y)$ is a lagged version of

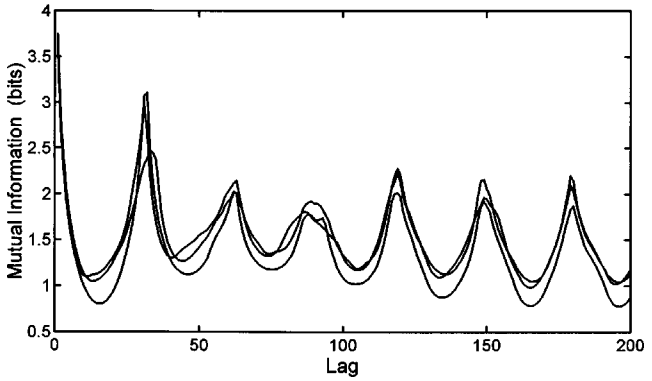

FIG. 7. Mutual information as a function of lag using Rössler data. Mutual information was calculated using an adaptive partition algorithm. The data used in Fig. 3 were used in these calculations. $N_{\text{data}}=100\,000$. Viewed at lag=18, the curves from the x, y, and z variables have the top-down order of x–z–y.

data set X, this rejection of the null hypothesis is anticipated.

Suppose that time series X is transformed by a monotone increasing function h_X where h_X may be nonlinear. Similarly suppose that time series Y is transformed by a monotone increasing function h_Y. The adaptive partition algorithm for calculating mutual information is then applied to calculate $I(h_X(X), h_Y(Y))$. These transforms are monotonic. Therefore while the values are changed, the relative ordering of elements in the time series are invariant. When the algorithm is applied, the location of the boundaries of axis partitions will be shifted but the occupancies of each element will be unchanged, that is, $P_X(i)$, $P_Y(j)$, and $P_{XY}(i,j)$ are unchanged. Therefore the value of mutual information is unchanged. This is summarized in the following result.

Theorem. Let X and Y be time series of equal length. Let h_X and h_Y be monotone increasing functions. If mutual information is calculated using the adaptive partition algorithm, then

$$I(X,Y) = I(h_X(X), h_Y(Y)).$$

V. FRASER-SWINNEY ALGORITHM

Fraser and Swinney [4,23] have constructed an alternative adaptive partition algorithm for calculating mutual information. As in the case of the previous algorithm, the calculation is directed to an estimate of the discrete form of the mutual information integral given in Eq. (1). Numerical approximation of the joint probability distribution P_{XY} constitutes the most demanding element of the computation. The Fraser-Swinney algorithm [4] does this by constructing a locally adaptive partition of the XY plane (see Fig. 8).

As a preliminary exercise leading to the construction of the algorithm, consider a sequence of partitions G_0, G_1, G_2, …, G_m. Each partition is a grid of 4^m elements generated by dividing the X and Y axis into 2^m equiprobable elements, that is the boundaries on the X and Y axis are positioned so that $P_X=P_Y=1/2^m$ for each element of the partition. G_0 is the entire XY plane. $R_m(K_m)$ denotes an element of the partition G_m.

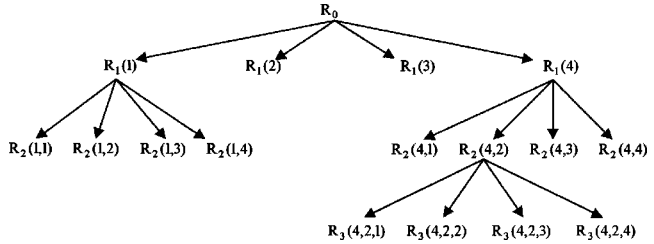

FIG. 8. Illustrative example of the adaptive partition employed by the Fraser-Swinney algorithm. In this hypothetical example, the substructure of elements $R_1(2)$ and $R_1(3)$ is approximately uniform and these elements are therefore not partitioned. Elements $R_1(1)$, $R_1(4)$, and $R_2(4,2)$ are partitioned into subelements because they meet the criterion for the presence of smaller scale structure.

A finer partition is used in areas of the XY plane where P_{XY} has nonuniform structure. For the hypothetical example in the diagram, P_{XY} is deemed to be approximately uniform on $R_1(2)$ and $R_1(3)$. The partitioning terminates with these elements. In contrast, $R_1(1)$ and $R_1(4)$ have locally nonuniform joint distributions and are partitioned. In this example, partitioning terminates at the G_2 level with the exception of element $R_2(4,2)$, which has a nonuniform joint distribution and is partitioned into four G_3 elements, $R_3(4,2,1)$–$R_3(4,2,4)$. The partitioning continues until the local joint distribution P_{XY} is approximately uniform.

In the case where P_{XY} is exactly uniform on $R_m(K_m)$, Fraser and Swinney [4] show that dividing the partition element into four subdivisions will have no effect on the contribution to mutual information obtained from that element. Terminating the partitioning process at level G_m is therefore justified in this case. As a practical matter, however, it is necessary to establish a criterion that can be used to terminate the partitioning process for some specific element $R_m(K_m)$ when P_{XY} is nearly, but not exactly, uniform on that element. In their paper, Fraser and Swinney construct a test for uniformity that uses a χ^2 test to examine structure on both the $m+1$ and $m+2$ generation partition of $R_m(K_m)$. Let $N=N(R_m(K_m))$ denote the number of XY pairs in element $R_m(K_m)$. Using analogous notation for the subdivisions, let $a_i=N(R_{m+1}(K_m,i))$ and let $b_{i,j}=N(R_{m+2}(K_m,i,j))$. By the Fraser and Swinney criterion, P_{XY} will be deemed to be effectively uniform on $R_m(K_m)$ and the partitioning process will be terminated on that element if both $\chi_3^2<1.547$ and $\chi_{15}^2<1.287$, where

$$\chi_3^2 = \left\{ \frac{16}{9}\left(\frac{1}{N}\right)\sum_{i=1}^{4}(a_i - N/4)^2 \right\},$$

$$\chi_{15}^2 = \left\{ \frac{256}{225}\left(\frac{1}{N}\right)\sum_{i=1}^{4}\sum_{j=1}^{4}(b_{i,j} - N/4)^2 \right\}.$$

