International Journal of Bifurcation and Chaos, Vol. 3, No. 2 (1993) 241–256

DETECTING NONLINEARITIES IN STATIONARY TIME SERIES

FLORIS TAKENS
Department of Mathematics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

Received December 27, 1992

In this review we survey methods for detecting nonlinearities in stationary time series. These methods are based on the estimation of so-called correlation integrals. These correlation integrals provide a way of analyzing time series and reveal aspects which are often complementary to the information one obtains from power spectra and autocorrelations. So we also focus our attention on the meaning and the estimation of the correlation integrals.

1. Introduction

In this paper we survey results, centered around the question of how to decide whether a given stationary time series is adequately described by a linear stochastic model or contains nonlinearities (or is even due to chaotic dynamics). This question has been considered before by Brock *et al.* [1987], LeBaron & Scheinkman [1989], see also Brock & Dechert [1988] (one now refers to the BDS test) and, from a different point of view, by Pijn [1990] and Theiler *et al.* [1991]. As we shall see, in all these approaches the *correlation integral* plays an essential role.

In order to make the above problem somewhat more concrete, we recall that it is known that deterministic dynamical systems can produce time series which look like random time series: a famous example is the logistic map $f(x) = 4x(1-x)$. This map defines a dynamical system with seemingly random or stochastic behavior in the following sense. If we take a point $x_0 \in [0, 1]$, then, with probability one, the time series $\{x_n\}$ with $x_n = f^n(x_0)$ is random in the sense that it is linearly uncorrelated: the average of $x_n \cdot x_{n+k}$, for any $k > 0$ is equal to the square of the average of the x_n. Such a time series is reproduced in Fig. 1.

For two recent surveys on dynamical systems with such "stochastic" behavior see Casdagli [1991] and Grasberger *et al.* [1991]. From these examples it follows that it is a nontrivial problem to determine whether a given, seemingly random, time series is really random or admits an explanation in terms of a simple deterministic mechanism (like this logistic map). The situation becomes even more

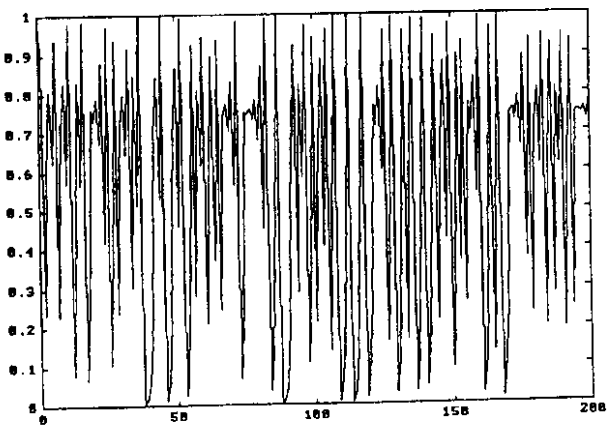


Fig. 1. A time series produced by the logistic map: the value of x_n (vertical) is given as function of the number n of iterates.

complex once we realize that the complexity of a time series may be due to a combination of nonlinearities (like in the logistic map) and random perturbations. Here we study the more restricted question whether a given time series shows an indication that it is not adequately described by a linear model. The methods however are inspired by what we know about chaotic dynamical systems and their seemingly random behavior.

To clarify the notion of a deterministic mechanism, we review in the next section some concepts of (chaotic) dynamical systems and the time series which they can produce. We consider the evolutions of a dynamical system (with finite dimensional state space) as deterministic evolutions (although the initial state may be random).

After this review we discuss the correlation integral in detail. This notion was introduced in the context of dynamical systems, but is applicable more generally. It provides a general method to obtain information about a time series and as such it is complementary to the spectral theory in which time series are analysed in terms of power spectra and autocorrelations (or autocovariance).

We then review the spectral theory of linear stochastic systems, their optimal (linear) predictors and the relation of these predictors with time series from (nonlinear) dynamical systems. We consider the time series produced by such stochastic linear systems as the prototype of random time series; in fact we think they can be considered as the "most random" time series with given autocorrelation (or autocovariance) coefficients.

Finally we come to the main subject and discuss the tests as considered by Scheinkman *et al.* and Pijn *et al.* These tests deal with the question whether a given time series has more structure than a completely stochastic time series with the same autocorrelation coefficients. This notion "more structure" may sound somewhat vague, but the test can also be interpreted as a way to determine whether one has to reject the possibility of a linear stochastic model or not. We also discuss a refinement of these tests which is in some sense a combination of both the test of Scheinkman *et al.* and the test of Pijn *et al.*

Arguments in the main sections are quite informal. For rigorous mathematical arguments see the appendix or the references.

At the end of this introduction I should say a few words about the difference between "nonlinear" and "chaotic." In extreme cases these notions are clear: in a time series produced by the above logistic map, there are nonlinear effects which produce chaotic behavior; there is no stochasticity (or randomness). On the other hand, if we take a time series, produced by a stochastic linear model (this notion will be defined later) and then take each value to the third power, we have a time series in which nonlinearity plays a role, but in which there is no question of "deterministic chaos." In general, if we have systems in which the time evolution is determined by rules involving both nonlinearity and stochasticity, it is hard, and may be even in principle impossible, to compare the effects of the nonlinearity and the stochasticity.

Another point which should be mentioned is that we are dealing here with general methods (or parameter free methods), i.e., we do not test whether, for example, a polynomial model (of given degree and order) gives a better fit. There are many investigations in this direction, especially in the application of neural nets, and in a number of cases they give much better results than the parameter free methods, but I think these parameter free methods have their value: first because of their theoretical interest, and second because there is often no clear justification for the assumption that possible nonlinearities have the special form implicitly assumed when using such special methods.

2. Chaotic Dynamical Systems

For simplicity of exposition, we consider here only time series with discrete time. The considerations carry over to the case of continuous time with only the obvious modifications.

A *dynamical system* is given by its *state space* X and by its *evolution map* $\varphi : X \to X$. For an initial state $x_0 \in X$ at time 0, the successive states at time 1, 2, etc are given by $x_1 = \varphi(x_0)$, $x_2 = \varphi^2(x_0)$, etc. In order to relate the notion of time series to such a dynamical system, we postulate that there is a *read-out function* $f : X \to \mathbb{R}$, which assigns to each state $x \in X$ the value $f(x) \in \mathbb{R}$ which is recorded or measured when the system is in the state x. So for the initial state x_0 we obtain the *time series* $\{y_n = f(\varphi^n(x_0))\}$. The sequence $\{\varphi^n(x_0)\}$ of successive states is called an *evolution* of the dynamical system. In what follows, we always assume that X is a finite dimensional vector space or a manifold, that the evolution map φ and the read-out function f are differentiable and that for each initial state x_0 the corresponding evolution $\{x_n = \varphi^n(x_0)\}$, for $n \geq 0$, remains in a bounded part of X.

We consider different kinds of evolutions. The main types are:

- stationary evolutions, i.e., evolutions $\{x_n = x_0\}$;
- periodic evolutions, i.e., evolutions $\{x_n\}$ such that for some $N > 0$ we have $x_n = x_{n+N}$ for all n;
- quasiperiodic evolutions, i.e., evolutions $\{x_n\}$ such that $x_n = F(\omega_1 n, \ldots, \omega_k n)$ for some multiperiodic function F (periodic with period 1 in each of its variables) and constants ω_1 to ω_k such that $1, \omega_1, \ldots, \omega_k$ are independent over the rationals;
- chaotic evolutions, i.e., evolutions $\{x_n\}$ such that there is a positive constant $A > 0$ so that for each $i \neq j$, there is some $m > 0$ with $\rho(x_{i+m}, x_{j+m}) > A$; ρ denotes the distance function in the state space X.

Apart from these types of evolutions we also consider evolutions which are asymptotically stationary, periodic or quasiperiodic: these are evolutions converging to stationary, periodic or quasiperiodic evolutions.

We do not claim that these are all the possible types of evolution. In fact any orbit in the Feigenbaum attractor, see Collet & Eckmann [1980], is neither of the above types. We could avoid this by defining a chaotic evolution as one which is neither (asymptotically) stationary, periodic or quasiperiodic, as in Ruelle & Takens [1971]. At present it is however common to include in any definition of chaoticity some form of sensitive dependence on initial conditions, as we have done in the above definition of a chaotic evolution. Also we believe that the above types of evolutions are the most common evolutions which occur in a persistent way. In any case, whenever necessary, we shall be more explicit about the different types of evolution which we consider.

