Journal of Statistical Physics, Vol. 73, Nos. 3/4, 1993

Measuring Statistical Dependences in a Time Series

Bernd Pompe[1]

Received December 14, 1992

We propose two methods to measure all (linear and nonlinear) statistical dependences in a stationary time series. Presuming ergodicity, the measures can be obtained from efficient numerical algorithms.

KEY WORDS: Entropy; time series analysis; mutual information; contingency; chaotic map; quadratic map.

1. INTRODUCTION

The measurement of statistical dependences is one of the fundamental problems in time series analysis. For example, given a finite data sequence, there is reason to look for a predictor if there are statistical dependences between past and future states. In data compressing coding systems statistical dependences between letters are used to reduce the bit rate (see, e.g., refs. 2, 13, and 22).

There are several quantities and algorithms to measure statistical dependences. All of them have their advantages and limitations. In Section 2 we give a short review of some well-known "classical" methods which facilitates an evaluation of our procedures.

In Section 3 we propose our first method. It is characterized by transforming the time series to its relative rank numbers, yielding a transformed series which is uniformly distributed with values in the interval $]0, 1]$, and then estimating a quantity called generalized mutual information (GMI). It is defined on the base of Rényi's generalized entropy of second order.[4,17] The transformation to (relative) rank numbers is crucial for our approach, because in this case we can guarantee that the GMI is nonnegative and equals zero if and only if there are no (linear or nonlinear) statistical

[1] Fachbereich Physik, Ernst-Moritz-Arndt-Universität, 17487 Greifswald, Germany.

0022-4715/93/1100-0587$07.00/0 © 1993 Plenum Publishing Corporation

dependences. If the one-dimensional distribution is not uniform, then we cannot generally conclude from a vanishing GMI to statistical independence. On the other hand, the GMI indicates determinism in the time series on a given coarse-grained level.

Shannon's information measure would always, i.e., for an arbitrary one-dimensional distribution of the series, provide a nonnegative mutual information, which equals zero iff there are no statistical dependences, and thus no transformation would be required. However, we prefer to use Rényi's information measure because it can more easily be estimated using the Grassberger–Procaccia–Takens algorithm (GPTA), which is well known from calculating the correlation dimension of a fractal measure.[9,20]

In contrast to the first method, our second method (see Section 4) directly works also for nonuniform one-dimensional distributions of the data. Hence no transformation of the original data is required. Moreover, the method inquires into statistical dependences between a past vector and a *future vector* of states, which is more general than in our first method, where we have only a scalar future state. The quantity estimated is called "mutual account." It could be considered as a kind of contingency which is properly standardized. The mutual account can be estimated from the data sequence by the GPTA and a modified GPTA. In the latter case we have to count a triple of $(D+d)$-, D-, and d-dimensional points, originating from an embedding of the data with delay coordinates, such that the distance of the second point to a D-dimensional projection of the first point and that of the third point to a d-dimensional projection of the first point are each less than a certain threshold. This algorithm works, like the GPTA, for small values of the threshold. The usefulness of our methods is illustrated for several simple examples, including white noise and chaotic 1D maps.

2. QUANTITIES MEASURING STATISTICAL DEPENDENCES

Preliminaries. Consider a *stationary* time series $\{X_t\}$, where X_t is a real-valued random variable representing the observation x_t made at time t. The series is assumed to be *time-discrete*, and X_t should be, as a rule, *discrete* as well: $x_t \in \{x(1), x(2),..., x(k)\}$. This is the typical case in practice, where a continuous signal is sampled with a certain sampling period and recorded using an analog digital converter with k quantization levels.

Let p_m denote the probability that $X_t = x(m)$. For a given time lag τ, X_t and $X_{t+\tau}$ are said to be *statistically independent* if

$$p_{mn}(\tau) = p_m p_n \qquad \text{for all} \quad m, n = 1, 2,..., k \tag{1}$$

with the joint probability $p_{mn}(\tau)=\text{prob}\{X_t=x(m),\ X_{t+\tau}=x(n)\}$. If (1) fails, then we have statistical dependences, and in this paper we propose methods to measure them. But first of all we briefly report on some well-known measures to motivate our approach.

Correlation Function. The most common measure of statistical dependences in a stationary time series $\{X_t\}$ is the *autocorrelation function* (ACF)

$$\text{cor}(X_t, X_{t+\tau})=\frac{\langle X_t X_{t+\tau}\rangle-\langle X\rangle^2}{\langle X^2\rangle-\langle X\rangle^2} \tag{2}$$

where $\langle X_t\rangle=\langle X_{t+\tau}\rangle=\langle X\rangle$ is the mean value of X_t. Of course, in the stationary case cor is a function of τ only. The ACF can be estimated directly from the data $\{x_t\}$ without any partitioning (quantization, coarse-graining, classification), which is advantageous compared to some other methods described below. Moreover, there are fast algorithms for estimating the ACF of a finite series $\{X_t\}_{t=1}^{T}$. For instance, the FFT (fast Fourier transformation) method yields $\text{cor}(X_t, X_{t+\tau})$ for different τ from two FFTs and a squared magnitude (see, e.g., ref. 16). However, the ACF measures only *linear* dependences, which are the only ones for very special distributions of $(X_t, X_{t+\tau})$, e.g., jointly normal. To learn anything about all (including nonlinear) statistical dependences, in general, the correlations of higher moments $\text{cor}(X_t^m, X_{t+\tau}^n)$ have to be considered as well. There are no statistical dependences iff these correlations vanish for all $m, n=1, 2,\ldots, k-1$.[17] Testing this for large k would be rather time consuming.

Mutual Information. Statistical dependence can be measured by the mutual information

$$I(\tau)=H_1-(H_2(\tau)-H_1) \tag{3}$$

with the Shannon entropies

$$H_1=-\sum_{n=1}^{k} p_n \log p_n \qquad \text{and} \qquad H_2(\tau)=-\sum_{m,n=1}^{k} p_{mn}(\tau)\log p_{mn}(\tau) \tag{4}$$

$I(\tau)$ represents the amount of information on $X_{t+\tau}$ that is contained in X_t and vice versa. We always have

$$0\leqslant I(\tau)\leqslant H_1 \tag{5}$$

where $I(\tau)=0$ iff X_t and $X_{t+\tau}$ are statistically independent, and $I(\tau)=H_1$ iff $X_{t+\tau}$ follows unequivocally from X_t.[17] Moreover, it can be rewritten as

$$I(\tau)=\sum_{m,n=1}^{k} p_{mn}(\tau)\log\frac{p_{mn}(\tau)}{p_m p_n} \tag{6}$$

Hence, it can be considered as the information we gain if the joint probabilities $\{p_m p_n\}_{m,n}$ (hypothesis on independence) are replaced by $\{p_{mn}(\tau)\}_{m,n}$ (real joint distribution).

The difficulty in calculating mutual information from a finite time series $\{X_t\}_{t=1}^{T}$ is in estimating joint probabilities p_{mn} from histograms or by any less "naive" procedure (see, e.g., ref. 19). If k—the number of possible values of X_t—is large (say $k^2 \gtrsim T/10$), a coarse-graining of the plane spanned by x_t and $x_{t+\tau}$ has to be used to have good statistics. However, the partition must be fine enough to follow changes of $p_{mn}(\tau)$ and not to underestimate $I(\tau)$.

Fraser and Swinney[7] proposed a rather sophisticated algorithm that covers the $(x_t, x_{t+\tau})$ plane by a sequence of partitions $\{G_i\}$ such that each partition is a rectangular grid generated by dividing each axis into $k=2^i$ *equiprobable* segments, and G_{i+1} is a refinement of G_i yielding smaller partition elements of the boxes of G_i that are characterized by a substructure. This method could be improved by averaging over mutual information calculated from partitions which are carefully shifted in a certain region of state space to average the influence of the position of the partition on the mutual information.[14] However, this procedure would be, in general, rather time consuming. It is the aim of this paper to give an alternative to these proposals (see Section 3). Our ideas are based on another quantity, which is given as follows.

Mean Square Contingency. We could also take the *mean square contingency* (see, e.g., ref. 17)

$$\varphi^2(\tau) = \sum_{m,n=1}^{k} \frac{[p_{mn}(\tau) - p_m p_n]^2}{p_m p_n} \tag{7}$$

in order to measure statistical dependences. Obviously, it vanishes iff X_t and $X_{t+\tau}$ are statistically independent. Moreover, we always have

$$0 \leqslant \varphi^2(\tau) \leqslant k-1 \tag{8}$$

where the upper bound is attained iff $X_{t+\tau}$ is a function of X_t, which is the deterministic case. The principal problems in calculating the contingency are the same as in the case of mutual information. However, in the following sections we propose two methods which do not need a partition.

