Test for nonlinear dynamical behavior in symbol sequences

Henning Voss* and Jürgen Kurths
Institut für Physik, Universität Potsdam, Postfach 60 15 53, D-14415 Potsdam, Germany
(Received 13 December 1996; revised manuscript received 14 October 1997)

We discuss the analysis of highly discretized data, given as a time sequence of measurements, and propose a test for nonlinear dynamical behavior. Statistical significance is achieved by means of surrogate data, for which the construction rules are given. The method of surrogate data as often applied to the analysis of time series, however, cannot be used in the present case of the analysis of symbol sequences. Therefore, a considerably different technique is developed. The test is applied to several model examples.
[S1063-651X(98)05906-6]

PACS number(s): 02.50.Fz

In time series analysis, an important question is whether a given time series results from a linear stochastic process (colored noise) or a process with a nonlinear dynamical behavior. This could be dynamics on a strange attractor or any other irregular dynamics that can be described by nonlinear differential equations or nonlinear maps. The answer to this question has immediate consequences for modeling and the physical interpretation of the data [1]. Sometimes the data of an experiment are highly discretized or even given by symbols rather than numbers, e.g., in biology, physiology, meteorology, and climatology. This implies a huge information loss, since a symbol cannot carry the same amount of information as a real number, and, more important, between symbols no order relation exists. Experimental data are always limited in length and resolution. Additionally, the data are often corrupted by noise. To yield reliable results in spite of these limitations, one has to apply statistical tests that take all limitations into account.

In this contribution we will elaborate on a method to discriminate between symbol sequences that can be described by colored noise and dynamical ones. A useful definition of ''symbolic noise,'' however, is still lacking [2]. The test is performed by means of the method of surrogate data [3], which is widely applied for data given by a time series.

At first, we derive a certain property of ''noisy sequences'' that can be used for discrimination. We find that one cannot use here the same concept for constructing surrogate data as in the usual time series case. So we propose another approach to construct surrogate data for symbol sequences, using a new discriminating statistic. The method will be illustrated by examining several examples of nonlinear deterministic and stochastic sequences. A brief discussion concerning inherent limitations in the analysis of symbol sequences, in contrast to time series analysis, is reported in the last part of the paper.

It is important to mention that the approach introduced here is directed towards the analysis of real-world data. We do not consider ''symbolic dynamics'' in the sense that to each infinitely long symbol sequence a unique phase space point is associated by a generating partition (for an overview cf. [4,5]). Real-world data are always disturbed by noise, and in this case the concept of a generating partition is usually not clearly defined [6]; furthermore, in the following we assume that only a finite sequence is given.

Given a symbol sequence or a time series that has been produced by a process whose properties are not known, there exist several questions one might wish to answer. To do this, one can test the data with respect to a corresponding null hypothesis. However, it is easy to demonstrate that the null hypotheses differ for symbol sequences and time series. The simplest question is whether there is evidence for any structure at all. In the case of time series analysis, the corresponding null hypothesis is as follows: The time series can be fully described by identically and independently distributed (IID) noise. In the case of symbol sequences this means that the symbol sequence can be fully described by a realization of a Bernoulli process, i.e., a memoryless Markov process. Both hypotheses can be tested using shuffled data as surrogates.

A more complicated question is whether or not the time series can be modeled by a linear stochastic process. The corresponding null hypothesis $\mathbf{H_0}$ is that the time series can be generated by linearly autocorrelated Gaussian noise, also called colored noise. Surrogate data to test this hypothesis can be generated by an autoregressive (AR) model [7], which preserves the linear correlations. The amplitude-adjusted Fourier transform (AAFT) algorithm of Theiler *et al.* [3] also accounts for arbitrary monotone transformations Θ superimposed on the time series.

To test for $\mathbf{H_0}$ in the case of *symbol sequences*, as a first attempt one could apply the AAFT algorithm to the symbol sequence S. Therefore, one has to map S to a set of real numbers to yield a time series, say Y. For example, if S consists of α different symbols, the time series Y could consist of the numbers $\{1, \ldots, \alpha\}$ only. However, application of the AAFT algorithm to Y fails, because in this case Θ is not a one-to-one relation and thus the amplitude adjustment cannot be performed in a unique way. It is important to note that this is not only a technical problem. The deeper reason for the failure of the method is that there does not exist an order relation for symbol sequences; they always have to be related to numbers in some nonunique way to calculate quantities of time series analysis. An essential consequence is that for symbol sequences containing α different symbols there are $\alpha(\alpha-1)/2$ possibilities to define an autocorrelation function

*FAX: +49-331-977-1142. Electronic address: hv@agnld.uni-potsdam.de

(ACF) [8]. Therefore, in the case of $\alpha \geqslant 3$, which is studied here, one has several equivalent ACF's. So, a test for something analogous to colored noise cannot rely on preserving one special correlation structure. Here we will exploit other basic properties of colored noise that can be applied to symbol sequences.

