Testing for nonlinearity in unevenly sampled time series

Andreas Schmitz and Thomas Schreiber
Physics Department, University of Wuppertal, D-42097 Wuppertal, Germany
(Received 28 April 1998; revised manuscript received 9 December 1998)

We generalize the method of surrogate data of testing for nonlinearity in time series to the case that the data are sampled with uneven time intervals. The null hypothesis will be that the data have been generated by a linear stochastic process, possibly rescaled, and sampled at times chosen independently from the generating process. The surrogate data are generated with their linear properties specified by the Lomb periodogram. The inversion problem is solved by combinatorial optimization. [S1063-651X(99)04004-0]

PACS number(s): 05.45.−a

I. INTRODUCTION

The vast majority of methods of time series analysis deals with data measured at times that are an integer multiple of the fixed sampling interval Δ. Unevenly sampled time series are often excluded although they are quite common in cases where measurements are restricted by practical conditions. For example, most astronomical observations cannot be made in the daytime and must often be interrupted in the night due to cloudy weather; data from the stock exchange has gaps during the weekends and holidays, etc. The reason for excluding such data is mainly technical: most methods cannot be easily generalized to the unevenly sampled case. This is particularly true for nonlinear methods of time series analysis [1].

For the nonlinear approach to analyzing a time series, we first have to ask for signatures of nonlinearity in the generating process. In this paper for the first time we present a statistical test for nonlinearity in unevenly sampled time series by extending the common concept of surrogate data [2], where randomized data sets are used to obtain a Monte Carlo approximation to the probability distribution of a suitable test statistics.

In order to take a Monte Carlo approach to nonlinearity testing, we have to be able to generate sequences that are random except for their linear correlations. Any additional structure realized in these surrogate time series can lead to spurious positive results of the statistical test. In certain situations, the problem of varying time intervals can be circumvented by interpolating the data to equally spaced sampling times. However, in a test for nonlinearity, one could then not distinguish between genuine structure and nonlinearity introduced spuriously by the interpolation process.

Besides the generation of surrogates, we have to be able to measure the degree of nonlinearity in the data. In contrast to the process of generating surrogates, interpolations are permitted here as part of the specification of a test statistic. A badly designed test statistic could in the worst case lower the discrimination power of the test, while still keeping it formally correct. In this paper we use a very simple test statistic that measures nonlinearity through deviations from time reversibility. The main emphasis is laid on the generation algorithm for the surrogates.

II. SURROGATE DATA

The null hypothesis in this paper is that the data have been generated by a linear stochastic process that is measured instantaneously and possibly rescaled. Let $x(t)$ be the outcome of a linear stochastic process. The time series $\{y_n\}$ is, according to the null hypothesis, generated by

$$y_n = f(x(t_n)), \quad n = 1, \ldots, N \tag{1}$$

where $f(\cdot)$ is a monotonic function. This excludes correlations or even a deterministic relationship between the sampling times t_n and the observable y_n.

Generating surrogate data sets with the same linear properties as the original data amounts to the conservation of the autocorrelation function. But even simple things such as autocorrelations are hard to maintain for unevenly sampled data. Any time interval can occur between successive points and it is possible to combine them to nearly arbitrary lags. One idea is to calculate autocorrelations by binning all possible intervals to the desired, discrete lags, a process that involves some nonlinearity. Using these autocorrelations for generating surrogates can lead to the spurious rejection of purely linear time series.

Standard surrogate methods make use of the Fourier transformation to conserve the autocorrelations of the original data. The method of *amplitude adjusted Fourier transformation* (AAFT, [2]) rescales the original time series to a Gaussian distribution first. Then, the Fourier phases are randomized and the Fourier transformation is inverted. Finally, a rescaling to the original distribution is performed. A refined method has been suggested in Ref. [3] where an iteration scheme is used to simultaneously conserve the spectrum and the distribution. It consists of alternating Fourier transformation and rescaling steps. Both methods cannot directly be applied to unevenly sampled data, because they utilize the Fourier transformation and its inverse.

In Ref. [4], a general approach to the constrained randomization of time series is described that allows the specification of almost arbitrary properties. We will use this method to implement the power spectrum without explicit use of the inverse Fourier transform. Thus we may estimate the power spectrum by the Lomb periodogram.