It should be noted that while the Fraser-Swinney algorithm uses a χ^2 criterion to control subdivisions of the XY plane locally, it does not, in contrast with the algorithm of the previous section, provide a global statistical assessment of an $I(X,Y)$ calculation that includes the probability of the

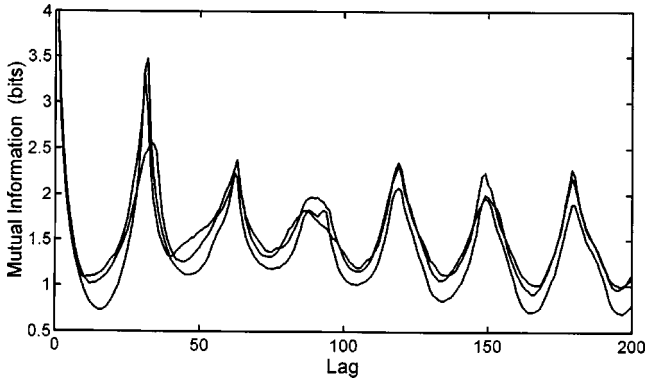

FIG. 9. Mutual information as a function of lag using the Rössler data of Fig. 3. Mutual information was calculated using the Fraser-Swinney algorithm when $N_D = 65\,536$. Viewed at lag = 18, the curves from the x, y, and z variables have the top-down order of x–z–y.

null hypothesis of statistical independence. The code implementing their algorithm distributed by Fraser and Swinney departs from the partition termination criterion outlined in the text of their paper. In their code, the probe for structure is conducted at only one sublevel and the partitioning process is terminated if $\chi_3^2 < 1.547$. Fraser's restatement of the algorithm in binary representation and the generalization to embedded data are summarized in Appendix B.

Results obtained when our implementation of the Fraser-Swinney algorithm with a single-level partition termination criterion of $\chi_3^2 < 1.547$ was applied to the Rössler data of Fig. 3 are shown in Fig. 9. In our implementation, as in the case of the Fraser-Swinney code, the length of data sets X and Y must be a power of 2. Visual comparison of the results obtained with the Fraser-Swinney algorithm and $N_{data} = 65\,536$ (Fig. 9) with the results obtained with the algorithm of Sec. IV and $N_{data} = 100\,000$ suggests that similar results were obtained. This point is emphasized in Fig. 10 which shows that superposition of the results obtained when $N_{data} = 65\,536$ for both algorithms. The values of lag corresponding to the first minimum of the mutual information versus lag function obtained with the two algorithms are either equal or differ by 1. The average difference in the value of mutual information is less than 4%.

We now have two candidate procedures for calculating $I(X,Y)$, the Fraser-Swinney algorithm and the globally adaptive partition algorithm presented in Sec. IV. A procedure for comparing the two methods is constructed in the next section.

VI. COMPARING ALGORITHMS

In the previous sections, two procedures for computing mutual information were presented. They are compared in this section. Two properties, accuracy and speed, are examined. A comparison of accuracy requires example cases where the true value of mutual information is known to a high accuracy. This can be provided by jointly Gaussian data sets. Two data sets are said to be jointly Gaussian if their joint probability density function centered at (m_x, m_y) has the form

$$P_{XY}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y(1-r^2)^{1/2}} \exp\left\{ \frac{-1}{2(1-r^2)} \left[\left(\frac{x-m_x}{\sigma_x} \right)^2 - 2r\left(\frac{x-m_x}{\sigma_x} \right)\left(\frac{y-m_y}{\sigma_y} \right) + \left(\frac{y-m_y}{\sigma_y} \right)^2 \right] \right\}.$$

m_x and σ_x are the mean and standard deviation of time series $\{X\}$. m_y and σ_y are defined analogously for $\{Y\}$, and r is the cross-correlation coefficient between $\{X\}$ and $\{Y\}$. For the case of jointly Gaussian data sets, the mutual information is analytically related to the correlation coefficient by