Precisely, *colored noise* has to be understood as a realization of a general linear process $\{X_t\}$, which is defined as $X_t = \Sigma_{u=0}^{\infty} g_u e_{t-u}$, where $\{e_t\}$ is a Gaussian IID process, and $\{g_u\}$ is a given sequence of constants satisfying $\Sigma_{u=0}^{\infty} g_u^2 < \infty$ [9]. We call a finite realization $X = \{x_t\}_{t=1}^{N}$ of such a process a *linear time series*. A linear time series can be modeled by the AR model $x_t = \Sigma_{i=1}^{p} a_i x_{t-i} + e_t$, whose coefficients can be fitted to the data.

We now want to test whether a given symbol sequence $S = \{s_t\}_{t=1}^{N}$ can be described by colored noise. This is the case if there exists a linear time series X and a *measurement partition* Φ [6], which leads to the given symbol sequence S. The measurement partition Φ divides the state space into a finite number of sets, each of which is labeled by a symbol a_i of an alphabet $\mathcal{A} = \{a_1, \ldots, a_\alpha\}$. Thus, the unknown time series is transformed into a symbol sequence by Φ: $\mathbb{R} \rightarrow \mathcal{A}$, $x_t \mapsto s_t$. Now the null hypothesis $\mathbf{H_0}$ reads: *A measurement partition Φ and a suitable linear time series $X = \{x_t\}_{t=1}^{N}$ exists, such that the given symbol sequence $S = \{s_t\}_{t=1}^{N}$ is generated by $\Phi(X)$*. In this case S can be thought of as ''linear symbolic noise'' and contains no significant nonlinear dynamical behavior.

Of course, one cannot uniquely retrieve the unknown time series X that leads to the given symbol sequence S. But it is sufficient to show that the symbol sequence can be produced by the measurement partition of *at least one* linear time series. To test for $\mathbf{H_0}$ now we use the following criterion: The autocovariance of a general linear process is symmetric with respect to the time lag, i.e., $\mathrm{cov}(\tau) = \mathrm{cov}(-\tau)$. Given an arbitrary process (without loss of generality with zero mean) the covariance can be written as $\mathrm{cov}(\tau) = \int_{-\infty}^{\infty} p_{xy}(\tau) xy \, dx \, dy$, where y is the process shifted by τ time units and $p_{xy}(\tau)$ is the probability density $p_{xy}(\tau) = p(x_t = x \text{ and } x_{t+\tau} = y)$. Hence, the covariances $\mathrm{cov}(\tau)$ depend only on the symmetric part $p_{xy}^{s}(\tau) = \frac{1}{2}\{p_{xy}(\tau) + p_{yx}(\tau)\}$ of the probability density $p_{xy}(\tau)$:

$$\mathrm{cov}(\tau) = \int p_{xy}(\tau) xy \, dx \, dy \tag{1}$$

$$= \int p_{yx}(-\tau) xy \, dx \, dy \tag{2}$$

$$= \frac{1}{2} \int \{p_{xy}(\tau) + p_{yx}(\tau)\} xy \, dx \, dy. \tag{3}$$

Since a linear process can be characterized completely by its mean and covariances, the symmetry of the covariances leads to vanishing asymmetric constituents of $p_{xy}(\tau)$. This property is used here for the test. Because the partition Φ does not depend on time, the symmetry of the probability densities $p_{xy}(\tau)$ is preserved in the transition probabilities for the symbols, $p_{ij}(\tau)$: $p_{ij}(\tau) = p(s_t = a_i \text{ and } s_{t+\tau} = a_j) = p_{ji}(\tau)$ for all i, j, τ. Therefore, all symbol sequences whose statistical properties can be described completely by colored noise have symmetric transition probabilities, i.e., $p_{ij}(\tau) = p_{ij}^{s}(\tau)$. It is important to mention that the inverse does not hold, i.e., from symmetric transition probabilities it does *not* follow in general that the underlying time series is linear, such that one can only reject, but not confirm, $\mathbf{H_0}$.