III. LOMB PERIODOGRAM

Let $\{y_n\}$ be a time series sampled at times $\{t_n\}$ that need not be equally spaced. The power spectrum can then be estimated by the Lomb periodogram [5]. This spectral estimator is discussed, e.g., in Ref. [6]. Here we give the final formula:

$$P(\omega)=\frac{1}{2\sigma^2}\left\{\frac{\left[\sum_n (y_n-\overline{y})\sin\omega(t_n-\tau)\right]^2}{\sum_n \sin^2\omega(t_n-\tau)}+\frac{\left[\sum_n (y_n-\overline{y})\cos\omega(t_n-\tau)\right]^2}{\sum_n \cos^2\omega(t_n-\tau)}\right\} \quad (2)$$

where τ is defined by

$$\tan(2\omega\tau)=\frac{\sum_n \sin 2\omega t_n}{\sum_n \cos 2\omega t_n} \quad (3)$$

and $\overline{y},\sigma^2$ are the mean and the variance of the data, respectively. The result can be derived by fitting a least squares model $y=a\cos\omega t+b\sin\omega t$ to the data for each given frequency ω. Therefore, Lomb periodograms are often referred to as *least squares periodograms*. For time series sampled at constant time intervals, $\Delta=t_n-t_{n-1}$ for all n, the Lomb periodogram $P(2\pi n/N\Delta)$ yields the standard squared Fourier transformation. Except for this particular case, there is no inverse transformation for the Lomb periodogram, which makes it impossible to use the standard surrogate data algorithms mentioned above.

IV. GENERAL CONSTRAINED RANDOMIZATION

In order to avoid inverting the Fourier transform, we follow the general approach of Ref. [4], where desired properties on the surrogates are formulated by constraints. These constraints are implemented as a cost function $E(\{y_n\})$ which is constructed to have a global minimum if the constraint is fulfilled. In our case, the constraint is given by the Lomb periodogram (2) of the data. This can for example be expressed as a cost function by

$$E=\left[\sum_{k=1}^{N_f}|P(k\omega_0)-P_{\text{data}}(k\omega_0)|^q\right]^{1/q}, \quad (4)$$

which is the discrepancy between the desired Lomb periodogram P_{data} of the original data and the actual periodogram P of the surrogate. We calculate P at N_f equally spaced frequencies $k\omega_0$, but other choices are possible. With the parameter q one can specify the distance measure between the two periodograms. For $q=2$, the L^2 distance is used. Higher ''penalties'' for large differences in single frequencies could be given by raising q above two. For the case $q\to\infty$ only the largest difference contributes to E and we get the maximum norm. We use $q=1$ throughout, which yields the averaged absolute difference of P_{data} and P. This choice is motivated by the fact that the power is already a squared quantity. As a further alternative, one could use the differences between the square roots or logarithms of P, which puts less stress on the peaks in the power spectrum than Eq. (4). Another freedom lies in the choice of the minimum frequency ω_0 and the number of frequencies N_f and one may have to consider different values for each individual application.

As in Ref. [4] we will look for the minima of this cost function (4) among all permutations of the time series $\{y_n\}$. Combinatorial minimization by complete enumeration is not feasible here since the computational effort grows exponentially with the length of the time series. Instead, we will use the method of simulated annealing [7,8] that is expected to find an approximate solution in polynomial time.

V. SIMULATED ANNEALING

The simulated annealing proceeds as follows. Starting with a random permutation of the original time series, the surrogate is successively modified by exchanging two values y_i and y_j with i,j chosen at random. Let the cost function before the modification be E_{old}, and after the exchange E_{new}. The modification will be accepted if it yields a lower value for the cost function, or else with a probability

$$p=\exp(-\Delta E/T), \quad \Delta E=E_{\text{new}}-E_{\text{old}}. \quad (5)$$

Otherwise it will be rejected and a different pair is selected for modification. Using this updating scheme has been proposed by Metropolis *et al.* [9] as a method to keep a model system in equilibrium at a given system temperature T. For minimization, the ''temperature'' is lowered slowly in order to reach the ground state that is given by the global minimum of E. For the present purpose we do not have to reach the proper global minimum. A state with a small but finite E will be sufficient.