3. FIRST METHOD: GENERALIZED MUTUAL INFORMATION

Basic Theorem. Consider a stationary discrete time series $\{X_t\}$ such that X_t could be one of k different values $x(n)$, $n = 1, 2,..., k$. We now ask for statistical dependences between a D-dimensional vector of "past" states $\mathbf{X}_t = (X_{t-\Theta_{D-1}},..., X_{t-\Theta_0})$ and one "future" state $X_{t+\tau}$. Here $D = 1, 2, 3,...$; $\Theta_{D-1} > \cdots > \Theta_1 > \Theta_0 = 0$; and $\tau \geqslant 0$. Denote the joint probabilities

$$p_{m_{D-1},...,m_0,n}(\tau) = \mathrm{prob}\{X_{t-\Theta_{D-1}} = x(m_{D-1}),..., X_{t-\Theta_0} = x(m_0), X_{t+\tau} = x(n)\}$$

where $m_{D-1},..., m_0$, $n = 1, 2,..., k$. For short we set $p_{m_{D-1},...,m_0,n}(\tau) \equiv p_{\mathbf{m}n}(\tau)$. Moreover, we use the abbreviation $p_{\mathbf{m}} \equiv \sum_{n=1}^{k} p_{\mathbf{m}n}(\tau)$, and write $\sum_{m_{D-1},...,m_0=1}^{k}$ as $\sum_{\mathbf{m}=1}^{k}$.

We now define a contingency

$$\varphi_D^2(\tau) \equiv \sum_{\mathbf{m},n=1}^{k} \frac{[p_{\mathbf{m}n}(\tau) - p_{\mathbf{m}} p_n]^2}{p_{m_{D-1}} \cdots p_{m_0} p_n} \tag{9}$$

It can be rewritten as

$$\varphi_D^2(\tau) = \sum_{\mathbf{m},n=1}^{k} \frac{p_{\mathbf{m}n}^2(\tau)}{p_{m_{D-1}} \cdots p_{m_0} p_n} - \sum_{\mathbf{m}=1}^{k} \frac{p_{\mathbf{m}}^2}{p_{m_{D-1}} \cdots p_{m_0}} \tag{10}$$

Note that (9) is not the straightforward generalization of (7), because we use $\prod_{i=0}^{D-1} p_{m_i}$ instead of $p_{m_{D-1},...,m_0} = p_{\mathbf{m}}$ in the denominator, which is crucial for our approach. From the definition (9) we immediately see that $\varphi_D^2(\tau) \geqslant 0$, and $\varphi_D^2(\tau) = 0$ iff $\mathbf{X}_t$ and $X_{t+\tau}$ are independent [i.e., iff $p_{\mathbf{m}n}(\tau) = p_{\mathbf{m}} p_n$ for all $\mathbf{m}, n$]. For $D = 1$, Eq. (9) can be replaced by (7) because $\varphi_1^2(\tau) = \varphi^2(\tau)$.

Now we assume that X_t is *uniformly distributed*, i.e., $p_{m_{D-1}} = \cdots = p_{m_0} = p_n = 1/k = \varepsilon$. Hence (10) can be rewritten as

$$\varphi_D^2(\tau) = \varepsilon^{-(D+1)} \sum_{\mathbf{m},n=1}^{k} p_{\mathbf{m}n}^2(\tau) - \varepsilon^{-D} \sum_{\mathbf{m}=1}^{k} p_{\mathbf{m}}^2 \tag{11}$$

Consider now the quantity

$$I_D^{(2)}(\tau) \equiv H_1^{(2)} - (H_{D+1}^{(2)}(\tau) - H_D^{(2)}) \tag{12}$$

where the generalized Rényi entropies of second order (see, e.g., refs. 17 and 4) are involved,

$$H_1^{(2)} = -\log \sum_n p_n^2, \quad H_D^{(2)} = -\log \sum_{\mathbf{m}} p_{\mathbf{m}}^2, \quad H_{D+1}^{(2)}(\tau) = -\log \sum_{\mathbf{m},n} p_{\mathbf{m}n}^2(\tau) \tag{13}$$

Note that $H_1^{(2)} = -\log \varepsilon$ due to the proposed uniform one-dimensional distribution. We call $I_D^{(2)}(\tau)$ the *generalized mutual information* (GMI), in analogy to the mutual information (3). However, the interpretation of the GMI cannot be performed in total analogy to the mutual information.

The contingency (11) is related to the GMI,

$$I_D^{(2)}(\tau)=\log\left(\frac{\varphi_D^2(\tau)}{\varepsilon^{-D}\sum_{\mathbf{m}} p_{\mathbf{m}}^2}+1\right)=\log\frac{\sum_{\mathbf{m},n} p_{\mathbf{m}n}^2(\tau)}{\varepsilon\sum_{\mathbf{m}} p_{\mathbf{m}}^2} \tag{14}$$

From (14) we see that $I_D^{(2)}(\tau)\geqslant 0$ because $\varphi_D^2(\tau)\geqslant 0$, and from (9), (14) it follows that $I_D^{(2)}(\tau)=0$ iff $\mathbf{X}_t$ and $X_{t+\tau}$ are independent. Moreover, we have $H_{D+1}^{(2)}(\tau)-H_D^{(2)}\geqslant 0$, which we deduce from

$$\sum_{\mathbf{m}} p_{\mathbf{m}}^2=\sum_{\mathbf{m}}\left(\sum_{n} p_{\mathbf{m}n}(\tau)\right)^2\geqslant\sum_{\mathbf{m},n} p_{\mathbf{m}n}^2(\tau)$$

The equality holds iff $X_{t+\tau}$ follows unequivocally from $\mathbf{X}_t$, i.e., iff $p_{\mathbf{m}n}=p_{\mathbf{m}}$ or 0 for all $n=1, 2,..., k$. Obviously $I_D^{(2)}(0)=H_1^{(2)}$. Thus we have shown the following theorem:

Suppose a stationary discrete time series X_t with k *uniformly distributed* states, i.e., $\text{prob}\{X_t=x(n)\}=1/k$, $n=1, 2,..., k$. Then the generalized mutual information (12) between a D $(=1, 2,...)$-dimensional vector $\mathbf{X}_t=(X_{t-\Theta_{D-1}},..., X_{t-\Theta_0})$ and $X_{t+\tau}$ satisfies the relation

$$0\leqslant I_D^{(2)}(\tau)\leqslant H_1^{(2)}=\log k \tag{15}$$

for time lags $\Theta_{D-1}>\cdots>\Theta_1>\Theta_0=0$ and $\tau\geqslant 0$. Moreover, we have $I_D^{(2)}(\tau)=0$ if and only if $\mathbf{X}_t$ and $X_{t+\tau}$ are statistically independent, and $I_D^{(2)}(\tau)=\log k$ if and only if $X_{t+\tau}$ follows unequivocally from $\mathbf{X}_t$.

Remarks. It should be noted that $I_D^{(2)}(\tau)$ as defined in (12) and (13) might be negative if X_t is not uniformly distributed.[17] Moreover, if $\mathbf{X}_t$ and $X_{t+\tau}$ are statistically independent, then $I_D^{(2)}(\tau)=0$ for an arbitrary distribution of X_t. However, if X_t is not uniformly distributed, then from $I_D^{(2)}(\tau)=0$ we cannot deduce that $\mathbf{X}_t$ and $X_{t+\tau}$ are statistically independent, as the following example shows: Suppose that the random vector (X, Y) has the joint distribution $p_{11}=-1/2+2p-p^2$, $p_{22}=1/2-p^2$, and $p_{12}=p_{21}=1/2-p+p^2$ with $1-1/\sqrt{2}<p<1/\sqrt{2}$. Then X and Y have the same distribution $p_1=p_{11}+p_{12}=p_{11}+p_{21}=p$ and $p_2=p_{22}+p_{12}=p_{22}+p_{21}=1-p$. With these expressions we get

$$I_1^{(2)}=2H_1^{(2)}-H_2^{(2)}=-2\log(p_1^2+p_2^2)+\log(p_{11}^2+p_{12}^2+p_{21}^2+p_{22}^2)=0$$

This means that the GMI equals zero for all $p\in[1-1/\sqrt{2}, 1/\sqrt{2}]$. On the other hand, X and Y are statistically independent (i.e., $p_{mn}=p_m p_n$ for $m, n=1, 2$) only for $p=1/2$. Consequently, the GMI is, in general, *no* appropriate measure of statistical dependence. That is why we propose to transform a nonuniform series to a uniform one and then to estimate the GMI. How to perform the transformation and some of its consequences will be discussed later in this section. First let us comment on an algorithm to estimate the sums over squared joint probabilities in (14).

The Grassberger/Procaccia/Takens Algorithm (GPTA). The GPTA was originally used to calculate fractal (information) dimensions of probability measures defined on chaotic (strange) attractors, or to estimate generalized metric entropies.[9,20] It is based on the following idea:

Suppose a stationary continuous time series $\{Y_t\}_{t=1}^{T}$ and a corresponding realization (data sequence) $\{y_t\}_{t=1}^{T}$. Consider a D-dimensional embedding of the data with time delay coordinates

$$\{\mathbf{y}_t\}_{t=1+\Theta_{D-1}}^{T} \tag{16}$$

where $\mathbf{y}_t = (y_{t-\Theta_{D-1}}, ..., y_{t-\Theta_0})$, $0 = \Theta_0 < \Theta_1 < \cdots < \Theta_{D-1}$. Let us assume that the series is standardized with real values between 0 and 1. Then cover the D-dimensional cube $[0, 1] \times \cdots \times [0, 1]$ with an ε-partition $\beta_{D,\varepsilon}$, which should denote a rectangular grid of k^D boxes generated by dividing each axis into k segments of equal size $\varepsilon = 1/k$,

$$\beta_{D,\varepsilon} \equiv \{B_{\mathbf{m}}\}_{\mathbf{m}=1}^{k} \qquad \text{with} \quad B_{\mathbf{m}} = B_{m_{d-1}} \times \cdots \times B_{m_0} \tag{17}$$

where $B_m =](m-1)\varepsilon, m\varepsilon]$ for $m = 1, 2, ..., k$.