As we remarked before, evolutions give rise to time series through the read-out function f. Also for time series we distinguish the above types. The definitions are the same, except that we have to replace x_n by $y_n = f(x_n)$ and ρ by the distance in $\mathbb{R}$. Generically, the type of an evolution is the same as the type of the corresponding time series. This is a consequence of a much more general result, often referred to as the *reconstruction theorem* [Packert *et al.*, 1980; Sauer *et al.*, 1991; Takens, 1981] which is of importance for later discussions in this paper, and which we formulate below. For a more complete discussion see Sauer *et al.* [1991], where the notion of genericity is also discussed and refined.

For a dynamical system on a state space X, which we assume to be a finite dimensional manifold or vector space, given by an evolution operator $\varphi : X \to X$ and a read-out function $f : X \to \mathbb{R}$ we define the reconstruction maps $\mathrm{Rec}_k : X \to \mathbb{R}^k$ by

$$\mathrm{Rec}_k(x) = (f(x), f(\varphi(x)), \ldots, f(\varphi^{k-1}(x))) \in \mathbb{R}^k .$$

The reconstruction theorem says that for generic pairs (φ, f), Rec_k defines an embedding of X in $\mathbb{R}^k$ whenever $k > 2 \cdot \dim(X)$. This means that for such k the state x is completely determined by $\mathrm{Rec}_k(x)$.

For a time series $\{y_n = f(\varphi^n(x_0))\}$, the image of the evolution $\{x_n\}$ under the reconstruction map Rec_k is obtained by taking the successive vectors $(y_i, \ldots, y_{i+k-1})$, also called *reconstruction vectors*. The reconstruction theorem then implies that, for k sufficiently big and (φ, f) generic, the limit set of these reconstruction vectors is homeomorphic to the so called ω-limit set of the evolution in X. (Usually this limit set is an attractor, but that is not essential for our considerations — note that even for time series which are not generated by a dynamical system, the limit set of the reconstruction vectors is often denoted by "the attractor".)

In less technical language, this reconstruction result means the following. The reconstruction vectors in $\mathbb{R}^k$ accumulate on a set which has the same shape as the set of states to which the "real" evolution is attracted (in the state space). "The same shape" is to be interpreted as homeomorphic, or equal up to a continuous deformation. So we can reconstruct this part of the state space (up to continuous deformations) from the time series without even knowing the state space itself. If we however do not know whether our time series is produced by a finite dimensional dynamical system, or whether our k is sufficiently big, then it is more difficult to interpret the set of points in $\mathbb{R}^k$ formed by the reconstruction vectors.

3. Correlation Integral

In this section we consider general time series, not necessarily produced by a dynamical system as defined above. Here and in what follows we will assume that our time series satisfy some mild conditions, namely that they are *bounded* and *stationary*. The first condition is clear (and only relevant for infinite time series). The second condition is more complicated. Informally speaking, a time series is stationary if its "character" does not change

with time, e.g., its "average" should not gradually increase (like the prices, in economic time series, due to inflation). A formal definition is the following. A bounded time series $\{y_n\}$ is *stationary* if for each k, and for each continuous $g : \mathbb{R}^k \to \mathbb{R}$, the average

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} g(y_i, \ldots, y_{i+k-1})$$

exists.

It is not hard to make examples of dynamical systems producing time series which are not stationary if one allows evolutions which do not stay in a bounded part of X. It is more surprising that this is also possible in cases where all evolutions are bounded in the future. It is however believed that such situations are not persistent, in the sense that by a small perturbation of the initial condition or evolution map this pathology can be removed. Also it has been proven that evolutions of many, even chaotic, dynamical systems give rise to stationary time series — for the evolutions on axiom A attractors this was proven in Ruelle [1976]. The situation is however not clarified in full generality.

Our definition of stationarity and assumption of boundedness imply that we have for each k a measure μ_k such that for each continuous function $g : \mathbb{R} \to \mathbb{R}$ we have

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} g(y_i, \ldots, y_{i+k-1}) = \int g d\mu_k .$$

Roughly speaking $\mu_k(A)$ is the fraction of the k-dimensional reconstruction vectors which happen to lie in A — one also calls μ_k the measure defined by the *relative frequencies* for k-dimensional reconstruction vectors. We mention here that, independently of the study of dynamical systems, the evolutions of reconstruction vectors, at least in dimension 2, were plotted by statisticians when analysing time series. See Tong [1990], page 216–217 and the references mentioned there.

For a bounded (and infinite) time series $\{y_n\}$ we define the *correlation integral* $P_k(\varepsilon)$ as the probability that two, randomly and independently chosen, k-dimensional reconstruction vectors are within distance ε in the supremum norm. In more detail: we define the distance $\rho_k(i, j)$ of the reconstruction vectors starting at i, j respectively as $\rho_k(i, j) = \max_{s=0,\ldots,k-1} |y_{i+s} - y_{j+s}|$. Using this, we define the fraction $P_{N,k}(\varepsilon)$ of the distances among the first N reconstruction vectors which are smaller than ε as

$$P_{N,k}(\varepsilon) = \frac{2}{N(N-1)} \text{ (number of pairs } (i, j) \text{ with } 0 \le i < j < N \text{ and } \rho_k(i, j) < \varepsilon) .$$

Then we define the correlation integral $P_k(\varepsilon)$ as the limit

$$P_k(\varepsilon) = \lim_{N\to\infty} P_{N,k}(\varepsilon) .$$

One can prove that this last expression has the following interpretation in terms of measures. Let $\mu_k \times \mu_k$ denote the product measure on $\mathbb{R}^k \times \mathbb{R}^k$, and let $U_\varepsilon(\Delta)$ denote the ε-neighborhood of the diagonal in $\mathbb{R}^k \times \mathbb{R}^k$. If the $\mu_k \times \mu_k$ measure of the boundary of $U_\varepsilon(\Delta)$ is zero, then the above limit converges to the $\mu_k \times \mu_k$ measure of $U_\varepsilon(\Delta)$. So that we have indeed that $P_k(\varepsilon)$ is the probability that two randomly and independently chosen points (with respect to μ_k) are within distance ε. In the exceptional case that the boundary of $U_\varepsilon(\Delta)$ is not a measure-zero set, the limit may not exist, but the liminf and the limsup exist and are contained between the measure of the interior and the measure of the closure of $U_\varepsilon(\Delta)$. In what follows we shall ignore this exceptional case.

This notion of the correlation integral was first introduced in connection with the time series produced by dynamical systems (see Grassberger & Procaccia [1983] and Takens [1983]), as a method to obtain information about invariants like dimension and entropy of attractors.

Only later was this same notion used by Scheinkmann and others to test whether a given time series might be modelled by a linear stochastic model or not. In this paper we are mainly interested in these last aspects, which will be the subject of Sec. 5. We conclude this section with a discussion of the correlation dimension in terms of this correlation integral. Although the correlation dimension is not our subject here, it may be helpful for getting an intuitive idea of the meaning of the correlation integral.

For this discussion we consider a somewhat more general situation: we do not need to restrict ourselves to $\mathbb{R}^k$ with the measure μ_k as defined above — the only thing we need is a set X on which a distance function ρ and a probability measure μ are defined.

Also in this context we define the correlation integral $P(\varepsilon)$ as the probability that two randomly (with respect to μ) and independently chosen points in X have distance less than ε. Now we observe

that if X is a q-dimensional cube, with the distance function ρ induced by the Euclidean distance and such that μ has a positive and continuous density with respect to the Lebesgue measure on $\mathbf{R}^q$, then $P(\varepsilon) \sim \varepsilon^q$, or, to put it more precisely,

$$q = \lim_{\varepsilon \to 0} \frac{\log(P(\varepsilon))}{\log(\varepsilon)}.$$

The same holds if X is a more general compact q-dimensional manifold. For this reason one can define the above limit, if it exists, as a dimension of X, depending both on the metric (as defined by the distance function ρ) and the measure μ. This is called the *correlation dimension*. Note that for measures concentrated on more complicated sets (fractals), this correlation dimension may take non-integer values. If we now think of X as a subset of $\mathbf{R}^k$ and of μ as a measure concentrated on X then we see that for smaller values of the correlation dimension, the values of $P(\varepsilon)$ converge slower to zero and hence are relatively bigger. For what follows it is important to note that if the correlation integral $P(\varepsilon)$ of a measure, defined on $\mathbf{R}^k$ converges to zero much slower than ε^k for $\varepsilon \to 0$, then the measure must be very unevenly distributed on $\mathbf{R}^k$ [like in the case where it is concentrated on a set of lower (correlation) dimension].