Simulated annealing has a rich literature that will not be reviewed here. An introduction can be found, for example, in [10]. Although some rigorous convergence results are available, in a given application it is very difficult in general to give an optimal scheme of lowering T. Here we use a cooling scheme as proposed for example in [6]. The temperature is lowered by a constant factor $\alpha<1$ to αT after N_{total} considered modifications or after $N_{\text{acc}}<N_{\text{total}}$ accepted updates. The parameters N_{total} and N_{acc} are chosen to be proportional to N. By using higher values for the parameters N_{total}, N_{acc}, and α, the cooling can be made slower. Slower cooling in general yields lower final values of E, i.e., higher accuracy of the periodogram, at the expense of computational time.

Two improvements that accelerate the annealing algorithm have been made. The first is to choose the two points that are candidates for an exchange with a probability that depends on their difference in magnitude, respectively, in rank. Let rank(y_i) be the position of y_i in the sorted array, going from 1 for the smallest to N for the largest value of the time series. Exchanging two points with a big difference in their rank (e.g., the smallest and the largest value) generally yields a larger change of the cost function E than exchanging

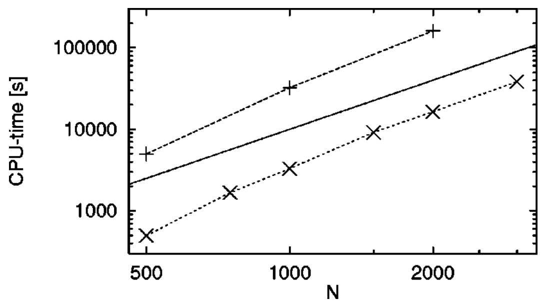

FIG. 1. CPU time used to generate one surrogate of length N. Time is given for a DEC alpha work station at 166 MHz (top) and a Pentium Linux PC at 450 MHz (bottom). The solid line is $\propto N^2$.

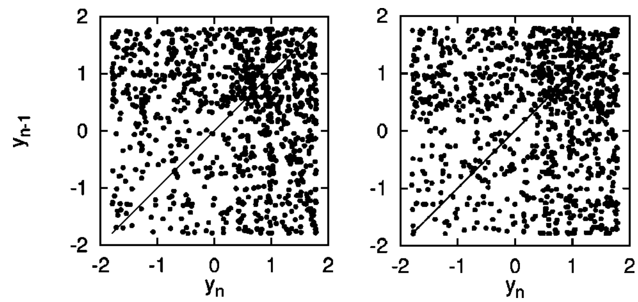

FIG. 2. Delay plot of an unevenly sampled Hénon map (left) and one surrogate (right). The test finds a significant difference in time asymmetry.

two points that do not differ much in their ranks. For good performance, it is desired to keep approximately the same acceptance rate through all temperatures. This can be achieved by choosing pairs of points $\{y_i, y_j\}$ with $i \neq j$ and probability $p_{ij}(d, \mu)$ where $d = |\text{rank}(y_i) - \text{rank}(y_j)| - 1$. The probability p is chosen to have a maximum for $d = 0$ and to decrease for higher d. The parameter μ characterizes the ''width'' of the distribution p and should be varied proportional to N/T. The exact shape of p does not seem to be of much importance. For example, we were not able to observe a significant difference in performance between exponential and Gaussian distributions. But in all cases we considered, a nonuniform p_{ij} of width $\propto N/T$ substantially accelerated the annealing process, so that higher accuracy is reachable with the same computational effort.

Calculating E is very time consuming for long time series and many frequencies. With a typical value of $N_f \propto N$ we have an algorithm of order N^2 for each annealing step. Additionally, the number of annealing steps is expected to grow at least linearly with N. For our applications, it is not necessary to recalculate all sums in Eq. (2) for every exchange, because we only change the values y_n while fixing the times t_n and frequencies $k\omega_0$. In Eq. (2), τ and the two denominators do not depend on y_n and can be stored in arrays for every frequency $k\omega_0$ at the beginning of the annealing process. The sums in the numerator do not change much either and only the two terms that correspond to pair to exchange $\{y_i, y_j\}$ have to be subtracted, recalculated and added again. This reduces the effort for the update of the Lomb periodogram to order N_f.