Consider a point $\mathbf{y}_{t_1}$ of the series (16) and let $B_{\mathbf{m}(t_1)}$ denote the box of $\beta_{D,\varepsilon}$ containing $\mathbf{y}_{t_1}$. We assume that for a sufficiently small value of ε the probability $p_{\mathbf{m}(t_1)}$ for finding any point of (16) in $B_{\mathbf{m}(t_1)}$ does not essentially change if the box is shifted in an ε-neighborhood of $\mathbf{y}_{t_1}$. Then this probability can be estimated from the time average, presuming ergodicity, as

$$p_{\mathbf{m}(t_1)} \simeq c_{D,\varepsilon/2}(t_1) = \lim_{T\to\infty} T^{-1} \sum_{t_2=1+\Theta_{D-1}}^{T} \mathscr{H}(\varepsilon/2 - \|\mathbf{y}_{t_1} - \mathbf{y}_{t_2}\|_{\max}) \tag{18}$$

where we use the Heaviside function $\mathscr{H}(x) = 0$ if $x \leqslant 0$ and $\mathscr{H}(x) = 1$ if $x > 0$, $\|\cdot\|_{\max}$ denotes the maximum norm in $\mathbb{R}^D$, and ε is assumed to be "small." Consider now a second time average over all probabilities $p_{\mathbf{m}(t_1)}$, which leads to the so-called *correlation integral*,[9,20]

$$\begin{aligned} C_{D,\varepsilon/2} &= \lim_{T\to\infty} T^{-1} \sum_{t_1=1+\Theta_{D-1}}^{T} c_{D,\varepsilon/2}(t_1) \\ &= \langle c_{D,\varepsilon/2}(t_1) \rangle_{t_1} \end{aligned} \tag{19}$$

Summarizing (18) and (19), we have

$$C_{D,\varepsilon/2} = \lim_{T\to\infty} C_{D,\varepsilon/2,T}$$

with

$$C_{D,\varepsilon/2,T} = N_{\text{total}}^{-1} \ \# \{(t_1, t_2) \text{ with } \|\mathbf{y}_{t_2} - \mathbf{y}_{t_1}\|_{\max} < \varepsilon/2\} \tag{20}$$

Herein $\#$ denotes the cardinality of the set, $1 + \Theta_{D-1} \leqslant t_1 < T$, $t_1 < t_2 \leqslant T$, and the total number of pairs is

$$N_{\text{total}} = \frac{(T - \Theta_{D-1})(T - \Theta_{D-1} - 1)}{2} \tag{21}$$

Due to the presumed ergodicity, the time average (19) can be alternatively written as a space average. Hence from (18) and (19) we obtain

$$C_{D,\varepsilon/2,T} \simeq \sum_{\mathbf{m}=1}^{k} p_{\mathbf{m}}^2 \qquad \text{for} \quad T \to \infty, \quad \varepsilon \to 0 \tag{22}$$

The estimation of $\sum_{\mathbf{m}} p_{\mathbf{m}}^2$ from the correlation integral (20) has computational advantages because we need no explicit partitioning of the state space. Of course, for $D = 1$ we could directly set $\sum_{\mathbf{m}} p_{\mathbf{m}}^2 = \varepsilon$ because of the assumed uniform one-dimensional distribution. Thus, according to (22), we would also set $C_{1,\varepsilon/2,T} \simeq \varepsilon$, though we obtain from a more sophisticated consideration that $C_{1,\varepsilon/2} = \varepsilon(1 - \varepsilon/4)$.

A similar algorithm works to estimate

$$\sum_{\mathbf{m},n=1}^{k} p_{\mathbf{m}n}^2(\tau) \simeq C_{D+1,\varepsilon/2,T}(\tau) \qquad \text{for} \quad T \to \infty, \quad \varepsilon \to 0 \tag{23}$$

using a $(D+1)$-dimensional embedding of the data: $\mathbf{y}_t = (y_{t-\Theta_{D-1}}, \ldots, y_{t-\Theta_0}, y_{t+\tau})$.

Now we can express the generalized mutual information (14) as

$$I_D^{(2)}(\tau) \simeq \log \frac{C_{D+1,\varepsilon/2}(\tau)}{C_{D,\varepsilon/2} C_{1,\varepsilon/2}} \qquad \text{for} \quad \varepsilon \to 0 \tag{24}$$

For practical implementations and discussions of statistical fluctuations in estimating the correlation integral we refer to refs. 1, 8, 10, 18, and 21.

Transformation to a Uniform Distribution. As we have shown above, our method would not work, in general, if the one-dimensional distribution of the time series is not the uniform distribution. Nevertheless, if we have no uniform distribution, then we propose to transform the original series

$$\{y_t\}_{t=1}^{T} \tag{25}$$

to a series

$$\{r_t\}_{t=1}^T \tag{26}$$

such that the empirical distribution of (26) is uniform, and then to apply our method to (26) instead of (25). The effect of this transformation can be compared to that of the method of Fraser,[7] because it would be the same to use a nonuniform partition with finer partition elements in the regions of phase space which have larger statistical weight, or to transform the data to a uniform distribution and then to use a uniform partition which corresponds to a fixed distance parameter ε in the GPTA.

If Y_t is continuous, then the transformation h: $y_t \rightarrow r_t$ is the distribution function of Y_t. This transformation is almost everywhere invertible and hence the statistical dependences of (25) are reflected in the transformed series (26) and vice versa. Moreover, due to the transformation h the generalized mutual information (24) becomes *invariant* with respect to any (possibly nonlinear but invertible) distortion g of the original signal. This is because the series (25) and a series $\{g(y_t)\}_{t=1}^T$ would provide the *same* uniformly distributed series (26).

Now let us suppose a more practical situation where a continuous signal is sampled with a certain sampling period and recorded using an analog digital convecter with k quantization levels. Hence (25) should be considered as a discrete sequence with $y_t \in \{1, 2,..., k\}$ rather than $y_t \in \mathbb{R}$. An invertible transformation to a uniform empirical distribution in $]0, 1]$ would be possible if all data in (25) are different, i.e., $y_{t_1} \neq y_{t_2}$ if $t_1 \neq t_2$. It can easily be done by transforming the data (25) to their relative rank numbers,

$$r_t = \frac{\#\{l \text{ with } 1 \leqslant l \leqslant T,\ y_l \leqslant y_t\}}{T} \quad \text{for} \quad t = 1, 2,..., T \tag{27}$$

using any sorting algorithm (e.g., quicksort).

However, the typical situation is that some values of a discrete series are equal. For instance, think of a time series which is recorded using an 8-bit analog/digital converter. Then necessarily there are equal values in the record if its length is $T > k = 256$, and the above transformation would not provide the desired uniform distribution. In this case we propose to distinguish arbitrarily between equal values of the original series (25). For instance, if $y_t \in \{1, 2,..., k\}$ and l values $y_{t_1}, y_{t_2},..., y_{t_l}$ of (25) are equal, then set $y_{t_2} \rightarrow y_{t_1} + 1/(l+1),..., y_{t_l} \rightarrow y_{t_1} + l/(l+1)$. Obviously we now have a series in which all data are different, and the corresponding sequence of relative rank numbers has the desired uniform distribution in $]0, 1]$. To make this arbitrariness of data transformation unimportant, we must guarantee that the distance parameter ε in the GPTA applied to the transformed series (26) is larger than $\varrho_{\max}\varepsilon_q$. Here ε_q denotes the relative quantization error (in our example $\varepsilon_q = 1/256$) of the original data and $\varrho_{\max} = |dh/dx|_{\max}$ is the maximum value of the one-dimensional distribution density of the original data. In practice we would set $\varrho_{\max} = l_{\mathrm{eq,max}}/T$, where $l_{\mathrm{eq,max}} = \max_t\{l_{\mathrm{eq},t}\}$, $l_{\mathrm{eq},t} = \#\{l \text{ with } 1 \leqslant l \leqslant T,\ y_l = y_t\}$, which is the maximum number of equal points in the series (25). In a forthcoming paper[12] we will discuss this problem in more detail.

Remarks and References. The expression (24) was already proposed by Grassberger *et al.*[10] They wrote that forecasting $\tau = 1$ time step ahead, e.g., would be possible *only* if the argument of the log in (24) is significantly larger than one, and any sort of correlations would imply $C_{D,\varepsilon}/C_{1,\varepsilon}^D > 1$. However, from the above considerations we see that, in general, $I_D^{(2)}(\tau)$ might be zero also if there are statistical dependences. Nevertheless, from the above theorem we conclude that the statement in ref. 10 is right if Y_t is uniformly distributed.