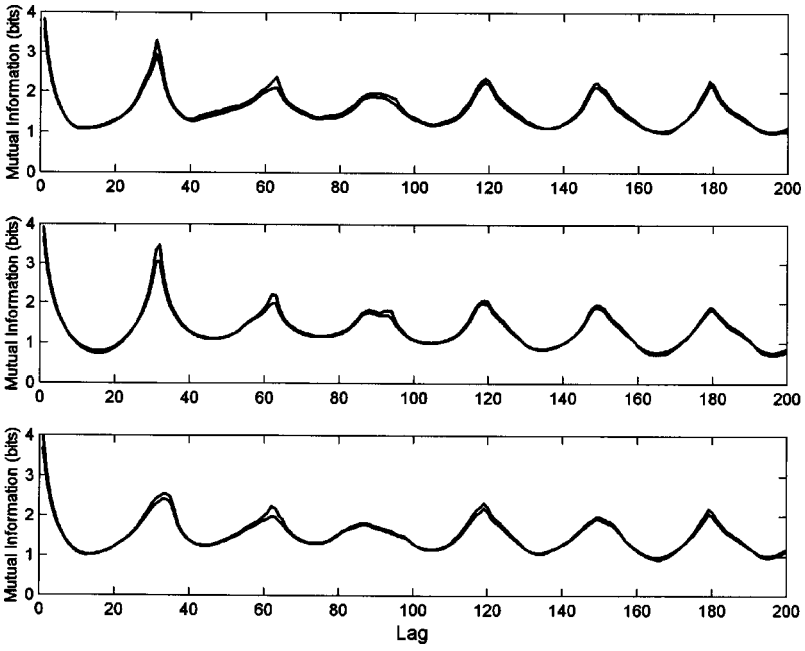

FIG. 10. Direct comparison of results obtained with the algorithm of Sec. IV and the Fraser-Swinney algorithm using Rössler data of Fig. 3. $N_D = 65\,536$. For those values of lag where the results of the two algorithms differ, the results of the algorithm of Sec. IV are below the results obtained with the Fraser-Swinney algorithm.

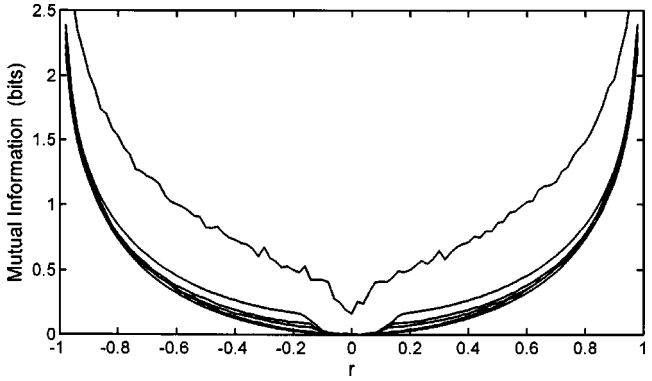

FIG. 11. Comparing the Fraser-Swinney algorithm, the algorithm of Sec. IV, and $-0.5\ln(1-r^2)$ for jointly distributed Gaussian data. Ninety-nine values of correlation r uniformly distributed on $(-1,1)$ were used. $N_D=8192$. For each value of r, 100 $\{X\}$, $\{Y\}$ data set pairs were generated. The algorithm's average value of mutual information is displayed. Viewed at $r=0.2$ the top-down ordering of the $I(X,Y)$ vs r functions is (i) the Fraser-Swinney algorithm with $\chi_3^2<1.547$, (ii) the algorithm of Sec. IV with $E_{XY}(i,j)\geqslant 5$, (iii) the algorithm of Sec. IV with $E_{XY}(i,j)\geqslant 10$, (iv) the algorithm of Sec. IV with $E_{XY}(i,j)\geqslant 15$, (v) the Fraser-Swinney algorithm with $\chi_3^2<5.000$, (vi) the analytical solution $-0.5\ln(1-r^2)$.

$$I(X,Y) = -0.5\ln(1-r^2).$$

A derivation of the relationship is given in Appendix A. The construction of a procedure for generating jointly Gaussian data sets with a specified correlation coefficient is also presented in that appendix.

Mutual information estimates obtained with the algorithm of Sec. IV and with the Fraser-Swinney algorithm are compared against $-0.5\ln(1-r^2)$ for the case of jointly distributed Gaussian data in Fig. 11. Ninety-nine values of r, uniformly distributed on $(-1,1)$ were used in these calculations. For each value of r, 100 jointly distributed $\{X\}$, $\{Y\}$ data set pairs of length 8192 were generated. The average value of mutual information for these pairs was determined using both algorithms. Multiple variants of each algorithm were used. The irregular $I(X,Y)$ vs r function seen in Fig. 11 was produced using the Fraser-Swinney algorithm when the subpartitioning process was terminated with the criterion $\chi_3^2<1.547$. With this criterion, an element of the partition is subdivided if the probability of nonuniform substructure is greater than 27%. This is the criterion implemented in their code. Calculations were also performed using $\chi_3^2<5.000$. This criterion results in the subdivision of an element of the partition only if the probability of nonuniform substructure is at least 80%. In this case, the results were much closer to $-0.5\ln(1-r^2)$. Three variants of the algorithm constructed in Sec. IV were used. In the first instance, the number of elements in the partition were chosen so that $E_{XY}(i,j)\geqslant 5$ for all elements. Recall that $E_{XY}(i,j)$ is the expected occupancy in partition element (i,j). Calculations also were performed with the Sec. IV algorithm with $E_{XY}(i,j)\geqslant 10$ and with $E_{XY}(i,j)\geqslant 15$. In the case of the Sec. IV algorithm, the value $I(X,Y)=0$ is returned whenever the null hypothesis of statistical independence is not rejected with a confidence level of at least 95%. This convention accounts for the transition to $I(X,Y)=0$ in the vicinity of $r=0$ for $I(X,Y)$ functions obtained with this algorithm. Viewed at $r=0.2$ the top-down ordering of the $I(X,Y)$ vs r functions is (i) the Fraser-Swinney algorithm with $\chi_3^2<1.547$, (ii) the algorithm of Sec. IV with $E_{XY}(i,j)\geqslant 5$, (iii) the algorithm of Sec. IV with $E_{XY}(i,j)\geqslant 10$, (iv) the algorithm of Sec. IV with $E_{XY}(i,j)\geqslant 15$, (v) the Fraser-Swinney algorithm with $\chi_3^2<5.000$, (vi) the analytical solution $-0.5\ln(1-r^2)$. The greatest numerical value of $I(X,Y)$ is obtained with the Fraser-Swinney algorithm with a subdivision criterion of $\chi^2<1.547$. This produces the greatest value of $I(X,Y)$ because the comparatively tolerant criterion of 27% introduces a numerical indication of small scale structure in the data (and hence a greater value of mutual information) that may not be present. With the more demanding criterion of $\chi^2<5.000$, a subdivision is introduced only if there is at least an 80% probability of nonuniform substructure. With this criterion there is less divergence between the algorithm-estimated value of mutual information and the analytically computed value of $-0.5\ln(1-r^2)$.