But even with the described modifications to the algorithm, annealing is quite computer time intensive. The CPU time used to generate one surrogate is shown in Fig. 1 for subsets of different lengths N of time series E (see example below). We calculated the Lomb periodogram at $N_f = N/2$ frequencies and used $q = 1$ in the cost function. As indicated by the solid line the whole algorithm is found to be of order N^2. Considering that the update of the cost function is of order N, the annealing scheme itself seems to be of order N. In any case, this is much faster than complete enumaration with an order exponential in N.

For the calculation of surrogates, simulated annealing is performed until E has fallen below a given value E_f, the desired accuracy of the Lomb periodogram. In Monte Carlo simulations of physical systems new configurations are usually derived from previous configurations. In order to avoid any correlations between the different surrogates, we prefer to start with a completely random permutation of the original time series for each surrogate. The starting temperature can roughly be determined by calculating the cost function for some randomly shuffled data sets and choosing T_0 as the average difference in cost between them. Once an adequate starting temperature is known, it can be used for further surrogates.

VI. TEST STATISTICS

So far, we have described how to produce randomized versions of unevenly sampled time series with given linear correlations, which is the main point of this paper. Let us now demonstrate how such surrogate sequences can be used in tests for nonlinearity. For this purpose, we have to be able to measure the degree of nonlinearity in a time series. Many statistics that have proven useful for evenly sampled time series (see, e.g., [11]) cannot easily be generalized to unevenly spaced data. This generalization, in general, is a topic of future research.

Here, as a first simple statistic, we choose a measure for time reversibility, which is a good indicator for nonlinearity. It is however not very enlightening about what source of nonlinearity there might be. For the data sorted in time order,

$$\gamma = \frac{1}{(\sigma^2)^{3/2}(N-1)} \sum_{n=2}^{N} \left(\frac{y_n - y_{n-1}}{t_n - t_{n-1}} \right)^3 \tag{6}$$

is calculated, which is just the mean of the slopes, taken to the third power. For a time series generated by a linear process, and for the surrogates, we expect $\gamma \approx 0$. In contrast, time series with nonlinearities can be asymmetrical in time and may yield values of $\gamma \neq 0$. To pay regard to deviations in both directions ($\gamma > 0$ and $\gamma < 0$), a *two-sided* test [12] has to be performed.

VII. EXAMPLES

To test the functionality of the surrogate test, we use 10 000 points of the Hénon map as a first example. From these, we pick $N = 1000$ points with their time indices chosen randomly. To generate surrogates, we calculate the Lomb periodogram for $N_f = 500$ frequencies in the interval $\nu \in [0, 0.5]$. A delay plot of the data and one surrogate is shown in Fig. 2. A little trace of the Hénon attractor can be found in the left figure that is built by pairs of values with time delay one. For the original time series we get

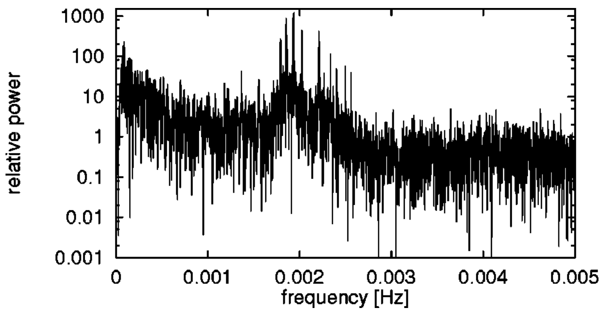

FIG. 3. Lomb periodogram of data set *E*. See text for details.