Moreover, it should be noted that Brock, Dechert, and Scheinkman (see refs. 18 and 3 and the references therein) have already considered the behavior of $C_{D+1,\varepsilon}(1)/C_{D,\varepsilon}$ as D varies, to measure the departures from independence between $\mathbf{Y}_t^D = (Y_{t-D+1},\ldots, Y_t)$ and $\mathbf{Y}_t^{D+1} = (Y_{t-D+1},\ldots, Y_t, Y_{t+1})$. Using the maximum norm, this ratio was considered as an estimate of the conditional probability that $\|\mathbf{y}_{t_2}^{D+1} - \mathbf{y}_{t_1}^{D+1}\|_{\max} < \varepsilon$ given that $\|\mathbf{y}_{t_2}^D - \mathbf{y}_{t_1}^D\|_{\max} < \varepsilon$. Further, they considered the equality

$$C_{D,\varepsilon} = C_{1,\varepsilon}^D \tag{28}$$

as a criterion for testing independence of a time series. This equality was found to hold for any $D = 2, 3,\ldots$ if the series is "independently and identically distributed" (IID). Then

$$B_{D,\varepsilon,T} = \sqrt{T}(C_{D,\varepsilon,T} - C_{1,\varepsilon,T}^D)$$

converges for $T \to \infty$ in distribution to the normal distribution $\mathcal{N}_{0,V}$ with mean 0 and the variance $V_{D,\varepsilon,T}$. The latter can be consistently estimated from the data, and the ratio $B_{D,\varepsilon,T}/(V_{D,\varepsilon,T})^{1/2}$ converges in distribution to $\mathcal{N}_{0,1}$. The IID hypothesis can be rejected if this ratio differs from zero, where the significance level can be read off from $\mathcal{N}_{0,1}$. However, Dechert (see ref. 3 and the references therein) has already recognized that from (28) we cannot, in general, conclude that the series is IID. Nevertheless, our theorem says that we can conclude from (28) to independence (on the coarse-grained level given by ε) if we have a uniform one-dimensional distribution. Moreover, we consider a profound information measure of the distance from the "purely stochastic case" and, on the other hand, from the "purely deterministic case."

If we use in (12) Shannon's information measure (4) instead of the generalized Rényi entropy (13), then we have an appropriate measure of statistical dependence also for nonuniform distributions of X_t. For $D=1$ this is $I(\tau)$ of Eq. (6). Pawelzik and Schuster[15] proposed an algorithm to estimate (6) on the basis of the so-called generalized correlation integrals

$$C^{(q)}_{D,\varepsilon}=\langle c_{D,\varepsilon}(t_1)^{q-1}\rangle_{t_1}^{1/(q-1)}$$

where $c_{D,\varepsilon}(t_1)$ is given by (18). For $q=2$ and a uniform marginal distribution this leads to our proposals and for $q\to 1$ this leads to that in ref. 15. However, we prefer to use the GMI (24) because it should have better statistics and somewhat less computation effort. The transformation of the data to a uniform marginal distribution is a well-known trick in order to enhance the significance of correlation dimension estimates (see, e.g., refs. 7 and 10), and it should improve the procedure in ref. 15 as well. However, the unification is not necessary in the case $q\to 1$, which is in contrast to our case $q=2$. But there is a second point; due to the transformation to a uniform marginal distribution, the GMI (as well as any other measure) becomes *invariant* with respect to any distortions of the signal, as was already mentioned above.

Example 1: White Noise. Consider a continuous time series with no statistical dependences (white noise). Moreover, X_t should be uniformly distributed in $[0,1]$. Then $(X_t, X_{t+\tau})$ is uniformly distributed in $[0,1]\times[0,1]$ for $\tau\geqslant 1$. Using an ε-partition of the unit square, we obtain $p_{mn}(\tau)=\varepsilon^2$, and from (14) it follows that $I_1^{(2)}(\tau)=0$. On the other hand, using the GPTA, we would obtain $C_{2,\varepsilon/2}(\tau)=\varepsilon^2(1-\varepsilon/2+\varepsilon^2/16)$ for $\varepsilon\leqslant 1$, which follows from simple calculation. Hence the GPTA would provide in the mean $I_1^{(2)}(\tau)\simeq\log[\varepsilon^{-2}C_{2,\varepsilon/2}(\tau)]=\log(1-\varepsilon/2+\varepsilon^2/16)$, which approaches zero with ε.

Example 2: Chaotic Signal. Consider a chaotic time series, representing successive measurements of a state variable of a chaotic dynamical system. Then we have

$$\lim_{\varepsilon\to 0}\lim_{D\to\infty}\log[C_{D,\varepsilon}/C_{D+1,\varepsilon}(\tau)]=\tau h^{(2)}$$

where $h^{(2)}$ is a positive finite value which represents the second-order metric entropy of the dynamical system generating the chaotic signal (see, e.g., ref. 6). (Here we assume that the time lags are regularly spaced, i.e.,

$\Theta_i = i\tau$ for $i = 0, 1, ..., D-1$.) Consequently from (24) we obtain $I_\infty^{(2)}(\tau) \simeq -\log\varepsilon - \tau h^{(2)}$ for sufficiently small values of ε. (For a "noisy" signal we have $h^{(2)} = \infty$, and for a nonchaotic deterministic signal $h^{(2)} = 0$.) If τ becomes large, then $I_\infty^{(2)}(\tau)$ typically decays with a rate less than $h^{(2)}$, because we have an information gain due to "foldings" which become relevant on a given coarse-grained level $\varepsilon > 0$.

Example 3: Quadratic Map $x_{t+1} = f(x_t) = 4x_t(1-x_t)$. This map is known to generate chaotic motions for almost all initial values $x_0 \in [0, 1]$ with respect to Lebesgue measure (see, e.g., ref. 5). The natural invariant measure of f is given by the density function $\varrho(x) = 1/\{\pi[x(1-x)]^{1/2}\}$ for $0 < x < 1$. For an application of our method to a typical chaotic time series $\{f^t(x_0)\}_t$ (e.g., $x_0 = 1/\pi$), we first have to transform the data to get a uniform one-dimensional distribution. This can be done analytically by the distribution function

$$y = h(x) = \int_0^x \varrho(z)\, dz = (2 \arcsin \sqrt{x})/\pi$$

Then we have $g \circ h = h \circ f$, with the tent map $g(y_t) = y_{t+1} = 2y_t$ if $0 < y_t < 1/2$ and $g(y_t) = 2 - 2y_t$ if $1/2 < y_t < 1$. The graph of g corresponds to the rank-delay representation r_{t+1} over r_t obtained from the original data according to (27), and the invariant measure of g is the Lebesgue measure on $[0, 1]$. [The invariant density of g is given by $\varrho_g(y) = \varrho(h^{-1}(y))\, |dh^{-1}/dy| = 1$, where h^{-1} is the inverse of h and $y \in (0, 1)$.]

Now we can analytically obtain the correlation integral $C_{2,\varepsilon/2}(1)$: First consider the points $(y_t, y_{t+1}) \in \mathbb{R}^2$ with $\varepsilon/4 < y_t < 1/2 - 3\varepsilon/8$ or $1/2 + 3\varepsilon/8 < y_t < 1 - \varepsilon/4$, and $\varepsilon/2 < 2/5$. For "small" values of ε this would be the main part of points, namely $1 - 5\varepsilon/4$ of all points. Obviously each of these points has $\varepsilon/2$ of all points as neighbors, with distance less than $\varepsilon/2$, using the maximum norm. Hence, the contribution of this part of points to the correlation integral is $\varepsilon/2 - 5\varepsilon^2/8$. The contributions of the rest of the points, which are characterized by $0 < y_t < \varepsilon/4$, $1 - \varepsilon/4 < y_t < 1$, and $1/2 - 3\varepsilon/8 < y_t < 1/2 + 3\varepsilon/8$, is easily found to be $5\varepsilon^2/4$. Summarizing, we obtain $C_{2,\varepsilon/2}(1) = \varepsilon(1/2 + 5\varepsilon/8)$. Moreover, we should set $C_{1,\varepsilon/2} = \varepsilon$ [see the arguments following (22)]. According to (24), we now obtain

$$I_1^{(2)}(1) \simeq \log C_{2,\varepsilon/2}(1)/\varepsilon^2 = \log \varepsilon^{-1}(1/2 + 5\varepsilon/8)$$

which approaches $-\log\varepsilon - \log 2 = H_1^{(2)} - \log 2$ for small ε. Note that $\log 2$ is the metric entropy of f, resp. g. On the other hand, it is well known that the correlation function (2) equals zero for $\tau = 1, 2, 3, ...$ (see, e.g., ref. 11). Hence this example demonstrates that, in general, cor is no appropriate indicator of statistical dependences—even in the deterministic case $X_{t+\tau}=f^{\tau}(X_t)$ the variables X_t and $X_{t+\tau}$ might be δ-correlated! On the other hand, the generalized mutual information describes the relation between X_t and X_{t+1} rather well, and even between X_t and $X_{t+\tau}$, $\tau>1$, as the following, more general example shows.

Example 4: Chaotic 1D Map. Suppose a 1D map $f\colon [0,1]\to[0,1]$, which should generate chaotic signals $\{x_t=f(x_{t-1})\}_{t=1}^{T}$ for almost all initial conditions x_0 with respect to a corresponding f-invariant measure. Moreover, assume that the delay representation for time lag 1 of the relative rank numbers of the orbit can be described for large values of the length T by the graph of a map g in the unit square $[0,1]\times[0,1]$. Then g has the Lebesgue measure μ_L as invariant measure. Suppose now that g is continuously differentiable at $y\in[0,1]$, except possibly at a finite number of points. Then we obtain for the absolute value of the slope $|g'(y)|>1$ for almost every $y\in[0,1]$, where $g'(y)=dg(y)/dy$. We conclude this from the g invariance of Lebesgue measure: $\mu_L(g^{-1}(B))=\mu_L(B)$ for any Borel set $B\subseteq[0,1]$. In terms of the invariant density ϱ_g this means

$$\varrho_g(y)=1=\sum_{y_i:\,g(y_i)=y}|g'(y_i)|^{-1}$$

For sufficiently small values of ε we find that a point $(y,g(y))\in[0,1]\times[0,1]$ has the relative part of $\varepsilon/|g'(y)|$ neighboring points with distance less than $\varepsilon/2$ (using maximum norm in $\mathbb{R}^2$). Thus we obtain in total $C_{2,\varepsilon/2}(1)=\varepsilon\int|g'(y)|^{-1}\,dy$.