TABLE I. The models to generate the time series, and the test results for different measurement partitions. The second column gives the number of elements of the alphabet related to the partition. The third column displays for which values of τ the quantity $\Psi(\tau)$ yields a significance of at least 98%. In case of ''none,'' the null hypothesis $\mathbf{H_0}$ cannot be rejected. The characters (a) to (f) refer to Figs. 1 and 2. For the continuous models Δt denotes the sampling interval.

Model	α	Significance
Hénon map [11] with $a = 1.4$, $b = 0.3$	3	1,2,4,5,6
	4	all (a)
	5	1,2,3,4,6
	6	none
Ikeda map [12]	3	1,4
with $a = 1.0$, $b = 0.9$, $\kappa = 0.4$, $\eta = 6.0$	4	1,2,4 (b)
	5	1,2,4,5
	6	2,4
Lorenz model [13]	3	2,3,4,5,6
with $\sigma = 10$, $\beta = 8/3$, $\rho = 28$,	4	2,3,4,5,6 (c)
$\Delta t = 0.05$	5	all
	6	all
Mackey-Glass equation [14]	3	3
with time delay $T = 17$, $\Delta t = 8$	4	2,3,5,6 (d)
	5	all
	6	2,3,5,6
AR(6) fit to the Hénon map	3	none
	4	none (e)
	5	none
	6	none
AR(6) fit to the Lorenz model	3	none
	4	none (f)
	5	none
	6	none

To test for $\mathbf{H_0}$ one has to test for the symmetry of the $p_{ij}(\tau)$. Symmetry can be quantified by the expression

$$\Psi(\tau) = \sum_{i,j=1}^{\alpha} \left(\frac{p_{ij}(\tau) - p_{ji}(\tau)}{2} \right)^2 \Big/ \, p_i p_j . \tag{4}$$

The quantity $\Psi(\tau)$ is zero if and only if $p_{ij}(\tau)$ is symmetric, and is restricted to the interval $[0, \alpha]$.

Because the asymptotic distribution of $\Psi(\tau)$ is not known, we estimate by a Monte Carlo method the *frequency* distribution of $\Psi(\tau)$ for an ensemble of surrogates (see below). Since this frequency distribution often differs considerably from Gaussianity, for computing significance levels we use the ''Monte Carlo probability,'' similar as in [10]. To test for a significance of p percent, we calculate $\Psi(\tau)$ for an

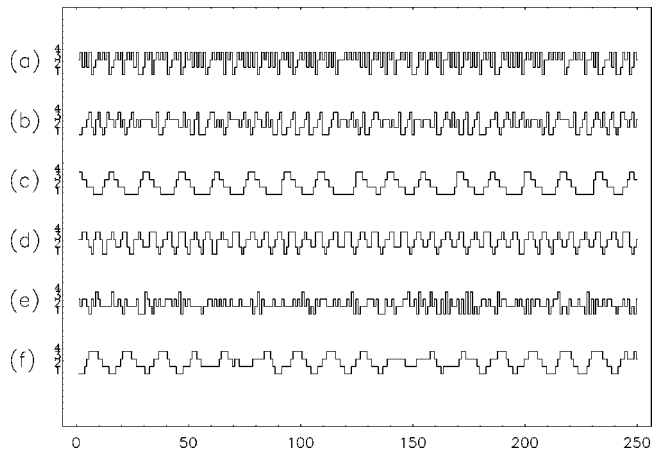

FIG. 1. Symbol sequences produced by a four-symbol measurement partition of the time series of Table I: (a) Hénon map, (b) Ikeda map, (c) Lorenz equations, (d) Mackey-Glass equation, (e) linearized Hénon time series, (f) linearized Lorenz time series. The first four sequences are nonlinear, the last two are linear ones.

ensemble of M surrogates with fixed parameter τ and cut the upper and lower $p/2$ percent of the resulting frequency distributions. If $\Psi(\tau)$ lies outside this frequency distribution, $\mathbf{H_0}$ is rejected at a significance level of p percent in the limit of an infinite M. For finite but large M the significance level varies by a small amount of the order of $1/M$, which we do not care about here.