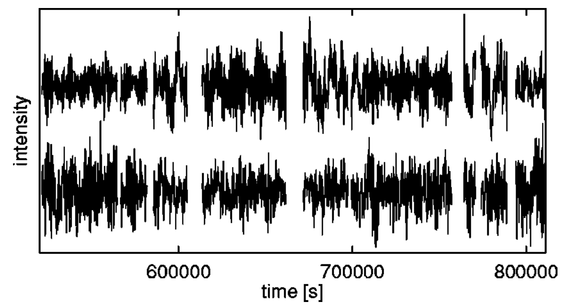

FIG. 4. The down-sampled data set *E* with one corresponding surrogate. Gaps of different sizes prevent reasonable interpolation.

$\gamma=-0.58$, while 19 surrogates gave values $-0.30<\gamma<0.18$. This corresponds to a 90% level of significance for the time series not being time reversible and hence nonlinear in the sense of the null hypothesis.

In order to verify the new test method we also applied the test to linear time series based on an AR(1) process $x_{n+1}=0.95x_n+\eta_n$ which is, however, invertibly rescaled by $y_n=x_n\sqrt{x_n}$. Again we picked 1000 points randomly from the original 10 000. A test performed with this data was unable to reject the null hypothesis, as expected since the null was true.

For a quantification of the power and the size [12] of the new method, many independent tests would be necessary. Such an extremely computer time-intensive study is beyond the scope of the present work.

Finally, the test is applied to experimental data. Data set *E* of the Santa Fe time series contest is a set of measurements of the time-integrated intensity of light observed from a variable star. It consists of 17 parts with different numbers of points, the time range of which partly overlaps and partly shows gaps. Inside the blocks, the data is evenly sampled with $\Delta=10$ s. Special interest in low frequency components makes it desirable to consider the time series as a whole. The Lomb periodogram of the data set is shown in Fig. 3.

For the surrogate test we further down-sampled the data by integrating over 12 successive measurements. Therefore, surrogates could be generated in reasonable time. The resulting time series is $N=2260$ points long with $\Delta=120$ s except for nine gaps taking up to 10 000 s, as shown in Fig. 4. The Lomb periodogram is calculated at $N_f=1130$ frequencies with up to $\nu_{max}=1/240$ Hz. The value for the time-reversibility statistic $\gamma=-0.56\times10^{-7}$ s^{-3} of the considered time series does not lie outside the interval $\gamma\in[-13\times10^{-7}$ s^{-3}, 29×10^{-7} s$^{-3}]$ spanned surrogates, and thus the null hypothesis cannot be rejected. One surrogate time series is also shown in Fig. 4.

Spurious high-frequency components could be introduced by discrepancies in the overlapping parts of the recording. To deal with that problem we deleted points from one of the two parts and repeated the test. No significant differences in the surrogates and its values for the test statistics were observed.

Taking a closer look at the individual parts shows considerable differences in their autocorrelations, which makes it dangerous to consider the whole data set as stationary. In contrast, the generated surrogates are stationary by construction. If one could detect significant differences with a nonlinear statistic, nonstationarity would be an equally likely explanation as nonlinearity.

VIII. SUMMARY

In this paper we presented a method for a test for nonlinearity for time series with uneven time intervals. Such a test consists of two main steps: generating surrogate data and calculating test statistics. The new method is able to achieve the first step using the constrained randomization scheme proposed in Ref. [4]. We offered only a first attempt on the second problem. More powerful test statistics are likely to be derivable from current nonlinear time series methods.

We would like to thank Daniel Kaplan, James Theiler, Peter Grassberger, and Holger Kantz for useful discussions. This work was supported by the SFB 237 of the Deutsche Forschungsgemeinschaft.

[1] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997).

[2] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).

[3] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[4] T. Schreiber, Phys. Rev. Lett. **80**, 2105 (1998).

[5] N. R. Lomb, Astrophys. Space Sci. **39**, 447 (1976).

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, 2nd ed. (Cambridge University Press, Cambridge, 1995).

[7] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[8] S. Kirkpatrick, J. Stat. Phys. **34**, 975 (1984).

[9] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1097 (1953).

[10] R. V. V. Vidal, *Applied Simulated Annealing*, Lecture Notes in Economics and Mathematical Systems Vol. 396 (Springer, Berlin, 1993).

[11] T. Schreiber and A. Schmitz, Phys. Rev. E **55**, 5443 (1997).

[12] J. Theiler and D. Prichard, Physica D **94**, 221 (1996).