A similar formula, in which g is replaced by $g^{\tau}=g\circ g^{\tau-1}$, $\tau=2,3,\ldots$, can be obtained for $C_{2,\varepsilon/2}(\tau)$. From (24), with $C_{1,\varepsilon/2}\simeq\varepsilon$, we finally obtain $I_1^{(2)}(\tau)\simeq-\log\varepsilon+\log\int_0^1|dg^{\tau}(y)/dy|^{-1}\,dy$ for $\varepsilon\to0$. This expression can be rewritten using the relation $|dg^{\tau}/dy|\simeq2^{\lambda\tau}$, which holds for large τ. Here λ is the Lyapunov exponent or metric entropy of order 1 of g (resp. f), measured in bits per time step, resp. per iteration. Now we obtain $I_1^{(2)}(\tau)\simeq-\log\varepsilon-\lambda\tau\log2$, which should hold for sufficiently small ε and large τ.

Example 5: Numerical Result for the Quadratic Map. The general result of Example 4 corresponds to that of Example 3, where $\lambda=1$ bit/iteration. Moreover, in the case of the tent map, which is conjugate to the quadratic map, we have $|dg^{\tau}/dy|=2^{\tau}$ for $\tau=1,2,3,\ldots$, and hence we expect $I_1^{(2)}(\tau)\simeq-\log\varepsilon-\tau\log2$, for sufficiently small ε. Figure 1 reflects the linear decay of the generalized mutual information. However, the curves were numerically obtained from the data $\{x_t\}_{t=1}^{T}$, $T=8192=2^{13}$, generated with the quadratic map $x_{t+1}=4x_t(1-x_t)$, $x_0=1/\pi$, applying the proposed first method: we first transformed the data according to (27) to

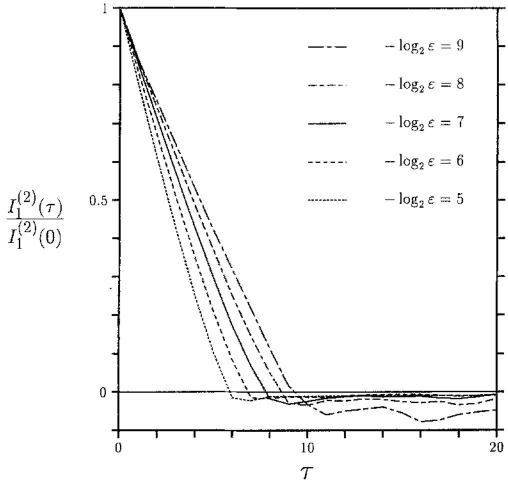

Fig. 1. Generalized mutual information of a chaotic time series generated by the quadratic map $x_{t+1} = 4x_t(1 - x_t)$ of Example 5, applying our first method.

get the series (26) of relative rank numbers having a uniform one-dimensional distribution. Then we estimated the generalized mutual information (12) for $D = 1$ using the GPTA to estimate (20) and using the relation (23).

In the figure we used a normalized representation $I_1^{(2)}(\tau)/I_1^{(2)}(0)$ with $I_1^{(2)}(0) = -\log \varepsilon$. Hence we would expect that $I_1^{(2)}(\tau)/I_1^{(2)}(0) = 1 + \tau/\log_2 \varepsilon$ for $\varepsilon \to 0$. However, if ε is fixed, e.g., at 2^{-9}, then ε cannot be considered to be small if $\tau \approx 9$ or $\tau > 9$, and the above formula, describing a linear decay of the mutual information, no longer holds. In this case the "initial" error ε is spread, on an average, over the whole interval $[0, 1]$ due to the exponentially expanding action of the chaotic map. Hence we would expect $I_1^{(2)}(\tau) = 0$ for $\tau > -\log_2 \varepsilon$, which is the situation of "white noise" considered in Example 1. There we have found that the use of the correlation integral from the GPTA to estimate the mutual information $I_1^{(2)}(\tau)$ leads to a systematic underestimation of the mutual information:

$$I_1^{(2)}(\tau)/I_1^{(2)}(0) \simeq \log(1 - \varepsilon/2 + \varepsilon^2/16)/\log \varepsilon^{-1} \approx \log(1 - \varepsilon/2)/\log \varepsilon^{-1} < 0$$

This explains the negative values of the mutual information in the cases $\varepsilon = 2^{-5}$ and 2^{-6}. For instance, take $\varepsilon = 2^{-5}$; then we obtain from the above formula a systematic error of $\log_2(1 - 2^{-6})/5 = -0.00454...$; which rather well matches the obtained deviations illustrated in Fig. 1. (For $\tau > 10$ the upper lines correspond to $\varepsilon = 2^{-5}$, 2^{-6}.) However, for smaller values of ε the deviations of the mutual information from zero are first of all caused by statistical errors due to the finite length of the time series. This leads

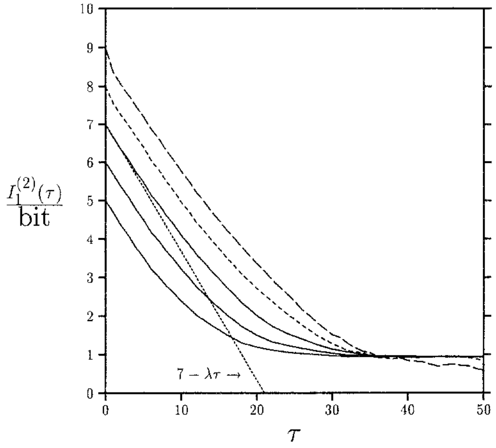

Fig. 2. Decay of the generalized mutual information of a time series obtained from a chaotic two-band attractor of the quadratic map of Example 5. Parameter: $I_1^{(2)}(0) = -\log_2 \varepsilon$.

especially for $\tau \geqslant 16$ and $\varepsilon = 2^{-9}$, 2^{-8} (the two lower lines in the figure) to considerable variations of the curves. When we used another realization of the chaotic orbit with the same length, then sometimes we found these variations to be much smaller.

Finally, Fig. 2 shows our results for the nonmixing chaotic map $x_{t+1} = 3.64 \times x_t(1 - x_t)$, which generates for "typical" initial values x_0 orbits on a chaotic two-band attractor with a Lyapunov exponent of $\lambda = 0.333 \pm 0.002$ bit/iteration. The two upper curves corresponding to $\varepsilon = 2^{-9}$ and 2^{-8} show again rather large statistical fluctuations, which are due to the finite data length of $T = 2^{13}$. However, for $\varepsilon = 2^{-5}$, 2^{-6}, and 2^{-7} (three lower curves) the decay rather well matches our general considerations of Examples 2 and 4: For small values of the time lag τ (1 to ~ 5) the decay rate is well described by λ (indicated by the dashed line at the mean curve with $\varepsilon = 2^{-7}$). Then we have a region of a slower decay which is due to "foldings" ($\sim 5 < \tau < \sim 35$), and finally the curves approximate 1 bit, which is the asymptotically remaining information on the band of the two-band attractor containing the future state ($\tau > \sim 35$). Actually we slightly underestimate the value of 1 bit, due to systematic errors. They can be derived similarily to that described in Example 1, recognizing that for $\tau \rightarrow \infty$ the points $(r_t, r_{t+\tau})$ of the rank delay representation can be considered as uniformly scattered over $([0, 0.5] \times [0, 0.5]) \cup ([0.5, 1] \times [0.5, 1])$ if τ is even, and over $([0, 0.5] \times [0.5, 1]) \cup ([0.5, 1] \times [0, 0.5])$ if τ is odd.

4. SECOND METHOD: MUTUAL ACCOUNT

Definition. Suppose a stationary discrete time series $\{X_t\}$, where X_t attains one of k possible different values. ($\{X_t\}$ could be considered as a coarse-grained version of a continuous series $\{Y_t\}$.) Let $\{p_{\mathbf{mn}}(\Theta, \tau)\}_{\mathbf{mn}}$ denote the joint probability distribution of $(\mathbf{X}_{t-\Theta}, \mathbf{X}_{t+\tau})$, and hence the distributions of $\mathbf{X}_{t-\Theta}$ and $\mathbf{X}_{t+\tau}$ are given by $p_{\mathbf{m}}(\Theta) = \sum_{\mathbf{n}} p_{\mathbf{mn}}(\Theta, \tau)$ and $p_{\mathbf{n}}(\tau) = \sum_{\mathbf{m}} p_{\mathbf{mn}}(\Theta, \tau)$, respectively.