An important observation on the correlation dimension is the following. If we replace the distance function ρ by a different one, denoted by ρ', such that for some constant K and all $x, y \in X$ we have

$$K^{-1}\rho(x, y) \leq \rho'(x, y) \leq K\rho(x, y)$$

then we have for both distance functions the same correlation dimension. This follows from the fact that, denoting the correlation integral with respect to ρ and ρ' by P and P' respectively,

$$P(K^{-1}\varepsilon) \leq P'(\varepsilon) \leq P(K\varepsilon).$$

This has an important consequence when combined with the previously mentioned reconstruction results. If $\{y_n\}$ is a time series, obtained from a dynamical system with evolution map φ and read-out function f which are generic in the sense discussed before so that for certain k, Rec_k is an embedding, then the correlation dimension of the corresponding measure μ_k on $\mathbf{R}^k$ is a quantity which does not depend on the read-out function f but describes an intrinsic property of the dynamics. This is due to the fact that different read-out functions, provided they are generic, give measures which can be transformed into each other by diffeomorphisms, and diffeomorphisms, restricted to bounded sets, give only a distortion of distances by a bounded factor.

Another consequence is that we may take on $\mathbf{R}^k$ any metric (derived from a norm) we want, e.g., the Euclidean metric defined by the norm $|x| = \sqrt{x_1^2 + \cdots + x_k^2}$ or the supremum norm defined by the norm $|x| = \max_{i=1,\ldots,k} |x_i|$. The different choices differ only by a factor which is bounded and bounded away from zero.

Besides these aspects, we have to mention that the limit, occurring in the definition of the correlation dimension, makes it hard, and often impossible, to determine this quantity numerically.

Since the correlation integral plays an important role in the present paper, we discuss the numerical estimation of this quantity, together with the estimate of the error of these estimates, in the appendix.

4. Spectral Theory

In this section we review the classical theory of time series which is based on the interpretation of the power spectrum and autocorrelation coefficients. We consider again a time series $\{y_n\}$ which we assume to be stationary. The power spectrum of such a time series is defined as the square of the absolute value of the Fourier transform of $\{y_n\}$. Especially when the time series is infinite, this definition needs some refinements (the possible convergence problems which are involved are taken care of by our stationarity assumption). Here we do not go into these details since we shall mainly use the autocorrelation coefficients, which contain the same information as the power spectrum and which are easier to define (and to interpret); for more details on this classical time series theory see, for example, Priestley [1981].

Before defining the autocorrelation coefficients we make the average of our time series equal to zero. Note that the average

$$\mathcal{E}(\{y_n\}) = \lim_{n \to \infty} n^{-1} \sum_{i=0}^{n-1} y_n$$

is defined for a stationary time series. From now on we assume that this average has been made zero by subtracting it from each term y_n. We then define

the *autocovariance coefficients* as

$$R_k = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} y_i \cdot y_{i+k}$$

for $k \geq 0$, and define the *autocorrelation coefficients*, assuming that R_0, the *variance* of the time series, is nonzero as $\rho_k = R_k/R_0$. (For a time series with an average different from zero, we define the autocovariance and autocorrelation coefficients as being equal to those of the corresponding time series with zero average). We note that the above limits also exist for stationary time series. Concerning the interpretation of the autocorrelation coefficients we have the following observations:

- for all k, we have $-1 \leq \rho_k \leq 1$ and if $\rho_k = 1$ then $|y_n - y_{n+k}|$ is zero on average; in this case not all $y_n - y_{n+k}$ have to be zero: it may be that $y_n - y_{n+k}$ just converges to zero. If $\rho_k = -1$, then $|y_n + y_{n+k}|$ is zero on average — in both cases the time series is (asymptotically) periodic;
- for time series which are periodic or which are asymptotic to a periodic time series, both R_k and ρ_k are periodic in k and if k is a multiple of the period, then $\rho_k = 1$;
- for a time series which is quasiperiodic, (i.e., which is of the form $y_n = F(\omega_1 n, \ldots, \omega_l n)$ for some continuous function $F : \mathbf{R}^l \to \mathbf{R}$ which is periodic with period 1 in each of its variables and $1, \omega_1, \ldots, \omega_l$ independent over the rationals), we also have both R_k and ρ_k quasiperiodic in k; ρ_k is never equal to one, but its limsup is one. The same holds for a time series which is asymptotically quasiperiodic;
- if ρ_k converges to zero for $k \to \infty$, the time series is chaotic — the converse is "usually true" though there are exceptions, see the appendix.

In the rest of this section we assume, unless stated otherwise, that the variance R_0 of the time series under consideration is equal to 1 so that we do not have to distinguish between autocovariance and autocorrelation coefficients.

We define a *linear predictor* to be a linear map from $\mathbf{R}^k$ to $\mathbf{R}$, given by coefficients $a_1, \ldots, a_k$. The prediction for y_n according to this predictor is then

$$\hat{y}_n = a_1 y_{n-1} + \cdots + a_k y_{n-k} .$$

So it is a *linear* rule to predict a value of the time series in terms of a finite (here k) number of past values. The number k is called the *order* of the predictor. We say that a linear predictor is *optimal*, given the order and given the time series, if the expectation value of the square of the prediction errors $(y_n - \hat{y}_n)^2$ is minimal. The variance of a linear predictor is this expectation value of the square of the prediction error. It is not hard to verify that the coefficients of a linear optimal predictor of order k and its variance can be calculated from the autocorrelation coefficients ρ_0 up to ρ_k and that, given the optimal linear predictor of order k and its variance, not assuming the time series has variance R_0 equal to one, the autocovariance coefficients $R_0, \ldots, R_k$ can be determined. This means that the information about a time series contained in the optimal linear predictors is the same as the information contained in the autocorrelation coefficients (and hence in the power spectrum). Although these things are known, for completeness we derive the above stated relations, between optimal linear predictors and autocorrelation coefficients, in the appendix.

Of course optimal linear predictors are also defined, for time series whose average is nonzero. In that case one has to allow for a constant term, i.e., one has to consider an affine map from $\mathbf{R}^{k+1}$ to $\mathbf{R}$ and one gets the prediction in the form

$$\hat{y}_n = a_0 + a_1 y_{n-1} + \cdots + a_k y_{n-k} .$$

Optimal linear predictors, as discussed above, are especially suitable for time series which admit a linear model (this notion will be discussed below). However we first discuss how well they work for time series produced by (nonlinear) deterministic dynamical systems, as discussed in Sec. 2. For these systems we distinguished several types of evolutions and saw that under generic conditions we had corresponding types of time series. If we have a periodic, or asymptotically periodic, time series, say with period q, then it is clear that for $k \geq q$ the optimal linear predictor of order k can be chosen as $\hat{y}_n = y_{n-q}$ and this predictor is "perfect" in the sense that its variance is zero (still in the case of an asymptotically periodic time series, the prediction errors are not zero). Next we consider quasiperiodic (or asymptotically quasiperiodic) time series. In this case the autocorrelation coefficients ρ_k reach values arbitrarily close to one. So for increasing k, the variances of the optimal linear predictors of order k converge to zero for $k \to \infty$. In this case we say that the linear predictors are "asymptotically perfect." Finally in the case of chaotic time series, we expect optimal linear predictors not to

be asymptotically perfect any more. It is not very hard to prove that any time series, for which the optimal linear predictors are not asymptotically perfect, is chaotic in the sense we defined. The converse is not true but counterexamples are probably exceptional; for more details see the appendix. In any case the notions of chaoticity and optimal linear predictors not being asymptotically perfect are strongly related.