TABLE II. Average normalized error in the estimation of mutual information.

Algorithm	Error
Algorithm of Sec. IV $E_{XY}(i,j)\geqslant 5$	1.91×10^{-3}
Algorithm of Sec. IV $E_{XY}(i,j)\geqslant 10$	1.55×10^{-3}
Algorithm of Sec. IV $E_{XY}(i,j)\geqslant 15$	3.15×10^{-13}
Fraser-Swinney algorithm $\chi_3^2<1.547$	2.48×10^{-1}
Fraser-Swinney algorithm $\chi_3^2<5.000$	0.97×10^{-3}

Following Hamilton [24], the following error measure was calculated:

$$\text{error} = \frac{\sum_{i=1}^{99} (I(X,Y)^{\text{analytical}} - I(X,Y)^{\text{algorithm}})^2}{\sum_{i=1}^{99} (I(X,Y)^{\text{analytical}})^2},$$

where $I(X,Y)^{\text{analytical}}$ denotes the value obtained using $-0.5\ln(1-r^2)$. The results are shown in Table II. It is seen that the magnitude of the error is low with both algorithms.

In addition to providing an explicit assessment of the probability of the null hypothesis of statistical independence, the algorithm of Sec. IV offers an additional advantage over the Fraser-Swinney algorithm. It is much faster. Comparison of computation times with data sets of different lengths is given in Table III. Both programs were run in MATLAB 6.5.0 (R13) on a Pentium 4 processor running at 2.53 GHz. The computation times of the algorithm of Sec. IV are typically on the order of 0.5% of the times required by the Fraser-Swinney algorithm. In addition to being more accurate than the $\chi_3^2<1.547$ criterion, the $\chi_3^2<5.000$ algorithm is faster because it introduces fewer subdivisions.

An approximate understanding of the sensitivity of the two algorithms to data set size can be obtained by examining the results presented in Fig. 12. That diagram shows the mu-

TABLE III. Comparative computation times for different algorithms.

N_{data}	Time algorithm of Sec. IV (sec)	Time Fraser-Swinney algorithm $\chi_3^2=1.547$ (sec)	Time Fraser-Swinney algorithm $\chi_3^2=5.00$ (sec)
4096	1.3	266.2	185.2
8192	2.7	544.0	392.4
16384	5.0	1169.5	851.0
32768	9.3	2549.5	1898.5
65536	24.1	5940.5	4533.5

tual information versus lag functions obtained from a single data set generated by the Rössler equations (x variable data). As already seen in Fig. 9, the results obtained when $N_D=65\,536$ are almost identical. More substantive differences are observed, however, when smaller data sets are used. When N_D is 4096 and 8192, the algorithm of Sec. IV produces output that is slightly less than, but largely parallel to, the results obtained when $N_D=65\,536$. For this algorithm, the value of lag giving the first minimum of mutual information was the same for all values of N_D tested. In contrast, when $N_D=4096$ and 8192, the Fraser-Swinney algorithm produces mutual information versus lag functions that present structures that are lost when more data are incorporated into the computations. In some instances, these structures can alter the identification of the lag giving the minimum value of mutual information.

VII. DISCUSSION

The Fraser-Swinney algorithm with the $\chi_3^2<5.000$ criterion outperforms that algorithm when $\chi_3^2<1.547$ is used both in terms of accuracy (Table II) and speed (Table III). A comparison of the Fraser-Swinney algorithm with the $\chi_3^2<5.000$ criterion against the algorithm of Sec. IV leads to the following conclusions. First, the algorithm of Sec. IV has a significant advantage over the Fraser-Swinney algorithm in providing a global test of the statistical independence null hypothesis. The Fraser-Swinney algorithm uses a χ^2 test locally to implement the partitioning protocol. It does not, however, return an assessment of the statistical independence of X and Y. Second, while the Fraser-Swinney algorithm is more accurate with data sets where $N_D=8192$ (Table II), the results of Fig. 12 suggest that the Fraser-Swinney algorithm requires large data sets even when the $\chi_3^2<5.000$ criterion is used. When smaller data sets are used the Fraser-Swinney algorithm presents structures that disappear when more data becomes available. If the object of the calculation is to use $I(x_i,x_{i+lag})$ functions to find the appropriate lag for embedding, then these local minima could give misleading results. Third, the algorithm of Sec. IV requires about 0.5% of the calculation time required by the Fraser-Swinney algorithm.

Limitations of this study should be noted. Additional algorithms could be considered. Following Silverman [25], Moon *et al.* [26] have used kernel density estimators to calculate probability densities. They argue that the resulting algorithm outperforms the Fraser-Swinney algorithm. Moon *et al.* also suggest that their algorithm can be improved by using *K-d* trees to partition the data. Caution must be exercised when evaluating this suggestion. Our exploratory calculations have shown that *K-d* tree partitions can be very sensitive to initial conditions. This sensitivity is addressed by Bradley and Fayyad [27] who published a procedure for