A main point of this procedure is the calculation of the surrogates. Given a symbol sequence S, we first generate surrogate data with the same transition probabilities $p_{ij}(\tau)$, which serve as a control (step 1), then we calculate surrogates with the same $p^s_{ij}(\tau)$ (step 2):

Step 1: For the production of surrogates we use Markov models fitted to the sequence S, since these reproduce the transition probabilities $p_{ij}(\tau)$ under fewest constraints. The fitting of a Markov model P of order k is performed in the following way: For $k=1$, estimate the conditional probability $p_{j|i}$ for the occurrence of the symbol a_j right after the symbol a_i. For larger k, use the conditional probability $p_{j|W}$ for the occurrence of the symbol a_j right after the k letter word W. This guarantees the same transition probabilities $p_{ij}(\tau)$ at least up to $\tau=k$ for S and the surrogates. To generate surrogates by the Markov model P, choose a word of length k at random from S and continue by using the $p_{j|i}$ and $p_{j|W}$, respectively, for the following symbols.

Step 2: Symmetrize the Markov model P. For $k=1$, the symmetrization of P is performed by setting $p^s_{j|i}=p^s_{ij}/p_i$. For larger k, use

$$p^s_{j|W}=\tfrac{1}{2}(p_{j|W}+p_{j|W^b}), \tag{5}$$

where W^b represents the word W in reverse order. The symmetrized Markov model P^s has transition probabilities $p_{ij}(\tau)$, which are symmetric at least up to $\tau=k$. Note that the symmetrization of the Markov model does not distort the frequency distribution of the symbols. To generate surrogates by the Markov model P^s, first choose a word of length k at random from the sequence *or the reverse sequence* and continue by using the $p^s_{j|i}$ and $p^s_{j|W}$, respectively.

We now demonstrate the potentials of the proposed method. We perform the test for symbol sequences by considering six different systems. These are two nonlinear maps, two nonlinear differential equations, and two noisy linear maps, given in Table I. The noisy linear maps are produced by a global linearization of the Hénon and the Lorenz time series, via an AR(6) model. We use rather short realizations of these systems, containing only 250 data points, and we encode the corresponding dynamics adopting different measurement partitions. For the special case of a four-symbol measurement partition the corresponding symbol sequences are shown in Fig. 1. Here, Φ is chosen as an equipartition of the state space, the alphabet $\mathcal{A}$ given by $\{1,2,3,4\}$. By using the Markov model P, we first generate 100 surrogates to test up to which time lag τ the models are appropriate. We find for Markov models of fourth order that for all sequences and all $\tau\leqslant 6$ the quantity $\Psi(\tau)$ lies within the frequency distribution of the surrogates (Fig. 2). Testing $\mathbf{H_0}$ by using the symmetrized Markov model P^s reveals clearly the nonlinear origin of the first four sequences. These results and also the

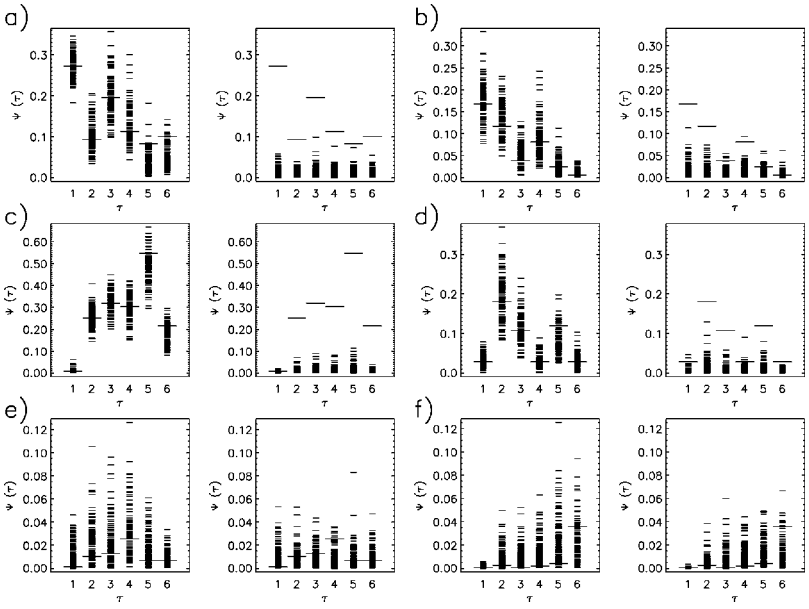

FIG. 2. The quantity $\Psi(\tau)$ for the sequences in Fig. 1(a) to 1(f), shown as —, and for 100 control (left) and symmetric (right) surrogates of sequences (a) to (f) shown as -, for time lags $\tau=1,\ldots,6$.

results for three-, five-, and six-symbol partitions, are summarized in Table I. Additionally, for other than equipartitions, we find a similar high sensitivity of the test, even for rather nonhomogeneous partitions of the phase space.