The linear predictors can also be used to produce stochastic time series which have the same correlation coefficients ρ_l, at least for values of $l \leq k$, as a given time series. One simply takes a time series generated by a model of the form

$$\tilde{y}_n = a_1\tilde{y}_{n-1} + \cdots + a_k\tilde{y}_{n-k} + \varepsilon_n\,, \quad \clubsuit$$

where $a_1, \ldots, a_k$ are the coefficients of an optimal linear predictor of order k for the given time series and where the ε_n are independently and identically distributed (*iid*) variables, taken from a distribution with mean zero and variance equal to the variance of the optimal predictor; we refer to terms ε_n as the *noise* in the model. We call a model of the form ♣ a *linear model*. Usually one takes the ε_n as normally distributed, but this would not be consistent with our convention of considering only bounded time series. One can however also take the ε_n from a uniform distribution. We point out that the coefficients of the optimal linear predictor can be determined from the spectrum or the autocorrelation coefficients of a time series, but that the spectrum or the autocorrelation coefficients give no information on what the probability distribution of the noise should be.

Time series generated by the above linear model have the first $k + 1$ autocorrelation coefficients $\rho_0, \ldots, \rho_k$ indeed equal to those of the original time series. The following autocorrelation coefficients are in general not zero: this would be incompatible with a theorem which says that for any integer l and reals $b_1, \ldots, b_l$ we have

$$\sum_{i,j=1,\ldots,l}^{n} b_i b_j \rho_{|i-j|} \geq 0\,.$$

One can show however that in a sense the ρ_l, for $l > k$, are as small as is permitted with the above restriction. In many examples it turns out that there is no significant difference between the autocorrelation coefficients of a time series and the autocorrelation coefficients of a time series made with a model of the above type with resonably low values of k.

We shall see in the next section how to construct, for a given time series, a model of type ♣ which gives time series that come closest to the original time series, and then show how to test whether the original time series and those of the ♣ model are significantly different.

5. BDS Test and Alternatives

As we observed before, the correlation integral was originally introduced as a means to calculate the correlation dimension of an attractor in a deterministic system, see Grassberger & Procaccia [1983], Takens [1983]. Later the correlation integral was used in Brock *et al.* [1987] to test whether a given time series might be iid (independently and identically distributed). The idea is as follows. If we consider a time series $\{y_n\}$ then, if it is iid, we know that the correlation integrals as defined in Sec. 3, have to satisfy $P_k(\varepsilon) = (P_1(\varepsilon))^k$. From a finite part of a time series, we can estimate these correlation integrals and in this way determine whether a finite segment of a time series gives an indication whether the hypothesis of iid should be rejected. Of course for this we need to know what the variance or standard deviation of our estimates for the correlation integral is — we shall discuss this in the appendix. As we remarked before, higher values of the correlation integrals are a sign for more structure in a time series. So in the case where the data are not independent, we expect an inequality $P_k(\varepsilon) \geq (P_1(\varepsilon))^k$. This is what one observes usually, but this is not a theorem: in the appendix we shall give an example where the opposite inequality holds.

In a situation where one has a time series which is not iid, but for which one wants to test whether it is described adequately by a linear model of the form like ♣ in the previous section, one proceeds as follows (we assume still, as in the previous section, that we have a time series with average zero). Let the time series be $\{y_n\}$ and the linear model, defined by an optimal linear predictor, be

$$\tilde{y}_n = a_1\tilde{y}_{n-1} + \cdots + a_k\tilde{y}_{n-k} + \varepsilon_n\,. \quad \clubsuit$$

Then one constructs the residues

$$r_n = y_n - (a_1 y_{n-1} + \cdots + a_k y_{n-k})$$

Now one can proceed with the residues as with the time series which was expected to be iid: if the residues are iid, then the linear model ♣ is indeed adequate. At this point we observe that if the original time series was obtained from a deterministic dynamical system as in Sec. 2, then still the

same holds for the residues. The residues can be obtained by changing the read-out function: If we have a dynamical system with state space X, evolution map φ and read-out function $f : X \to \mathbf{R}$ such that $x_n = \varphi^n(x_0)$ and $y_n = f(x_n)$ then we obtain a new read-out function in the following way. Define $F : X \to \mathbf{R}$ by

$$F(x) = f\varphi^k(x) - a_1 f\varphi^{k-1}(x) - \dots - a_{k-1} f\varphi^1(x) - a_k f(x).$$

Then it is clear that $r_n = F(x_{n-k})$, so that, except for a reindexing, the time series defined by F is the same as the time series of the residues. This indicates that the nonlinear or deterministic structure remains if we pass to the residues. On the other hand one can see from examples that the deterministic structure may be much harder to detect in the residues than in the original time series. Such an example is given in the appendix; see also Theiler *et al.* [1991].

In general the above procedure can be applied to any stationary time series. The order of the predictor should be such that the autocorrelation coefficients of the given time series and of a time series produced by the linear model are not significantly different. A different criterion for the right order of the optimal linear predictor is that the autocorrelation coefficients of the residues are sufficiently close to zero, compared with the standard deviation of the estimation procedure for these autocorrelation coefficients.

An alternative approach was used by Pijn [1990], in his thesis, and by Theiler *et al.* [1991]. Instead of *modifying* the original time series, he obtains a new, and random, time series with the same power spectrum, and hence the same correlation coefficients, and then uses the correlation integrals to see whether the original and the random time series are significantly different or not. His way of producing the random time series with the same power spectrum is as follows. We first compute the Fourier transform of our time series. The coefficients are complex numbers; we denote the coefficient for the frequency ω by $a(\omega)$. Since we started with a real time series, we have $a(\omega) = a(-\omega)$. Now, keeping the absolute values equal, we change the arguments of the Fourier coefficients in a random way, but so that the above relation $a(\omega) = a(-\omega)$ remains. Then we apply the inverse Fourier transform. The new time series has by construction the same power spectrum as the time series with which we started. It can be shown that the residues, as defined above, of this new time series are iid and have even a normal distribution. The test consists of verifying whether the correlation integrals of the time series, the original one and the one with randomized Fourier coefficients are significantly different or not.

In fact this procedure was not used originally by Pijn as an alternative for the BDS test, but as a way to see whether the plots of the correlation dimensions, were consistently significantly different from those of the randomized signals. This method has the advantage that we do not distort the possible deterministic structure in the original data. A disadvantage is that we compare with a random series whose residues are normally distributed. We remove this in the second alternative for the BDS test which we present below.

In this final alternative we proceed as follows. For the original time series, we calculate an optimal linear predictor as before and calculate the residues $\{r_i\}$. As before we denote the coefficients of the linear predictor by $a_1, \dots, a_k$. These residues, together with the linear model defined by the optimal linear predictor, are now used to produce a random time series $\{w_n\}$ as follows:

$$w_n = a_1 w_{n-1} + \dots + a_k w_{n-k} + r_{\varepsilon(n)}$$

where $r_{\varepsilon(n)}$ is chosen randomly from the set $\{r_i\}$. In this way we obtain a time series $\{w_n\}$ which has the same autocorrelation coefficients ρ_k, at least for low values of k, as the original time series while the residues are not automatically normal, but adapted to the original time series. Especially where we have reason to expect residues not to be normally distributed, like (according to some authors) in economic time series, this last method may have some advantages. In the next section we show an example where the variance in the estimation of the correlation integrals is quite different for this method, as compared with the method of Pijn and Theiler *et al.* As before, the order of the optimal linear predictor one uses here should not be too low. In fact one can take the order here rather high since we do not have the problem of getting an unwanted distortion of the deterministic structure.

The above discussion of the different alternatives may look somewhat confusing. So I want to conclude this section by discussing the basic principles behind the different alternatives.

The *correlation integral* is used as a device to distinguish time series which are otherwise, as far as

one can see from power spectra or autocorrelation coefficients, of the same type.

There are two classes of *transformation on time series* which were used in the present context:

- removing the autocorrelation coefficients, i.e., making them zero, without destroying the "deterministic structure" — this is done by taking the residues with respect to an optimal linear predictor; as the order of the predictor becomes higher, it will become more difficult to detect any deterministic structure;
- removing the "deterministic structure" without changing the autocorrelation coefficients — there are two ways of realizing this:

— random phase shifts in the Fourier coefficients — in this way the Fourier spectrum remains exactly the same and we get a time series as with a linear model with normally distributed noise;

— producing a new time series from a linear stochastic model, obtained from an optimal linear predictor and noise taken from the residues of the original time series; in this case the spectrum is only approximately preserved, but by taking the order of the optimal linear predictor higher this approximation can be made better.