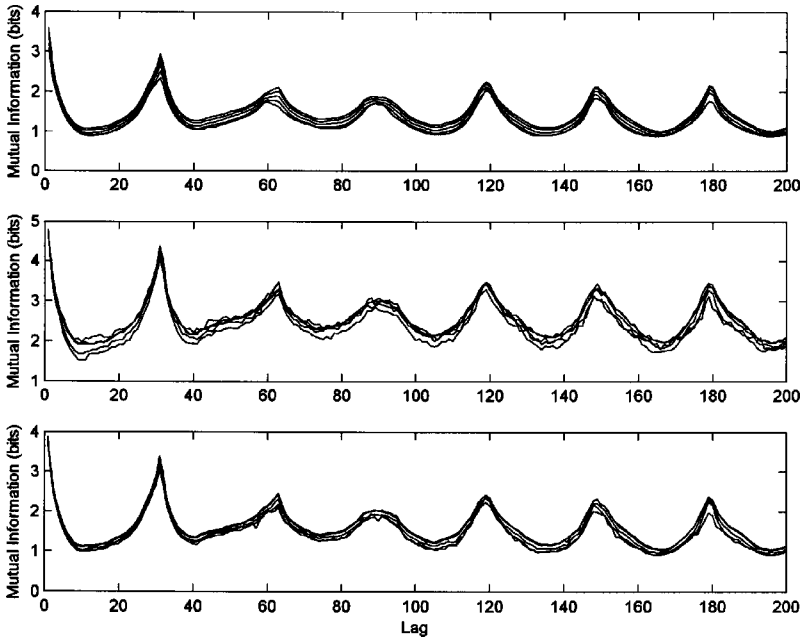

FIG. 12. Mutual information versus lag for data sets of different sizes. Mutual information versus lag was computed using both algorithms for $N_D=4096$, 8192, 16 384, 32 768, and 65 536. The data were generated by the Rössler equations, and x-variable output was used in the calculations. Functions calculated with $N_D=65\,536$ are at the top of each set of curves. Functions calculated with $N_D=4096$ are at the bottom of each set of curves. The top set of curves was calculated using the algorithm of Sec. IV. The middle set of results was calculated using the Fraser-Swinney algorithm with $\chi^2=1.547$. The results in the lowest panel were calculated with the Fraser-Swinney algorithm and $\chi^2=5.000$.

computing initial conditions based on a procedure for estimating the modes of a distribution.

Instead of partitioning phase space as is done in the algorithms discussed above, Pawelzik and Schuster [28] used the first order correlation integral to calculate probability densities and entropies. These entropies are then used to calculate mutual information. We consider here application of the technique to embedded time series data, $X_k=(x_k,x_{k+lag},x_{k+2lag},\ldots x_{k+(m-1)lag})$ and $Y_k=(y_k,y_{k+lag},y_{k+2lag},\ldots y_{k+(m-1)lag})k=1,\ \ldots,\ N-m+1$. Application to scalar data is trivially obtained by taking the embedding dimension, m, to be one for X and Y, and thus dimension 2 for the joint space. The density of X in the neighborhood of X_k is approximated by the first order correlation integral,

$$p_{X_k}(r)=\frac{1}{N_V-1}\sum_{j\neq k}\Theta(r-|X_j-X_k|),$$

where Θ is the Heaviside function, N_V is the number of embedding vectors, and r is the neighborhood size being considered. This density differs from that used earlier because it counts the number of points in possibly overlapping neighborhoods. The densities used in the algorithms discussed earlier involved nonoverlapping neighborhoods created by the partitioning process. This leads to a slightly different expression for the entropy which, in this case, is given by

$$H(X,r)=-\frac{1}{N_V}\sum_{k=1}^{N}\ln p_{X_k}(r).$$

In some implementations, finite sample corrections due to Grassberger [29] are included. The entropies of the Y data as well as the joint entropy are calculated similarly, and these are used to obtain the mutual information from the relation $I(X,Y)=H(X)+H(Y)-H(X,Y)$.

Quian Quiroga *et al.* [30] used the Pawelzik-Schuster algorithm with the Grassberger corrections in a study of synchronization of rat electrocorticograms (ECoG). They studied three multichannel ECoG records in a rat model of genetic absence epilepsy and compared activity between left and right hemispheres. They concluded that except for mutual information their linear and nonlinear measures provided qualitatively similar results. The authors felt that the small number of data points ($N=1000$) was responsible for the failure of mutual information to provide robust estimates of interhemispheric synchronization. These data were reanalyzed by Duckrow and Albano [31] using a modified Fraser-Swinney algorithm. The data were embedded and interleaved as described in Appendix B and the resulting binary representations were used as inputs in the Fraser-Swinney algorithm. Using embedding dimensions from 1 to 10 and Lags from 1 to 30, the results consistently showed the ranking that Quian Quiroga *et al.* found using other measures of synchronization. Results obtained by Duckrow and Albano using these data and a uniform partition algorithm showed a behavior similar to that found by Quian Quiroga when they used the Pawelzik-Schuster algorithm.

Yet another approach to calculating mutual information has been published by Kilminster *et al.* [32] who have shown that the Radon transform can be used to estimate joint probability density functions which can then be used to estimate mutual information. They argue that, in contrast with standard methods, this procedure preserves fractal structure. Since completing this manuscript, our attention has been directed to a valuable paper by Kraskov, Stögbauer, and Grassberger [33] on estimating mutual information. The Kilminster *et al.*, Moon *et al.*, and Kraskov *et al.* algorithms could be compared against the Fraser-Swinney algorithm and the algorithm of Sec. IV in an expanded study.

ACKNOWLEDGMENTS

We would like to thank E. Kurali for directing us to information concerning jointly Gaussian data sets. The comments and discussions with A. I. Mees and T. I. Schmah are acknowledged with gratitude. Particular thanks are directed to CDR R. S. Hernandez, USN for encouragement, support and project leadership. This research was supported by Grant No. 601135N.4508.518.A0247 from the Office of Naval Research and the Navy Bureau of Medicine to the Naval Medical Research Center. The opinions and assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the Navy Department or the naval service at large.