Here we performed six different tests for six different parameters τ, each with a single-test significance of 98%. It is beyond the scope of this paper to investigate how the total significance level, where all results are summed in one quantity, has to be calculated. Since the results are not independent of each other, this would be a formidable task, but we think that often the single-test significances can give useful hints for some underlying structure.

With this method we only test for linearity but not for stochasticity, i.e., a symbol sequence resulting from a partition of a *linear deterministic* time series would not yield significance in all cases. In contrast, if the AAFT method is applied directly to such a time series, it can yield significance. This has been already pointed out in [15]; the method of Theiler tests for the negation of ''linearity *and* stochasticity,'' which is ''nonlinearity *or* determinism.'' If one wants to exclude such misleading results in testing against linearity, one also has to test for determinism. Usual tests for determinism (e.g., [16]) are based on smoothness properties of the trajectories in the embedding space. Dealing with symbol sequences does not allow one to use analogous methods. Note also that sequences built from a two-letter alphabet necessarily have symmetric transition probabilities and cannot be tested by this method. In this case one still could apply methods based on word distributions [17], which, however, regard other hypothesis and usually need a larger amount of data than discussed here.

To summarize, we have introduced a method for distinguishing ''linear symbolic noise'' from nonlinear dynamics in experimental symbolic data. We consider a symbol sequence as linear symbolic noise, if it could result from a measurement partition of a linear *time series*. To test for this, we introduce a method of surrogate data suited for testing symbol sequences. The method works well for different measurement partitions, and the test shows a high sensitivity even for very short sequences. For possible applications we think of the analysis of symbol sequences obtained from physiological experiments [18], especially from experiments of neurophysiology [19] and cognitive complexity [20].

We thank Dr. A. Witt for useful discussions. H.V. acknowledges financial support from the Max-Planck-Gesellschaft.

[1] *Nonlinear Modeling and Forecasting, SFI Studies in the Science of Complexity*, edited by M. Casdagli and S. Eubank (Addison-Wesley, Redwood City, 1992).
[2] W. Li, J. Stat. Phys. **60**, 823 (1990).
[3] J. Theiler *et al.*, Physica D **58**, 77 (1992).
[4] R. Badii and A. Politi, *Complexity*, Cambridge Nonlinear Science Series (Cambridge University Press, Cambridge, 1997).
[5] E. Ott, T. Sauer, and J. A. Yorke, *Coping with Chaos*, Wiley Series in Nonlinear Science (John Wiley & Sons, New York, 1994).
[6] J. P. Crutchfield and N. H. Packard, Physica D **7**, 201 (1983).
[7] J. Kurths and H. Herzel, Physica D **25**, 165 (1987).
[8] H. Herzel and I. Große, Physica A **216**, 518 (1995).
[9] M. B. Priestley, *Spectral Analysis and Time Series* (Academic Press, London, 1981).
[10] P. E. Rapp *et al.*, Phys. Lett. A **192**, 27 (1994).
[11] M. Hénon, Commun. Math. Phys. **50**, 69 (1976).
[12] S. Hammel, C. K. R. T. Jones, and J. Maloney, J. Opt. Soc. Am. B **2**, 552 (1985).
[13] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).
[14] M. C. Mackey and L. Glass, Science **197**, 287 (1977).
[15] M. Paluš, in *Time Series Prediction—Forecasting the Future and Understanding the Past*, edited by A. S. Weigend and N. A. Gershenfeld (Addison-Wesley, Reading, MA, 1994).
[16] D. T. Kaplan and L. Glass, Phys. Rev. Lett. **68**, 427 (1992).
[17] T. Schürmann and P. Grassberger, Chaos **6**, 414 (1996).
[18] *Nonlinear Analysis of Physiological Data*, edited by H. Kantz, J. Kurths, and G. Mayer-Kress (Springer, Berlin, 1998).
[19] P. Tass *et al.*, Phys. Rev. E **54**, R2224 (1996).
[20] R. Engbert *et al.*, Phys. Rev. E **56**, 5823 (1997).