All the different tests can be considered as producing two time series and then testing whether they are significantly different, using the correlation integrals. These two time series are equal in many aspects, but in one the deterministic structure is still present and in the other the deterministic structure is destroyed. In this way we think the last of the three possibilities mentioned is optimal in the sense that it combines the advantages of the BDS test and the test of Pijn *et al.*

6. An Example

In this section we discuss the analysis, as described above, for a time series of length 500 obtained from an experiment which was performed at the department of Chemistry at the Technical University Delft. We show below this time series, together with the time series obtained by

- randomizing the phases of the Fourier coefficients;
- using a linear model defined by an optimal linear predictor of order 20 and noise from the residues as discussed in Sec. 5;
- calculating the residues.

Already from these diagrams it seems to be clear that the real time series is much more "systematic" than the random ones (the second and third).

We now use the method of correlation integrals to establish the difference between the original time series and the two random ones. We estimate the correlation integrals by taking 500 pairs of reconstruction vectors in a "random" way from the set of all possible pairs and counting the number of pairs whose distance is less than a fixed value. (Below we explain why we did not take all pairs of reconstruction vectors.) As we discuss in the appendix, it is hard to give a good value for the variance of the error of this estimate for the correlation integral. For this reason we use a Monte Carlo method: the construction of the random time series ((b) and (c) in the diagram) can be repeated so as to obtain different samples of the same process. For the diagrams below, we constructed 25 such samples, estimated for each the correlation integrals (using 500 pairs of reconstruction vectors), and determined from this the average and variance of these estimations. The results are shown in the following diagrams.

As we see, the main difference between the use of the two random series is that in the last case the variance of the estimation error is much bigger. This indicates that it is indeed difficult to predict this variance (without using a Monte Carlo method) since in both cases the spectral properties are the same. The fact that these variances are bigger in the second case means that in the second case there is less reason to claim that the linear model does not give an adequate description of the original time series. This is in agreement with the fact that in this second case we used a better model: not only the spectral properties, but also the noise, was derived from the original time series.

We see that these results indicate that, even in the second case, the linear model can be rejected because, for embedding dimensions around 6 to 9 and at a length scale between 0.1 and 0.2, the estimates of the correlation integrals for the original time series differ from the average estimate for the random time series by about ten times the standard deviation or more. So not only can we reject the linear models, but we also see on what length scale the differences become apparent and that on length scales above the order 0.25 the second linear model is acceptable, but the first (with Gaussian noise) is not. The fact that the discrepancies show up mainly for embedding dimensions around 6 to 9 may be harder to interpret.

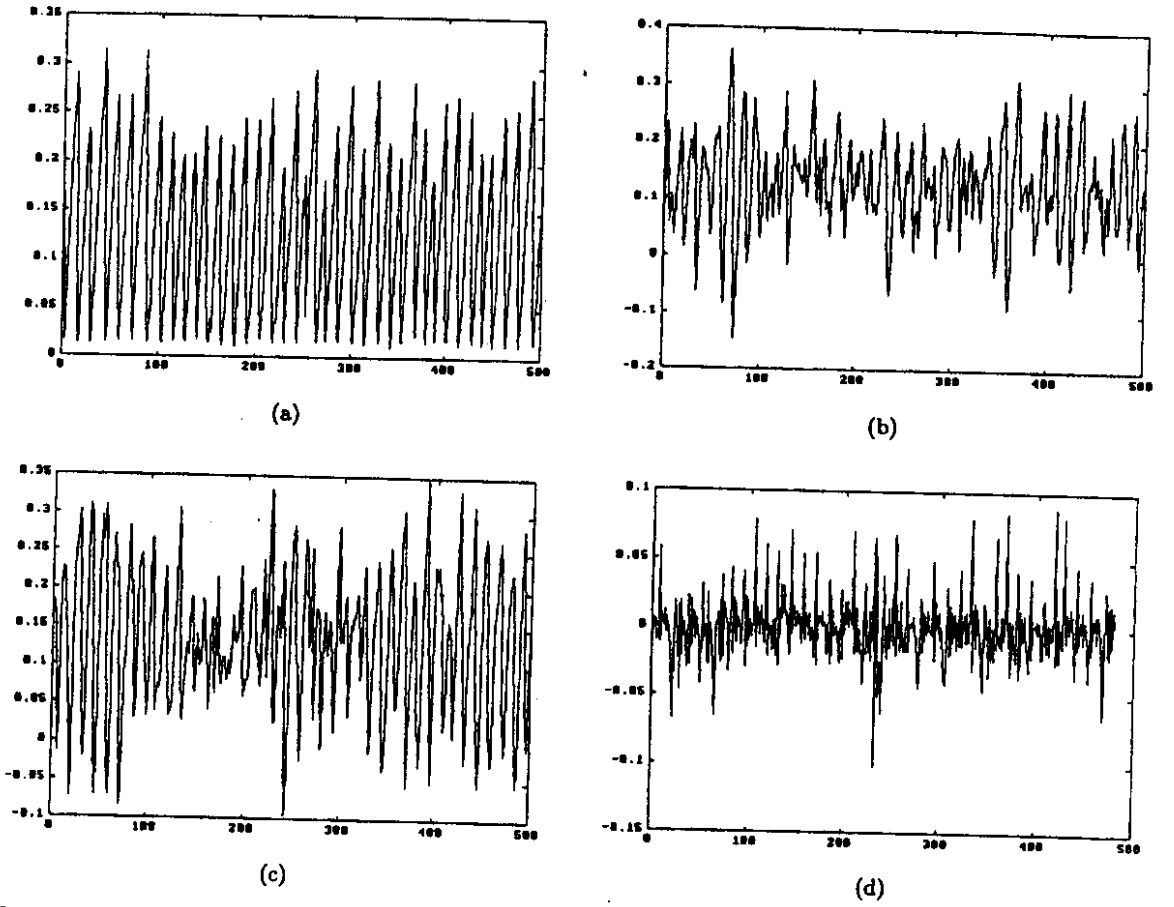


Fig. 2. Time series (a), two random time series with the same spectral properties, one obtained by randomizing Fourier coefficients (b) and one by using a linear model with noise (c), and the time series of residues (d).