APPENDIX A: JOINTLY GAUSSIAN DATA SETS AND THE MUTUAL INFORMATION OF JOINTLY GAUSSIAN DATA SET PAIRS

We construct here a procedure for generating jointly Gaussian data sets $\{Y^1\}$ and $\{Y^2\}$ from two independent Gaussian data sets $\{X^1\}$ and $\{X^2\}$. This is followed by a demonstration showing that the mutual information of two jointly Gaussian data sets with a cross-correlation coefficient r is $-0.5\ln(1-r^2)$.

For simplicity of presentation we consider the special case of data sets that have zero mean and equal variance. The procedure can be extended to the more general case. Let $\{X^1\}=(x_1^1,x_2^1,x_3^1,\ldots x_N^1)$ and $\{X^2\}=(x_1^2,x_2^2,x_3^2,\ldots x_N^2)$ be Gaussian distributed with zero mean and the same variance σ^2. It is further assumed that they are uncorrelated, that is, their cross-correlation coefficient r is equal to zero. Given the assumption of zero correlation, their joint probability distribution is the product of their individual probability distributions,

$$P_{X^1X^2}(x^1,x^2)=\frac{1}{2\pi\sigma^2}\exp\{-[(x^1)^2+(x^2)^2]/2\sigma^2\}$$
$$=\frac{1}{2\pi|\Sigma_x|^{1/2}}\exp\left\{-x^T\sum_X^{-1}x/2\right\},$$

where Σ_x is the (X^1,X^2) covariance matrix,

$$\sum_{x} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}.$$

Two data sets $\{Y^1\}=(y_1^1,y_2^1,y_3^1,\ldots y_N^1)$ and $\{Y^2\}=(y_1^2,y_2^2,y_3^2,\ldots y_N^2)$ with zero means, equal variance σ^2, and cross-correlation r are jointly Gaussian if their joint probability density function is

$$P_{Y^1Y^2}(y^1,y^2) = \frac{1}{2\pi|\Sigma_y|^{1/2}}\exp\{-y^T\Sigma_y^{-1}y/2\}.$$

Σ_y is the (Y^1,Y^2) covariance matrix,

$$\sum_{Y} = \sigma^2\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}(1-r^2)^{1/2}, \quad \sum_{Y}^{-1} = \frac{1}{(1-r^2)\sigma^2}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix},$$

$$\left|\sum_{Y}\right|^{1/2} = \sigma^2(1-r^2)^{1/2}. \tag{A1}$$

Matrix A is a two-dimensional linear transformation relating $\{X^1\}$ and $\{X^2\}$, independent Gaussian random variables, to $\{Y^1\}$ and $\{Y^2\}$, jointly distributed Gaussian variables,

$$\begin{pmatrix} x_j^1 \\ x_j^2 \end{pmatrix} = A\begin{pmatrix} y_j^1 \\ y_j^2 \end{pmatrix}.$$

Let A be given by

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Using this representation for A, the relationship, $x=Ay$, and the expression for Σ_Y^{-1} above makes it possible to solve for b, c, and d in terms of a and r. There are an infinity of A's that depend on the choice of a. We use here the simplest case, $a=1$,

$$A = \begin{pmatrix} 1 & 0 \\ \dfrac{r}{\sqrt{1-r^2}} & \dfrac{-1}{\sqrt{1-r^2}} \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 1 & 0 \\ r & -\sqrt{1-r^2} \end{pmatrix}.$$

In the next step, we need to establish the relationship cited in the text between mutual information $I(Y^1,Y^2)$ and r, the cross-correlation coefficient. In this derivation, we use the property that $\{Y^1\}$ and $\{Y^2\}$ are jointly distributed, have correlation r, and are related to independent Gaussian data sets $\{X^1\}$ and $\{X^2\}$ by linear transformation A. The derivation begins with the integral representation for mutual information expressed in terms of the joint and individual probability density functions. The integrals are taken from $-\infty$ to $+\infty$,

$$I(Y^1,Y^2) = \iint P_{Y^1Y^2}(y^1,y^2)\ln\left\{\frac{P_{Y^1Y^2}(y^1,y^2)}{P_{Y^1}(y^1)P_{Y^2}(y^2)}\right\}dy^1dy^2.$$

By construction, Y^1 and Y^2 are jointly Gaussian with equal variances. Y^1 and Y^2 are Gaussian distributed, giving the following expression for mutual information:

$$I(Y^1,Y^2) = \iint \frac{e^{-y^T\Sigma_Y^{-1}y/2}}{2\pi|\Sigma_Y|^{1/2}}\ln\left\{\frac{\dfrac{e^{-y^T\Sigma_Y^{-1}y/2}}{2\pi|\Sigma_Y|^{1/2}}}{\dfrac{e^{-(y^1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\dfrac{e^{-(y^2)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}}\right\}dy^1dy^2.$$

Given the previously stated expression for $|\Sigma_Y|^{1/2}$, and the relationship between x and y, we can transform this into integrals over x^1 and x^2:

$$I(Y^1,Y^2) = \iint \frac{e^{-x^Tx/2\sigma^2}}{2\pi\sigma^2}\ln\left\{\frac{e^{-x^Tx/2\sigma^2}}{e^{-(y^1)^2/2\sigma^2}e^{-(y^2)^2/2\sigma^2}(1-r^2)^{1/2}}\right\}dx^1dx^2$$

which can be simplified to

$$I(Y^1,Y^2) = \iint \frac{e^{x^Tx/2\sigma^2}}{2\pi\sigma^2}\left\{\frac{1}{2\sigma^2}[r^2(x^1)^2 - r^2(x^2)^2 - 2r\sqrt{1-r^2}x^1x^2] - \ln\sqrt{1-r^2}\right\}dx^1dx^2.$$