Now we define a quantity which is somewhat similar to the contingency (7),

$$\Phi_D^2(\tau) \equiv \sum_{\mathbf{m},\mathbf{n}=1}^{k} [p_{\mathbf{mn}}(\Theta, \tau) - p_{\mathbf{m}}(\Theta)\, p_{\mathbf{n}}(\tau)]^2 \tag{29}$$

It can be rewritten as

$$\Phi_D^2(\tau) = S_{\mathbf{mn}}^2 - 2S_{\mathbf{mn},\mathbf{m},\mathbf{n}} + S_{\mathbf{m}}^2 S_{\mathbf{n}}^2 \tag{30}$$

using the abbreviations

$$S_{\mathbf{mn}}^2 = \sum_{\mathbf{m},\mathbf{n}=1}^{k} p_{\mathbf{mn}}^2(\Theta, \tau), \qquad S_{\mathbf{m}}^2 = \sum_{\mathbf{m}=1}^{k} p_{\mathbf{m}}^2(\Theta), \qquad S_{\mathbf{n}}^2 = \sum_{\mathbf{n}=1}^{k} p_{\mathbf{n}}^2(\tau) \tag{31}$$

$$S_{\mathbf{mn},\mathbf{m},\mathbf{n}} = \sum_{\mathbf{m},\mathbf{n}=1}^{k} p_{\mathbf{mn}}(\Theta, \tau)\, p_{\mathbf{m}}(\Theta)\, p_{\mathbf{n}}(\tau) \tag{32}$$

To get a logarithmic scale we derive from (29) a new quantity, which we call "mutual account,"

$$A_D(\tau) \equiv \log \left[\frac{\Phi_D^2(\tau)}{S_{\mathbf{m}}^2 S_{\mathbf{n}}^2} + 1 \right] \tag{33}$$

It can be rewritten as

$$A_D(\tau) = \log \left[\frac{S_{\mathbf{mn}}^2}{S_{\mathbf{m}}^2 S_{\mathbf{n}}^2} + 2\left(1 - \frac{S_{\mathbf{mn},\mathbf{m},\mathbf{n}}}{S_{\mathbf{m}}^2 S_{\mathbf{n}}^2}\right) \right] \tag{34}$$

Properties. From (29) and (33) we see that $A_D(\tau) \geqslant 0$, where the equality holds if and only if $\mathbf{X}_{t-\Theta}$ and $\mathbf{X}_{t+\tau}$ are statistically independent [i.e., iff $p_{\mathbf{mn}}(\Theta, \tau) = p_{\mathbf{m}}(\Theta)\, p_{\mathbf{n}}(\tau)$ for all $\mathbf{m}$ and $\mathbf{n}$].

Moreover, we always have $S_{\mathbf{mn}}^2 \leqslant S_{\mathbf{n}}^2$ and $S_{\mathbf{mn},\mathbf{m},\mathbf{n}}/S_{\mathbf{m}}^2 \in [p_{\mathbf{n},\min}, p_{\mathbf{n},\max}]$, where $p_{\mathbf{n},\min} = \min\{p_{\mathbf{n}}\}$ and $p_{\mathbf{n},\max} = \max\{p_{\mathbf{n}}\}$. From this we obtain the following estimation of an upper bound $A^*(\tau)$ of the mutual account (34):

$$\log\left(2 + \frac{1 - 2p_{\mathbf{n},\max}}{S_{\mathbf{n}}^2}\right) \leqslant A^*(\tau) \leqslant \log\left(2 + \frac{1 - 2p_{\mathbf{n},\min}}{S_{\mathbf{n}}^2}\right) \tag{35}$$

provided that the argument of the log is positive. If $p_{\mathbf{n},\max} \gg p_{\mathbf{n},\min}$ and $1 \approx 2p_{\mathbf{n},\max}$, then formula (35) is not a valuable estimation. However, the typical situation in the following will be that $p_{\mathbf{n},\max} \ll 1$, and even $p_{\mathbf{n},\max} \to 0$ for $\varepsilon \to 0$, where ε is a distance parameter in the algorithm to estimate the mutual account which is described below. Moreover, for $p_{\mathbf{n},\max} \to 0$ we also have $S^2_{\mathbf{n}} \to 0$ and hence we obtain from (35)

$$A^*(\tau) \simeq \log\left(2 + \frac{1}{S^2_{\mathbf{n}}}\right) \simeq \log\left(\frac{1}{S^2_{\mathbf{n}}}\right) \qquad \text{for} \quad p_{\mathbf{n},\max} \to 0 \tag{36}$$

Consider now the deterministic case where $\mathbf{X}_{t+\tau}$ follows unequivocally from $\mathbf{X}_{t-\Theta}$. Thus we have $p_{\mathbf{mn}}(\Theta, \tau) = p_{\mathbf{m}}(\Theta)$ for exactly one $\mathbf{n} = \mathbf{n}(\mathbf{m})$ and $p_{\mathbf{mn}}(\Theta, \tau) = 0$ for the other $\mathbf{n} \neq \mathbf{n}(\mathbf{m})$. Hence we obtain $S^2_{\mathbf{mn}} = S^2_{\mathbf{m}}$, $S_{\mathbf{mn},\mathbf{m},\mathbf{n}} = \sum_{\mathbf{m}} p^2_{\mathbf{m}}(\Theta)\, p_{\mathbf{n}(\mathbf{m})}(\tau)$, and again $S_{\mathbf{mn},\mathbf{m},\mathbf{n}}/S^2_{\mathbf{m}} \in [p_{\mathbf{n},\min}, p_{\mathbf{n},\max}]$. Thus the mutual account can be estimated in the deterministic case as in (35), resp. (36).

Summarizing, we have shown the following theorem:

Suppose a stationary discrete time series $\{X_t\}$, where the random variable X_t attains one of k possible different values. Then the mutual account, defined in (33), between a D-dimensional vector of "past" states

$$\mathbf{X}_{t-\Theta} = (X_{t-\Theta_{D-1}}, \ldots, X_{t-\Theta_0}), \qquad D = 1, 2, 3, \ldots$$

$$\Theta_{D-1} > \cdots > \Theta_1 > \Theta_0 = 0$$

and a d-dimensional vector of "future" states

$$\mathbf{X}_{t+\tau} = (X_{t+\tau_0}, \ldots, X_{t+\tau_{d-1}}), \qquad d = 1, 2, 3, \ldots, \quad 0 < \tau_0 < \tau_1 < \cdots < \tau_{d-1}$$

satisfies the relation

$$0 \leqslant A_D(\tau) \leqslant A^*(\tau) \tag{37}$$

where the upper bound $A^*(\tau)$ can be estimated according to (35), resp. (36). Moreover, $A_D(\tau) = 0$ if and only if $\mathbf{X}_{t-\Theta}$ and $\mathbf{X}_{t+\tau}$ are statistically independent. In the deterministic case, where $\mathbf{X}_{t+\tau}$ follows unequivocally from $\mathbf{X}_{t-\Theta}$, we have $A_D(\tau) = A^*(\tau)$.

If we ask for only one future state (i.e., if $d = 1$ and hence $\tau = \tau_0 = \tau$) and if X_t is uniformly distributed with $p_n = \varepsilon$, then we obtain $S_{\mathbf{mn},\mathbf{m},\mathbf{n}} = \varepsilon S^2_{\mathbf{m}}$, $S^2_{\mathbf{n}} = \varepsilon$, and hence from (14) we deduce $A_D(\tau) = I^{(2)}_D(\tau)$.

Consider now the case $D = d = 1$. Here we have $\mathbf{m} = m_0 = m$, $\mathbf{n} = n_0 = n$, and $S^2_m = S^2_n$. Thus we obtain for the mutual account

$$A_1(\tau) = \log\left[\frac{S^2_{mn}}{(S^2_m)^2} + 2\left(1 - \frac{S_{mn,m,n}}{(S^2_m)^2}\right)\right] \tag{38}$$

Moreover, for $\tau=0$ we have $p_{mn}=p_m$ if $m=n$ and zero elsewhere, and thus $S^2_{mn}=S^2_m$, $S_{mn,m,n}=\sum_m p^3_m=S^3_m$. This motivates us to set $A_1(0)=\log[1/S^2_m+2-2S^3_m/(S^2_m)^2]$.

Estimation of $A_D(\boldsymbol{\tau})$. Suppose a continuous ergodic time series $\{Y_t\}^T_{t=1}$. Then the sums (31) can be estimated from a data sequence using the GPTA with a $(D+d)$-, D-, and d-dimensional embedding of the original data, respectively (see Section 3). However, estimating (32) is somewhat more difficult and will be discussed now.

Given the data sequence $\{y_t\}^T_{t=1}$, construct a triple of $(D+d)$-, D-, and d-dimensional points at the instants t_1, t_2, and t_3,

$$\mathbf{y}_{t_1-\boldsymbol{\Theta},t_1+\boldsymbol{\tau}}=(y_{t_1-\Theta_{D-1}},\ldots,y_{t_1-\Theta_1},y_{t_1-\Theta_0},y_{t_1+\tau_0},y_{t_1+\tau_1},\ldots,y_{t_1+\tau_{d-1}}) \tag{39}$$

$$\mathbf{y}_{t_2-\boldsymbol{\Theta}}=(y_{t_2-\Theta_{D-1}},\ldots,y_{t_2-\Theta_1},y_{t_2-\Theta_0}) \tag{40}$$

and

$$\mathbf{y}_{t_3+\boldsymbol{\tau}}=(y_{t_3+\tau_0},y_{t_3+\tau_1},\ldots,y_{t_3+\tau_{d-1}}) \tag{41}$$

They are defined for $\Theta_{D-1}<t_1\leqslant T-\tau_{d-1}$, $\Theta_{D-1}<t_2\leqslant T$, and $1-\tau_0\leqslant t_3\leqslant T-\tau_{d-1}$. Note that $\mathbf{y}_{t-\boldsymbol{\Theta}}$ and $\mathbf{y}_{t+\boldsymbol{\tau}}$ can be considered as a D- and d-dimensional projection of $\mathbf{y}_{t-\boldsymbol{\Theta},t+\boldsymbol{\tau}}$, respectively. Assume that the data are standardized with real values between 0 and 1, and cover the D- and d-dimensional embedding of the data according to (40) and (41) with ε-partitions $\beta_{D,\varepsilon}=\{B_{\mathbf{m}}\}$ and $\beta_{d,\varepsilon}=\{B_{\mathbf{n}}\}$ in a similar way as in (17).