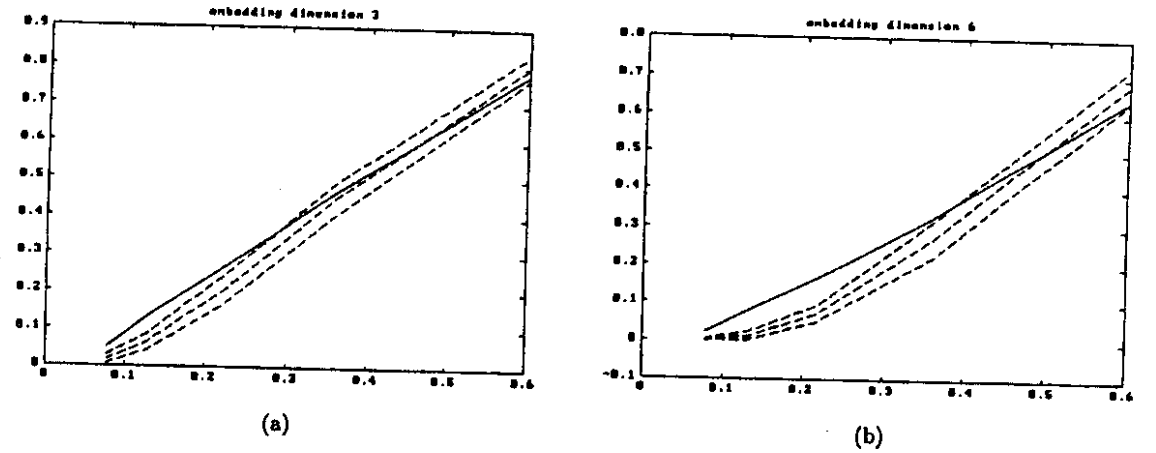
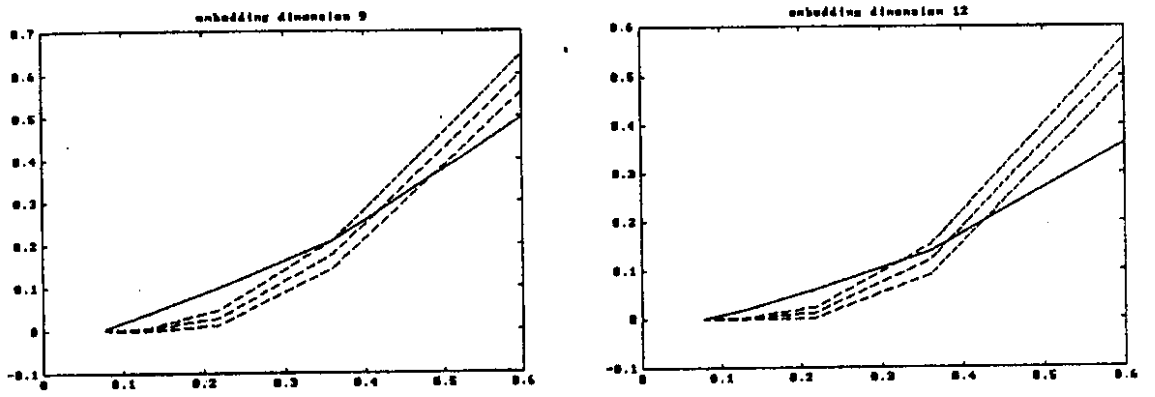
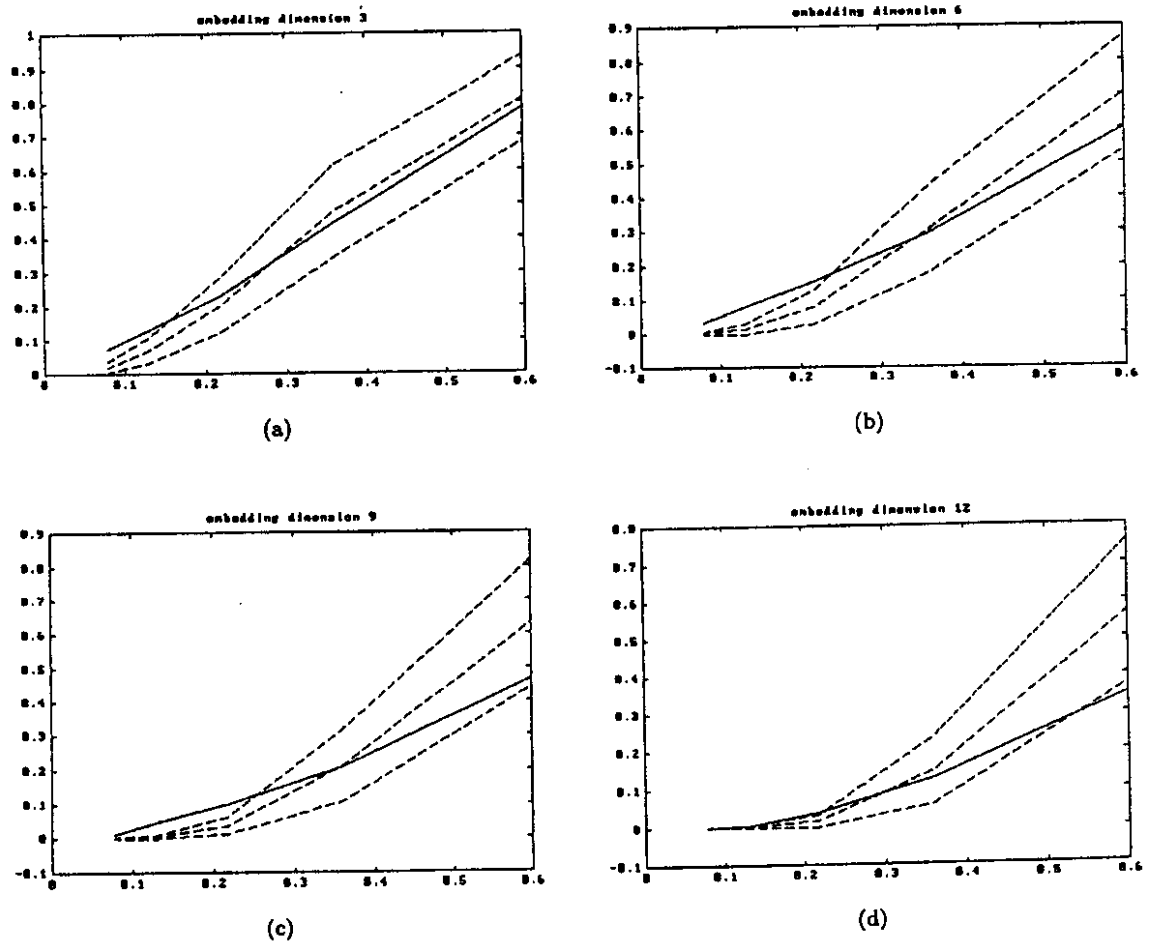


Fig. 3. Estimates of correlation integrals — the length of the reconstruction vectors is the embedding dimension k — horizontally we give the length scale ε (the time series were first rescaled in such a way that the difference between the minimal and maximal value of the original time series became one) vertically we give the estimated correlation integrals $P_k(\varepsilon)$: the solid line is for the original time series, the dashed lines for the randomized time series — the average estimate and the averages plus and minus twice the standard deviation; the randomized time series are here obtained by randomizing the phases of the Fourier coefficients.



(c) (d)

Fig. 3. (*Continued*)



(a) (b)

(c) (d)

Fig. 4. Same as the previous diagram, but using random time series obtained from a linear model, defined by an optimal linear predictor of order 20 and noise from the series of residues.

Finally we have to say something about the fact that we did our estimations only using 500 pairs each time. We know in general that for a series of N reconstruction vectors, we get the best estimate for the correlation integral when using all the $N(N-1)/2$ pairs of different reconstruction vectors. But we also know that even then the variance of the estimate is of the order $1/N$. In other words, these $N(N-1)/2$ pair are so strongly dependent that they really count as approximately N *independent* pairs. This is the reason that we expect that, when taking not all the pairs, but only N pairs, one does not lose very much. In any case, even if we would get a much better estimate when using all the pairs, if we conclude, as in the above case, that the linear model has to be rejected, there is no reason that such a conclusion has to be modified when using all pairs: using all pairs will make the variances smaller and hence will lead to an even stronger rejection. So only in a case where we could not reject the linear model, might we consider the much heavier calculations using all pairs of different reconstruction vectors to see whether even then the results do not indicate a difference.

Acknowledgments

I thank Dr. H. Dehling for informing me about statistical aspect of the problems discussed in this paper, in particular for introducing me to the ideas of the U-statistics.

References

Billingsley, P. [1968] *Convergence of Probability Measures* (Wiley).

Brock, W. A. & Dechert, W. D. [1988] "Theorems on distinguishing deterministic from random systems," in *Dynamic Economic Modelling*, Proc. of the Third International Symposium on Economic Theory and Econometrics, ed. Barnett, Berndt, White (Cambridge University Press).

Brock, W. A., Dechert, W. A. & Scheinkman J. A. [1987] "A test for independence based on the correlation dimension," *Dept. of Economics* (University of Wisconsin, Houston and Chicago).

Casdagli, M. [1991] "Chaos and deterministic versus Stochastic Nonlinear Modelling," *J. R. Statist. Soc.* **54**, 303–328.

Collet, P. & Eckmann J.-P. [1980] *Iterated Maps on the Interval as Dynamical Systems* (Birkhäuser).

Denker, M. & Keller G. [1986] "Rigorous statistical procedures for data from dynamical systems," *J. Stat. Physics* **44**, 67–93.

Grassberger, P. & Procaccia I. [1983] "Measuring the strangeness of strange attractors," *Physica* **D9**, 189–208.

Grasberger, P., Schreiber T. & Schaffrath C. [1991] "Nonlinear time series analysis," *Int. J. Bifurcation and Chaos* **1**, 521–547.

Hoeffding, W. [1948] "A class of statistics with asymptotically normal distribution," *Ann. Math. Stat.* **19**, 293–325.

Halsey, T. C., Jensen, M. H., Kadanov, L. P. & Procaccia I. [1986] "Fractal measures and their singularities: The characterization of strange sets," *Phys. Rev.* **A33**, 1141–1151.

Priestley, M. B. [1981] *Spectral Analysis and Time Series* (Academic Press).

Pijn, J. P. M. [1990] *Quantitative Evaluation of EEG Signals in Epilepsy, Thesis*, Amsterdam (Vakgroep Experimentele Dierkunde).

Packert, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw R. S. [1980] "Geometry from time series," *Phys. Rev. Lett.* **45**, 712.