Consider the integral

$$\iint \frac{e^{-x^Tx/2\sigma^2}}{2\pi\sigma^2}\left\{\frac{1}{2\sigma^2}[r^2(x^1)^2 - r^2(x^2)^2]\right\}dx^1dx^2.$$

The two terms are of equal magnitude and opposite sign, and the double integral is therefore equal to zero. Similarly consider

$$\iint \frac{e^{-x^Tx/2\sigma^2}}{2\pi\sigma^2}\left\{\frac{1}{2\sigma^2}(-2r\sqrt{1-r^2}x^1x^2)\right\}dx^1dx^2.$$

Each integral is of an odd function over the range $-\infty$–$+\infty$ and is therefore equal to zero. The integral for mutual information simplifies to

$$I(Y^1,Y^2) = -\iint \frac{e^{-x^Tx/2\sigma^2}}{2\pi\sigma^2}\{\ln\sqrt{1-r^2}\}dx^1dx^2.$$

Using

$$\int_{-\infty}^{+\infty} e^{-z^2/2\sigma^2} = (2\pi)^{1/2}\sigma$$

gives

$$I(Y^1,Y^2) = -\ln\sqrt{1-r^2} = -\frac{1}{2}\ln(1-r^2).$$

APPENDIX B: BINARY REPRESENTATION OF *XY* PARTITIONING AND GENERALIZATION TO EMBEDDED DATA

Section V discussed the local adaptive partitioning used by Fraser and Swinney to calculate mutual information. The space being partitioned is that of the joint distribution of $X=\{x_1,x_2,\ldots x_N\}$ and $Y=\{y_1,y_2,\ldots y_N\}$, a subset of the XY plane which may be considered a two-dimensional embed-

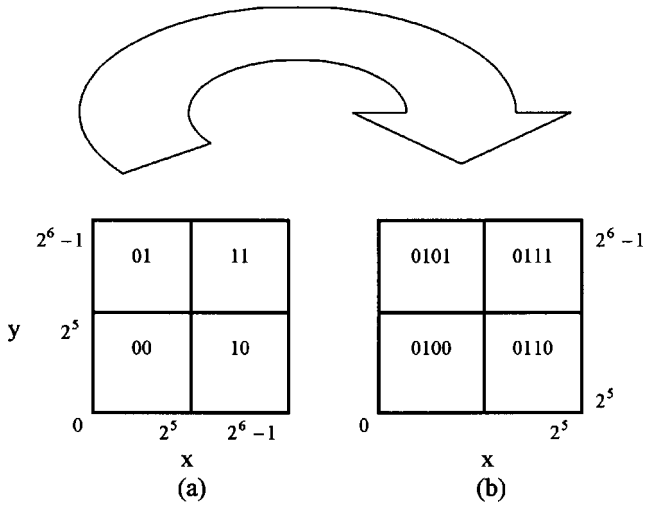

FIG. 13. (a) Partition of $0 \leqslant x, y \leqslant 2^6-1$ into four quadrants. (b) Partition of quadrant 01 (upper left) into four subquadrants.

ding space whose elements are $(x_i, y_i), i=1,2,\ldots N$. The following steps are used to implement the procedure:

1. Let the number of elements of both X and Y be $N=2^n$ (the binary logic of the algorithm requires $N=2^n$).

2. Rank order both X and Y with no repeated elements so that they both map to permutations of the integers 0, 1, …, 2^n-1. To avoid repeated elements, one may assign higher ranks to numbers appearing earlier in the series. Call these rank-ordered lists $X^R=\{X_1^R, x_2^R, \ldots x_N^R\}$ and $Y^R=\{y_1^R, y_2^R, \ldots y_N^R\}$. X^R and Y^R are equiprobable.

3. Transform the elements of X^R to binary. Since the $0 \leqslant x_k^R \leqslant 2^n-1$, these binary representations have at most n bits, i.e., $x_k^R = a_k^{n-1} a_k^{n-2} \ldots a_k^0$. Here, a_k^{n-1} is the most significant bit of x_k^R, a_k^{n-2} the second most significant, etc. Perform the same transformation on the elements of Y^R to get $y_k^R = b_k^{n-1} b_k^{n-2} \cdots b_k^0$.

4. Interleave the bits of x_k^R and y_k^R to get

$$z_k^R = (a_k^{n-1} b_k^{n-1} a_k^{n-2} b_k^{n-1} \cdots a_k^0 b_k^0). \tag{B1}$$

The two left-most elements of z_k^R are the most significant bits of x_k^R and y_k^R, respectively, the next two are the next most significant bits, etc. For example, suppose $(x_k^R, y_k^R) = (5, 47)$. Then, using the binary representations, $5=000101$ and $47=101111$, the interleaved representation of (x_k^R, y_k^R) is

$$(x_k^R, y_k^R) \Rightarrow z_k^R = (010001110111).$$

A crucial advantage of this representation derives from the observation that the successive bit pairs provide a tree representation for the location of (x_k^R, y_k^R) in the two-dimensional embedding space. To see this, label the axes of a two-dimensional embedding space by x and y and consider the region $0 \leqslant x, y \leqslant 2^5-1$. If this region is subdivided into four quadrants as in Fig. 13(a), then the bottom-left quadrant contains all those vectors with six-bit y's whose most significant bits are 0 and with y's whose most significant bits are also zero, the bottom-right quadrant contains all those x's whose most significant bits are 1 and those y's whose most significant bits are 0, etc. The location of any interleaved point in this subdivision is thus labeled by its first two elements; the (x_k^R, y_k^R) in our example is in quadrant 01. If this quadrant is again subdivided into four, the next two bits of z_k^R specify its location in the new subdivision [Fig. 13(b)], and so on.