Consider now an arbitrary point $\mathbf{y}_{t_1-\boldsymbol{\Theta},t_1+\boldsymbol{\tau}}$. Then its projections $\mathbf{y}_{t_1-\boldsymbol{\Theta}}\in B_{\mathbf{m}(t_1)}$ and $\mathbf{y}_{t_1+\boldsymbol{\tau}}\in B_{\mathbf{n}(t_1)}$. The probability $p_{\mathbf{m}(t_1)}$ for finding a point $\mathbf{y}_{t_2-\boldsymbol{\Theta}}$ in $B_{\mathbf{m}(t_1)}$ can be estimated in the presumed ergodic case analogous with (18),

$$p_{\mathbf{m}(t_1)}(\boldsymbol{\Theta})\simeq c_{D,\varepsilon/2}(t_1)=\lim_{T\to\infty}T^{-1}\sum_{t_2=1+\Theta_{D-1}}^{T}\mathscr{H}(\varepsilon/2-\|\mathbf{y}_{t_1-\boldsymbol{\Theta}}-\mathbf{y}_{t_2-\boldsymbol{\Theta}}\|_{\max,D}) \tag{42}$$

where $\|\cdot\|_{\max,D}$ denotes the maximum norm in $\mathbb{R}^D$. Similarly, we estimate the probability $p_{\mathbf{n}(t_1)}$ for finding any point $\mathbf{y}_{t_3+\boldsymbol{\tau}}$ in $B_{\mathbf{n}(t_1)}$

$$p_{\mathbf{n}(t_1)}(\boldsymbol{\tau})\simeq c_{d,\varepsilon/2}(t_1,\boldsymbol{\tau})=\lim_{T\to\infty}T^{-1}\sum_{t_3=1}^{T-\tau_{d-1}}\mathscr{H}(\varepsilon/2-\|\mathbf{y}_{t_1+\boldsymbol{\tau}}-\mathbf{y}_{t_3+\boldsymbol{\tau}}\|_{\max,d}) \tag{43}$$

Consider now the time average

$$\begin{aligned}K_{D,d,\varepsilon/2}(\boldsymbol{\tau})&=\lim_{T\to\infty}T^{-1}\sum_{t_1=1+\Theta_{D-1}}^{T-\tau_{d-1}}c_{D,\varepsilon/2}(t_1)\,c_{d,\varepsilon/2}(t_1,\boldsymbol{\tau})\\&=\langle c_{D,\varepsilon/2}(t_1)\,c_{d,\varepsilon/2}(t_1,\boldsymbol{\tau})\rangle_{t_1}\end{aligned} \tag{44}$$

which can be written as

$$K_{D,d,\varepsilon/2}(\boldsymbol{\tau}) = \lim_{T \to \infty} K_{D,d,\varepsilon/2,T}(\boldsymbol{\tau})$$

with

$$K_{D,d,\varepsilon/2,T}(\boldsymbol{\tau}) = N_{\text{total}}^{-1} \ \# \left\{ (t_1, t_2, t_3) \text{ with } \| \mathbf{y}_{t_1 - \boldsymbol{\Theta}} - \mathbf{y}_{t_2 - \boldsymbol{\Theta}} \|_{\max, D} < \frac{\varepsilon}{2} \right.$$
$$\left. \text{and } \| \mathbf{y}_{t_1 + \boldsymbol{\tau}} - \mathbf{y}_{t_3 + \boldsymbol{\tau}} \|_{\max, d} < \frac{\varepsilon}{2} \right\} \qquad (45)$$

and N_{total} is now the total number of triples (t_1, t_2, t_3).

Making use of the presumed ergodicity, (44) can be obtained alternatively from a space average. Considering also (42) and (43), we thus get

$$K_{D,d,\varepsilon/2}(\boldsymbol{\tau}) \simeq \sum_{\mathbf{m},\mathbf{n}=1}^{k} p_{\mathbf{mn}}(\boldsymbol{\Theta}, \boldsymbol{\tau}) \, p_{\mathbf{m}}(\boldsymbol{\Theta}) \, p_{\mathbf{n}}(\boldsymbol{\tau}) \qquad \text{for} \quad \varepsilon \to 0 \qquad (46)$$

Summarizing our proposals, we obtain from (34), (22), and (46)

$$A_D(\boldsymbol{\tau}) \simeq \log \left[\frac{C_{D+d,\varepsilon/2}(\boldsymbol{\tau})}{C_{D,\varepsilon/2} C_{d,\varepsilon/2}} + 2 \left(1 - \frac{K_{D,d,\varepsilon/2}(\boldsymbol{\tau})}{C_{D,\varepsilon/2} C_{d,\varepsilon/2}} \right) \right] \qquad \text{for} \quad \varepsilon \to 0 \quad (47)$$

where the correlation integrals are defined as in (19), but now with the embedding according to (39)–(41) to estimate $C_{D+d,\varepsilon/2}(\boldsymbol{\tau})$, $C_{D,\varepsilon/2}$, and $C_{d,\varepsilon/2}$, respectively.

For $d = 1$ we get $\boldsymbol{\tau} = \tau_0 = \tau$, and formula (47) could be compared with that of the generalized mutual information in (24). The additional term in (47) vanishes if the one-dimensional distribution of the time series is uniform.

In the most simple case, where $D = d = 1$, we get $\boldsymbol{\tau} = \tau_0 = \tau$, and

$$A_1(\tau) \simeq \log \left[\frac{C_{2,\varepsilon/2}(\tau)}{C_{1,\varepsilon/2}^2} + 2 \left(1 - \frac{K_{1,1,\varepsilon/2}(\tau)}{C_{1,\varepsilon/2}^2} \right) \right] \qquad \text{for} \quad \varepsilon \to 0 \qquad (48)$$

which should be compared with (38).

Example 6: Mutual Account of a Chaotic 1D Map. Suppose a 1D map f: $[0, 1] \to [0, 1]$, as in Example 4 of Section 3. f should generate a chaotic sequence $\{x_t = f(x_{t-1})\}_{t=1}^{\infty}$ for almost all initial conditions x_1 with respect to a corresponding f-invariant measure, which is assumed to be absolutely continuous with respect to Lebesgue measure. We denote the corresponding density function by ϱ. Moreover, we assume that f is continuously differentiable almost everywhere.

In order to estimate the mutual account (48), let us first derive an expression for $C_{1,\varepsilon/2}$. Obviously we have, for sufficiently small values of ε, a relative part $\varrho(x)\,dx$ of points in the sequence $\{x_t\}_{t=1}^{\infty}$, each of which has a relative part of $\varepsilon\varrho(x)$ neighbors with distance less than $\varepsilon/2$. Hence we obtain in total

$$C_{1,\varepsilon/2} \simeq \varepsilon \int \varrho^2(x)\,dx \qquad \text{for} \quad \varepsilon \to 0 \tag{49}$$

To get an expression for $C_{2,\varepsilon/2}(1)$, we have to distinguish between the two cases $|f'(x)| \equiv |df(x)/dx| > 1$ and $|f'(x)| < 1$. The relative part $\varrho(x)\,dx$ of points in the sequence $\{(x_t, f(x_t))\}_{t=0}^{\infty}$ has, in the first case, a relative part of $\varepsilon\varrho(x)/|f'(x)|$ points of $\varepsilon/2$-neighbors, and in the second case a relative part of $\varepsilon\varrho(x)$. Hence we obtain in total

$$C_{2,\varepsilon/2}(1) \simeq \varepsilon \left[\int_{|f'(x)|>1} \varrho^2(x)\,|f'(x)|^{-1}\,dx + \int_{|f'(x)|<1} \varrho^2(x)\,dx \right] \qquad \text{for} \quad \varepsilon \to 0 \tag{50}$$

Finally we have to estimate $K_{1,1,\varepsilon/2}(1)$: Obviously the relative part $\varrho(x)\,dx$ of points (x_{t_1}, x_{t_1+1}) from the sequence $\{(x_t, f(x_t))\}_{t=0}^{\infty}$ has a relative part of $\varepsilon\varrho(x)$ of points x_{t_2} of $\{x_t\}_{t=1}^{\infty}$ with $|x_{t_1} - x_{t_2}| < \varepsilon/2$, and a relative part $\varepsilon\varrho(f(x))$ of points x_{t_3} of $\{x_t\}_{t=1}^{\infty}$ with $|x_{t_1+1} - x_{t_3}| < \varepsilon/2$. Hence we obtain in total

$$K_{1,1,\varepsilon/2}(1) \simeq \varepsilon^2 \int \varrho^2(x)\,\varrho(f(x))\,dx \qquad \text{for} \quad \varepsilon \to 0 \tag{51}$$

Inserting (49)–(51) in (48) would provide the mutual account $A_1(1)$.