Ruelle, D. [1976] "A measure associated with axiom A attractors," *Amer. J. Math.* **98**, 619–654.

Ruelle, D. & Takens, F. [1971] "On the nature of turbulence," *Comm. Math. Phys.* **20**, 167–192.

Scheinkman, J. A. & LeBaron, B. [1989] "Nonlinear dynamics and stock returns," *J. of Business* **62**, 311–337.

Sauer, T., York, J. A. & Casdagli, M. [1991] "Embedology," *J. Statistical Physics* **65**, 579–616.

Takens, F. [1981] "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence* (Warwick 1980, LNM **898**, Springer-Verlag).

Takens, F. [1983] "Invariants related to dimension and entropy," in *Atas do 13° colóquio bras. de mat.* (IMPA, Rio de Janeito).

Tong, H. [1990] *Nonlinear time series* (Clarendon Press).

Theiler, J., Galdrikian, B., Longtin, A., Eubank, S. & Farmer, J. D. [1991] "Using surrogate data to detect nonlinearity in time series," in *Nonlinear Prediction and Modeling* eds. Casdagli, M. & Eubank S. (Addison-Wesley).

Appendix

In this appendix, which is more technical than the main sections of the paper, we give mathematical details of a number of the arguments and results which we discussed in this paper.

A.1. *Estimating the correlation integral*

The correlation integral was defined in Sec. 3 as the limit

$$P_k(\varepsilon) = \lim_{N\to\infty} P_{N,k}(\varepsilon)\,,$$

where the $P_{N,k}(\varepsilon)$ were defined in terms of the k-dimensional reconstruction vectors. We shall first assume that these reconstruction vectors can be considered as independent and random samples from the distribution μ_k on $\mathbb{R}^k$. Simplifying the notation somewhat, we deal with the following situation. For some probability measure μ on $\mathbb{R}^k$, or rather on some metric space X with distance function ρ, we want to estimate the probability that two randomly and independently chosen points are within distance ε — this probability is the correlation integral. This estimate must be based on a set of N points, which we assume to be a random and independent sample for the probability distribution μ, and their mutual distances.

Estimation of quantities like these is treated in the theory of *U-statistics* as introduced by Hoeffding [1948]. We first present the results and then give a sketch of the arguments involved.

An unbiased and minimal variance estimate for the correlation integral, based on an independent sample of N points x_i as above, is given by the following formula

$$\hat{P} = \frac{2}{N(N-1)} \sum_{i<j} g(x_i, x_j)$$

where g is the function which is one if the distance between x_i and x_j is smaller than ε and zero otherwise. This is of course the same as $P_{N,k}(\varepsilon)$ as defined in Sec. 3.

The variance of $\hat{P}$, as a function of the size N of the sample is $(4/N)\mathrm{var}(g_1) + O(N^{-2})$, where g_1 is the function which assigns to each $x \in X$ the μ-measure of the ε-ball around x — $\mathrm{var}(g_1)$ denotes the variance of the function g_1 with respect to the measure μ. Note that with N points we have $N(N-1)/2$ distances. So if these distances could be considered as independent, the variance of $\hat{P}$ should be of the order of N^{-2}.

We recall here that we assumed that the points x_i formed a random and *independent* choice with respect to the probability distribution μ, and this assumption is usually *not justified* in our context of reconstructed data. At the end of this subsection, we discuss how a justified estimate for this variance can be obtained.

We deduce the above formula for the variance of our estimator $\hat{P}$. Let θ be the average, with respect to the measure μ, of the function g_1 as introduced above. Then the variance is the average of $(\hat{P}-\theta)^2$, where the average is taken over all the N-tuples of randomly and independently chosen samples from μ. Writing $\mathcal{E}$ for taking averages, we have

$$\begin{aligned}
\mathcal{E}((\hat{P}-\theta)^2) &= \left(\frac{2}{N(N-1)}\right)^2 \cdot \mathcal{E}\left(\sum_{i<j} g(x_i, x_j) - \theta\right)^2 \\
&= \left(\frac{2}{N(N-1)}\right)^2 \cdot \mathcal{E}\Bigg\{ \sum_{i<j,k<l,i\neq k\neq j,i\neq l\neq j} (g(x_i, x_j)-\theta)(g(x_k, x_l)-\theta) \\
&\quad + \sum_{i<j,i<k,j\neq k} (g(x_i, x_j)-\theta)(g(x_i, x_k)-\theta) + \sum_{i<j,k<j} (g(x_i, x_j)-\theta)(g(x_k, x_i)-\theta) \\
&\quad + \sum_{i<j,k<j,i\neq k} (g(x_i, x_j)-\theta)(g(x_k, x_j)-\theta) + \sum_{i<j,j<k} (g(x_i, x_j)-\theta)(g(x_j, x_k)-\theta) \\
&\quad + \sum_{i<j} (g(x_i, x_j)-\theta)^2 \Bigg\}.
\end{aligned}$$

Since θ is the average value of $g(x_i, x_j)$ for $i \neq j$, the first summand is zero. In the second to the fifth summand, there is always one repeating index. Now for fixed x_i the average (over x_j and x_k) of $(g(x_i, x_j)-\theta)(g(x_i, x_k)-\theta)$ is $(g_1(x_i)-\theta)^2$. Taking the average over x_i in the last expression we get $\mathrm{var}(g_1)$. The same holds if the positions of the repeating indices are different. So all the terms in the second to fifth summand

are on average equal to $\mathrm{var}(g_1)$. Rewriting these summands as

$$\sum_{i\neq j\neq k\neq i} (g(x_i, x_j) - \theta)(g(x_i, x_k) - \theta)$$

we see that there are $N(N-1)(N-2)$ terms. This gives a contribution to the total average which equals

$$\frac{4N(N-1)(N-2)}{(N(N-1))^2}(\mathrm{var}(g_1))$$

$$= \frac{4}{N}\mathrm{var}(g_1) + O(N^{-2}).$$

Finally, the last summand has $N(N-1)/2$ terms and hence only gives a contribution of the order $O(N^{-2})$. This completes the proof.

Now we come back to the problem that the points x_i may be dependent. This happens in particular if we consider reconstruction vectors of a time series which is obtained by oversampling a continuous time signal. Estimation of the variance in this situation was considered in Denker & Keller [1986], but the method described there is very time consuming and it is still only valid asymptotically for large values of N. In the present context there is a way out, making use of the Monte Carlo method, as noted in Theiler *et al.* [1991] i.e., computing the values of $\hat{P}$ for a number of different segments of length N of the same type series. This assumes that we can continue the time series as far as we want. For experimental time series, and in particular for economic time series this is not realistic, but for numerically generated time series this is not a problem.

It turns out that for the last two tests considered in Sec. 5, we compare with numerically generated time series for which the Monte Carlo method can be used. In fact in the two cases one has to proceed slightly differently. When we apply random phase shifts on the Fourier coefficients, we get a time series of the same length. We can however repeat this several times. When we use a linear model with the "noise" taken from the set of residues, we can easily continue the time series as far as we want. We discuss an example of such an analysis in Sec. 6.

A.2. *Optimal linear predictors*

In this subsection we show how autocorrelation coefficients and optimal linear predictors are related. We assume we have a time series $\{y_i\}_{i\geq 0}$, which we assume to be stationary. From this time series we construct a Hilbert space $\mathcal{H}$. The elements of this Hilbert space are equivalence classes of time series $\{\tilde{y}_i\}_{i\geq k}$ for some k. Sums (or differences) of such time series are obtained by adding (or subtracting) values with the same index i; two time series are equivalent if their difference has variance zero. We want to define the inner product of $\{\tilde{y}_i\}$ and $\{\tilde{y}'_i\}$ as $\lim_{i\to\infty} i^{-1}\sum_{j\leq i}\tilde{y}_j\tilde{y}'_j$. A problem is that this limit may not exist. This can be solved in the following way: We start with our original time series $Y = \{y_i\}$ which we assume is stationary. We then construct from this the time series $\sigma^j(Y)$, where σ is the shift operator — the ith element of $\sigma(Y)$ is y_{i+1}. Inner products of these time series exist because of the stationarity of Y, they are just the autocovariances of Y. Then we define $\mathcal{H}$ as the completion of the finite sums of scalar multiples of these $\sigma^j(Y)$'s.