The technique of interleaving may also be used to implement time-delay embedding. Consider the m-dimensional embedding of X with a specified lag

$$X = (x_k, x_{k+\mathrm{lag}}, x_{k+2\mathrm{lag}}, \ldots x_{k+(m-1)\mathrm{lag}}).$$

Using the notation of Eq. (1), the m-dimensional embedding vector X_k may be represented as

$$X_k \rightarrow u_k = (a_k^{n-1} a_{k+\mathrm{lag}}^{n-1} \cdots a_{k+(k-1)\mathrm{lag}}^{n-1}) \times (a_k^{n-2} a_{k+\mathrm{lag}}^{n-2} \cdots a_{k+(k-1)\mathrm{lag}}^{n-2}) \cdots (a_k^0 a_{k+\mathrm{lag}}^0 \cdots a_{k+(m-1)\mathrm{lag}}^0), \tag{B2}$$

a number that uniquely represents X_k. A similar embedding and interleaving of Y gives

$$Y = (y_k, y_{k+\mathrm{lag}}, y_{k+2\ \mathrm{Lg}}, \ldots y_{k+(m-1)\mathrm{lag}})$$

and

$$Y_k \rightarrow v_k = (b_k^{n-1} b_{k+\mathrm{lag}}^{n-1} \cdots b_{k+(k-1)\mathrm{lag}}^{n-1}) \times (b_k^{n-2} b_{k+\mathrm{lag}}^{n-2} \cdots b_{k+(k-1)\mathrm{lag}}^{n-2}) \cdots (b_k^0 b_{k+\mathrm{lag}}^0 \cdots b_{k+(m-1)\mathrm{lag}}^0)$$

The interleaved sets, $\{u_k\}$ and $\{v_k\}$, each consists of 2^n numbers, each number specified by $n \times m$ bits. To calculate the mutual information of X and Y, $\{u_k\}$ and $\{v_k\}$ are converted to decimal and used as inputs in either the Fraser-Swinney algorithm or the algorithm of Sec. IV.

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[2] N. J. I Mars and F. H. Lopes da Silva, in *Methods of Analysis of Brain Electrical and Magnetic Signals. EEG Handbook*, Revised Series, edited by A. S. Gevins and A. Rémond (Elsevier Science Publishers, Amsterdam, 1987), Vol. I, pp. 297–307.
[3] P. E. Rapp, A. M. Albano, T. I. Schmah, and L. A. Farwell, Phys. Rev. E **47**, 2289 (1993).
[4] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[5] C. J. Cellucci, A. M. Albano, and P. E. Rapp, Phys. Rev. E **67**, 066210 (2003).
[6] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996).
[7] C. W. J. Granger, Econometrica, **37**, 424 (1969).
[8] M. J. Kaminski, M. Ding, W. A. Truccolo, and S. L. Bressler, Biol. Cybern. **85**, 145 (2001).
[9] J. Xu, Z.-R. Liu, R. Liu, and Q.-F. Yang, Physica D **106**, 363

(1997).
[10] T. Inouye, K. Shinosaki, and A. Yagasaki, Electroencephalogr. Clin. Neurophysiol. **55**, 290 (1983).
[11] T. Inouye, K. Shinosaki, A. Iyama, and Y. Matsumoto, Electroencephalogr. Clin. Neurophysiol. **86**, 224 (1993).
[12] F. H. Lopes da Silva, J. P. Pijn, and P. Boeijinga, Brain Topogr **2**, 9 (1989).
[13] F. Chen, J. Xu, F. Gu, X. Yu, X. Meng, and Z. Qiu, Biol. Cybern. **83**, 355 (2000).
[14] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
[15] J. B. Bendat and A. G. Piersol, *Measurement and Analysis of Random Data* (John Wiley, New York, 1966), p. 284.
[16] J. H. Cocatre-Zilgien and F. Delcomyn, J. Neurosci. Methods **41**, 19 (1992).
[17] F. Mosteller and J. W. Tukey, *Data Analysis and Regression* (Addison-Wesley, Reading, MA, 1977), p. 49.
[18] Y. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1992), p. 76.
[19] J. P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[20] T. Sauer, J. A. Yorke and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[21] W. G. Cochran, Biometrics **10**, 417 (1954).
[22] L. Ott, M. T. Longnecker, and R. L. Ott, *An Introduction to Statistical Methods and Data Analysis* (Wadsworth, New York, 1998).
[23] A. M. Fraser, IEEE Trans. Inf. Theory **35**, 245 (1989).
[24] J. D. Hamilton, *Time Series Analysis* (Princeton University Press, Princeton, NJ, 1964).
[25] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York, 1986).
[26] Y.-L. Moon, R. Rajagopalan, and U. Lall, Phys. Rev. E **52**, 2318 (1995).
[27] P. S. Bradley and U. M. Fayyad, in *Proceedings of the Fifteenth Conference on Machine Learning*, edited by I. Brasko and S. Dzeroski (Morgan Kaufman, New York, 1998).
[28] K. Pawelzik and H. G. Schuster, Phys. Rev. A **35**, 481 (1987).
[29] P. Grassberger, Phys. Lett. A **128**, 369 (1988).
[30] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, Phys. Rev. E **65**, 041903 (2002).
[31] R. B. Duckrow and A. M. Albano, Phys. Rev. E **67**, 063901 (2003).
[32] D. Kilminster, D. Allingham, and A. Mees, Ann. Inst. Stat. Math. **54**, 224 (2002).
[33] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).