Supposing a uniform distribution, which is characterized by $\varrho(x) = 1$, we easily obtain from (49) $C_{1,\varepsilon/2} \simeq \varepsilon$ and from (51) $K_{1,1,\varepsilon/2}(1) \simeq \varepsilon^2$. From (50) and the fact that f must be expanding almost everywhere (see remark in Example 4 of Section 3), we finally obtain $C_{2,\varepsilon/2} \simeq \varepsilon \int |f'(x)|^{-1}\,dx$. Using this expressions in (48), we get for the mutual account $A_1(1) \simeq -\log\varepsilon + \log\int |f'(x)|^{-1}\,dx$. The same consideration can be done for f^τ, $\tau = 1, 2, 3, \ldots$, yielding the more general expression

$$A_1(\tau) \simeq -\log\varepsilon + \log\int |df^\tau(x)/dx|^{-1}\,dx$$

which holds if f is uniformly distributed in $[0, 1]$. This result coincides with that of Example 4 for the generalized mutual information $I_1^{(2)}(\tau)$, which is not surprising, given the background of the remarks following the above theorem.

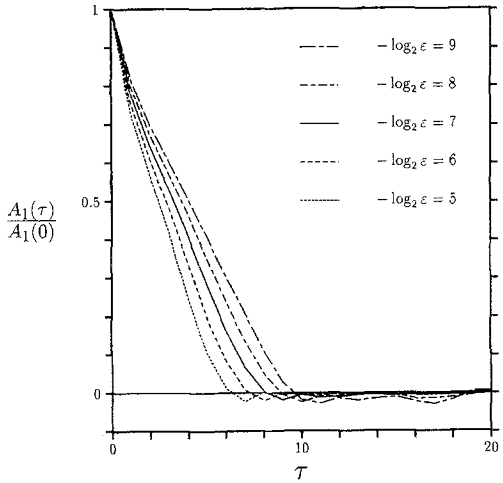

Fig. 3. Mutual account of a chaotic time series generated by the quadratic map of $x_{t+1} = 4x_t(1-x_t)$ of Exampe 5, applying our second method.

Example 7: Mutual Account of the Quadratic Map—Numerical Result. For the quadratic map of Example 3, Eqs. (46), (50), and (51) are not useful expressions, because they diverge for the corresponding invariant density $\varrho(x) = 1/(\pi[x(1-x)]^{1/2})$. Nevertheless, we can apply in this case the proposed numerical method to estimate $A_1(\tau)$ via formulas (48), (44), and (20). Figure 3 shows the corresponding results for the same data we used in Example 5 (Fig. 1) to estimate the generalized

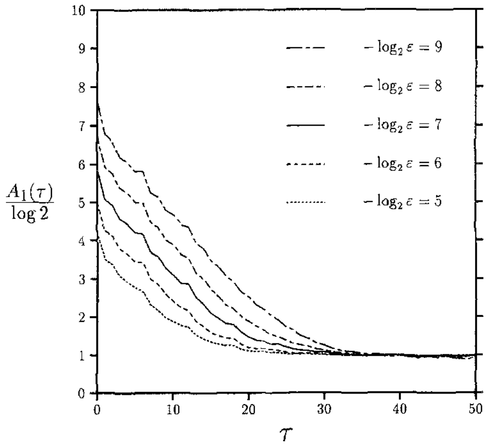

Fig. 4. Decay of the mutual account of a time series obtained from a chaotic two-band attractor of the quadratic map of Example 5.

mutual information. A comparison of both figures shows that the mutual account also rather well describes the decay of statistical dependences. For $\tau > -\log \varepsilon$ it approaches almost zero as well, where the deviations from zero are first of all due to statistical errors. (For a more detailed argument see Example 5.)

Moreover, Fig. 4 shows our results for the same signal used in Fig. 2. The distance parameter ε is measured in units of the distance between maximum and minimum of the data. Again we obtain asymptotically "1 bit," which is the information on the band of the two-band attractor containing the future state which is far away. The decay of the mutual account for small values of the time lag τ behaves like a zigzag, reflecting the fact that the two bands of the chaotic attractor are of different magnitude. This is an effect which cannot occur in Fig. 2 because there the date have a uniform marginal distribution due to the applied transformation.

5. CONCLUSIONS

We have proposed two methods to measure statistical dependences in a time series. The methods are applicable, in principle, to continuous ergodic time series. However, in practice, it would be sufficient to have discrete data with an accuracy of, say, at least 8 bits (i.e., the data should attain at least about $2^8 = 256 = \varepsilon^{-1}$ possible different values), and the data record should have a length of at least some thousands of points.

Of course, the data requirements depend on several circumstances. For instance, if we ask for statistical relations on a precision level ε, then our data have to be recorded with an accuracy of more than $-\log_2 \varepsilon$ bit. If the one-dimensional distributions of the data is not uniform, then the accuracy of the data record might have to be even much more than $-\log_2 \varepsilon$ to make our first method applicable.

Finally we want to mention that our methods also can be applied to measure cross-statistical dependences between different time series instead of relations within a single time series. In this case we have to transform each time series to a uniform one-dimensional distribution to apply our first method. In a forthcoming paper[12] we will describe fast algorithms for the proposed methods.

ACKNOWLEDGMENTS

I want to express my gratitude to Matthias Heilfort for his computer program implementations, from which the results of Fig. 3 were obtained, and for many stimulating discussions. Moreover, I am indebted to

Christoph Bandt for useful comments, and, last but not least, to a referee for pointing out that the methods work also for nonmixing time series and for directing my attention to the Brock–Dechert–Scheinkman algorithm.

REFERENCES

1. St. Bingham and M. Kot, Multidimensional trees, range searching, and a correlation dimension algorithm of reduced complexity, *Phys. Lett.* **140A**:327–330 (1989).
2. G. E. P. Box and G. M. Jenkins, *Time Series Analysis—Forecasting and Control* (Prentice Hall, Englewood Cliffs, New Jersey, 1976).
3. W. A. Brock, Causality, chaos, explanation and prediction in economics and finance, in *Beyond Belief: Randomness, Prediction and Explanation in Science*, J. L. Casti and A. Karlqvist, eds. (CRC Press, Boca Raton, Florida, 1991).
4. L. L. Campbell, A coding theorem and Rényi's entropy, *Information Control* **8**:423–429 (1965); Definition of entropy by means of a coding problem, *Z. Wahrsch. Verw. Geb.* **6**:113–118 (1966).
5. P. Collet and J.-P. Eckman, *Iterated Maps on an Interval as Dynamical Systems* (Birkhäuser, Boston, 1980).
6. J.-P. Eckman and D. Ruelle, Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.* **57**:617–656 (1985).
7. A. M. Fraser and H. L. Swinney, Using mutual information to find independent coordinates for strange attractors, *Phys. Rev. A* **33**:1134–1140 (1986); A. M. Fraser, Information and entropy in strange attractors, *IEEE Trans. Information Theory* **35**: 245–262 (1989).
8. P. Grassberger, Finite sample corrections to entropy and dimension estimates, *Phys. Lett. A* **128**:369–373 (1988); An optimized box-assisted algorithm for fractal dimensions, *Phys. Lett. A* **148**:63–68 (1990).
9. P. Grassberger and I. Procaccia, On the characterization of strange attractors, *Phys. Rev. Lett.* **50**:346 (1983); Measuring the strangeness of strange attractors, *Physica* **9D**:189–208; P. Grassberger, Generalized dimensions of strange attractors, *Phys. Lett.* **97A**:227–230 (1983).
10. P. Grassberger, Th. Schreiber, and C. Schaffrath, Nonlinear time sequence analysis, *Int. J. Bifurcation Chaos* **1**:521–547 (1991).
11. S. Grossmann and S. Thomae, Invariant distributions and stationary correlation functions of one-dimensional discrete processes, *Z. Naturforsch.* **32a**:1353–1363 (1977).
12. M. Heilfort and B. Pompe, in preparation.
13. N. S. Jayant and P. Noll, *Digital Coding of Waveforms—Principles and Applications to Speech and Video* (Prentice-Hall, Englewood Cliffs, New Jersey, 1984).
14. R. W. Leven and B. Pompe, Über die Möglichkeit der Zustandsvorhersage chaotischer Systeme, *Ann. Phys.* (Leipzig) **43**:259–278 (1986).
15. K. Pawelzik and H. G. Schuster, Generalized dimensions and entropies from a measured time series, *Phys. Rev. A* **35**:481–484 (1987); K. Pawelzik, *Nichtlineare Dynamik und Hirnaktivität* (Verlag Harri Deutsch, Thun, Frankfurt am Main, 1991), pp. 84–85.
16. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, New Jersey, 1978).
17. A. Rényi, *Probability Theory* (North-Holland, Amsterdam, 1970).
18. J. A. Scheinkman and B. LeBaron, Nonlinear dynamics and stock returns, *J. Business* **62**:311–337 (1989).

19. B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).
20. F. Takens, Invariants related to dimension and entropy, in *Atas do 13º Coloquio Brasileiro de Mathematica* (1983).
21. J. Theiler, Efficient algorithm for estimating the correlation dimension from a set of discrete points, *Phys. Rev. A* **36**:4456–4462 (1987); Statistical precision of dimension estimators, *Phys. Rev. A* **41**:3038–3051 (1990).
22. H. Tong, *Non-linear Time Series—A Dynamical System Approach* (Clarendon Press, Oxford, 1990).