From now on we assume that the average of the time series Y is zero and that its variance is one. We denote $\sigma^j(Y)$ by Y_j. Observe that the shift operator σ determines an isometry in $\mathcal{H}$ mapping Y_j to Y_{j+1}. All the Y_j are unit vectors and the inner products between them are the autocorrelation coefficients: $\langle Y_i, Y_j\rangle = \rho_{|i-j|}$, so the whole geometry of the configuration of vectors Y_0 up to Y_{-k} is determined by the autocorrelation coefficients ρ_1 up to ρ_k.

In this context the optimal linear predictor of order k,

$$\hat{y}_n = a_1 y_{n-1} + \cdots + a_k y_{n-k}$$

can be considered as an element $\hat{Y}$ of $\mathcal{H}$, which is a linear combination of Y_{-1} to Y_{-k}, with coefficients a_1 to a_k, and which is optimal in the sense that $||\hat{Y} - Y_0||^2$ is minimal. This means that we have to take $\hat{Y}$ equal to the projection of Y_0 on the linear space spanned by the vectors Y_{-1} to Y_{-k}. The variance of this linear predictor is the square of the distance of Y_0 to this linear subspace.

We now derive explicit formulas for the coefficients of the optimal linear predictor and its variance in terms of the autocorrelation coefficients.

$$||\hat{Y} - Y_0||^2 = \left\| Y_0 - \sum_i a_i Y_{-1} \right\|^2$$

$$= 1 - 2\sum_i a_i\rho_i + \sum_{i,j} a_i a_j \rho_{|i-j|}.$$

In order to find the optimal values of $a_1, \ldots, a_k$, we differentiate with respect to a_i and put the resulting

expression equal to zero:

$$-2\rho_i + 2\sum_j a_j \rho_{|i-j|} = 0 \text{ or } \rho = Aa,$$

where A is the matrix with i, jth element $\rho_{|i-j|}$ and where a and ρ are the vectors with components a_j, ρ_j, respectively. Using this notation we find by substitution the variance

$$||\hat{Y} - Y_0||^2 = 1 - \rho^T A^{-1} \rho.$$

This assumes that A is invertible, but this is automatically the case if $Y_0, \ldots, Y_{-k}$ are not linearly dependent, and that is what we shall assume — see below.

From the above considerations it is clear that the variance of the optimal linear predictor of order k, defined by $\hat{Y}$ is zero if and only if $\hat{Y} = Y_0$ is linearly dependent on the vectors Y_{-1} to Y_{-k}. But in that case, since the shift operator defines a linear map in $\mathcal{H}$, Y_1 also is linearly dependent on Y_0 to Y_{-k+1} and hence linearly dependent on Y_{-1} to Y_{-k}. Hence all the vectors Y_i are linearly dependent on Y_{-1} to Y_{-k}, the Hilbert space $\mathcal{H}$ is only finite dimensional and all the prediction problems are trivial. In what follows we assume this does not happen, so we assume that all the vectors Y_j are linearly independent. In fact in the case where we have linear dependence, the optimal predictors are not unique: the vector $\hat{Y}$ is still unique, but the coefficient a_1 to a_k which express $\hat{Y}$ as a linear combination of Y_{-1} to Y_{-k} are not unique.

Finally we observe that one can show that from an optimal linear predictor of order k and its variance, one can deduce the autocovariance coefficients $R_0, \ldots, R_k$. In fact these are the autocovariance coefficients of a time series produced by the linear model, defined by the optimal linear predictor.

A.3. *Chaos, predictability and autocorrelation*

We mentioned in Sec. 4 that there are time series, even stationary ones, which are chaotic, in our sense, but for which the optimal linear predictors are asymptotically perfect. A very simple example is the time series $\{y_n\}$ with $y_n = 0$ except for the case where n is 2^m for some integer m, in which case $y_n = 1$. Here the optimal linear predictor is $\hat{y}_n = 0$ and the variance of the errors is zero (the density of the values of n for which $y_n = 1$ is zero). In this example the Hilbert space $\mathcal{H} = \{0\}$. A similar example with $\mathcal{H}$ nontrivial and with nonzero variance, R_0 is obtained by adding a periodic signal, e.g., by adding $\sin(n)$ to y_n — this gives a time series which is chaotic according to our definition but for which the autocorrelation coefficients ρ_k do not converge to zero for $k \to \infty$. These examples may look very artificial, but there are rather simple dynamical systems, described by ordinary differential equations, which produce time series of the same nature. We describe such a system. This is an example with continuous time (since we use a differential equation); this is because the phenomenon is, to our knowledge, much more unusual in the case of deterministic systems with discrete time.

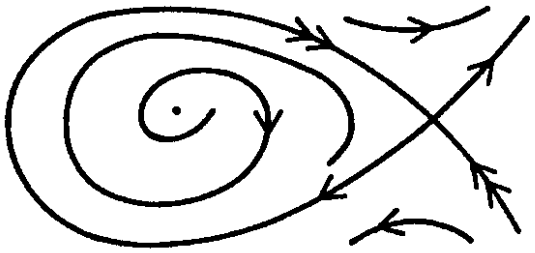


Fig. 5. Saddle with homoclinic loop.

On the plane $\mathbf{R}^2$, we consider a differential equation with a saddle point which has a homoclinic loop as in Fig. 5. We assume that at the saddle point the contracting eigenvalue is dominating (so that in the neighborhood of the saddle the divergence is negative). The homoclinic loop is attracting from one side. If we follow an orbit which is attracted to the homoclinic loop then we observe the following. The orbit spends most of its time in a small neighborhood of the saddle point and from time to time goes around the homoclinic loop. The successive times the orbits spend near the saddle grow exponentially. So here also, if we have to predict the position of the orbit, we always predict it to be in the saddle. Since the excursions along the homoclinic loop become rarer and rarer, the variance of the prediction error is zero.

This example is of course exceptional in the sense that a homoclinic loop is nongeneric, but still the example occurs in generic one parameter families.

A.4. *The deterministic structure in the residues*

We show, in a numerical example, how the deterministic structure becomes less recognizable in the

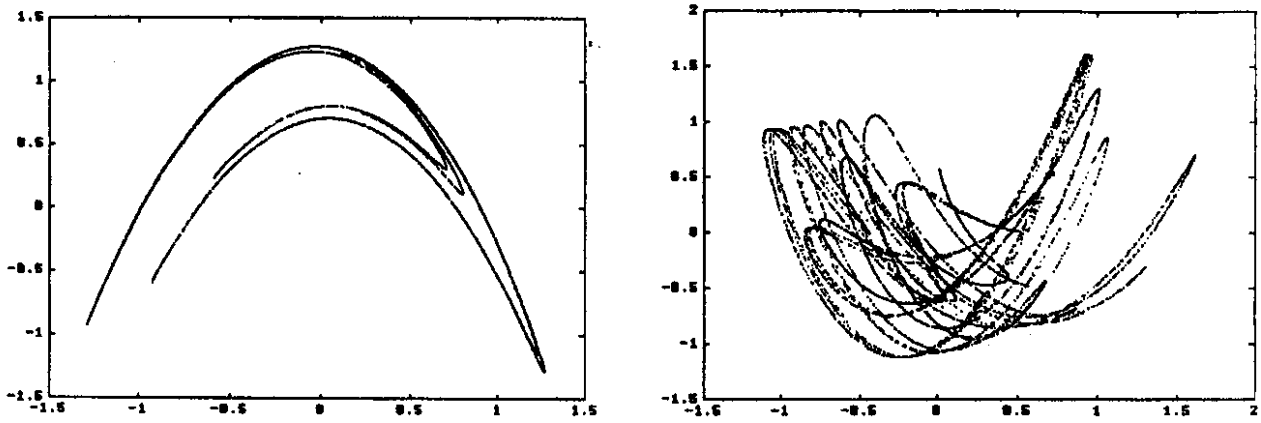


Fig. 6. Two-dimensional reconstructions of the Hénon map (a) and of its residuals (b).

residues. For this we take a time series of the Hénon map and the time series obtained by taking the residues from an optimal predictor of order 5. Below we show the two-dimensional reconstructions of these time series. It is clear that in the second case it is much harder to detect the fine-structure due to the deterministic origin from the correlation integrals $P_2(\varepsilon)$ for values of ε which are not very small. Similar pictures can be found in Theiler *et al.* [1991].