Change of the nature of a test when surrogate data are applied

Enno Mammen* and Swagata Nandi†
Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany

(Received 8 May 2003; revised manuscript received 1 December 2003; published 30 July 2004)

Surrogate data is a well-known method in nonlinear time series analysis and it has been widely used in testing nonlinearity. Fourier transform-based surrogates are artificially generated time series which share the linear properties of the observed series. They can be used for the generation of critical values for test statistics. In this paper we will show that the variance of these critical values may be of the same order as the variance of the test statistic itself. This changes the nature of the test because the test rejects if the test statistic divided by the critical value exceeds 1. An example is a test for normality that checks higher-order empirical cumulants. We will show that such a test is transformed to a test on (circular) stationarity.

DOI: 10.1103/PhysRevE.70.016121 PACS number(s): 05.10.Ln, 02.70.Uu

I. INTRODUCTION

Surrogate data testing is by now a well-established method in nonlinear time series analysis. In this paper we will discuss the application of surrogate data to testing. We will show that the nature of a test may change drastically by the application of surrogate data. After the application of surrogates, use of a test statistic that measures a certain type of deviation from the null hypothesis may result in a test that looks for a quite different type of alternative. This happens because surrogate data tests depend on two random quantities: on the test statistic and on the quantiles estimated by the surrogate data. In most other resampling methods, the estimated quantiles asymptotically stabilize. Then, the stochastic performance approximately depends only on the stochastic distribution of the test statistic. But we will see that in the case of surrogate resamples, this does not hold. This is the main point of this paper. In particular, we will argue that it is not sufficient only to study the distribution of the test statistic to analyze surrogate data tests. One has to consider the joint distribution of the test statistic and the surrogate critical value to understand when the test statistic is larger than the surrogate critical value. In this paper we will do this for a series of different test statistics.

The term "surrogate data" was first introduced by [1], but the basic idea appeared in a number of earlier publications (see [2–4]). The method in [3] is called multivariate scaling analysis. The method of surrogate data is a resampling method. Additional data are artificially generated and they are used to get an impression of a test statistic on a hypothesis. The classical algorithm is based on Fourier transform (FT) of the data. One computes the discrete Fourier transform (DFT) of the data, randomizes the phases, and then inverts the transform. In the literature, this method is also known as phase scrambling or Fourier bootstrap. We call it FT-based surrogates or simply surrogates. A discussion can be found in Sec. II. The most prominent application of surrogate data is testing for nonlinearity of a time series. The basic idea is to use surrogates for generating random data that only share linear properties with the observed time series, and to compare measures of nonlinearity for the generated random data and the observed process. It is then argued that if there is no significant difference, then there is no reason to reject linear stochastic modeling of the data. Such findings can be used to discuss whether the data are generated by some chaotic system. Instead of considering the hypothesis that the underlying dynamics is chaotic, it has been formulated the other way around. The null hypothesis is that the data are generated by some linear stochastic process. In particular, the more restrictive null hypothesis, that the data follow a Gaussian-linear stochastic process, has been considered by several authors. For such a null hypothesis, different test statistics have been used. Typically, test statistics have a direct interpretation, e.g., they may measure a distance from the null hypothesis by some empirical quantity. If this quantity significantly differs from 0, then the null hypothesis is rejected. When the distribution of the quantity is constant over the null hypothesis, then the null hypothesis is rejected by a level α test if the quantity exceeds the $1-\alpha$ quantile of this distribution ("the critical value"). If a test statistic is nonpivotal, its distribution varies over the null hypothesis. Then a classical approach is to approximate this distribution by a normal distribution with estimated variance. Another more recent approach is based on resampling. Artificial data are generated that (approximately) follow a fitted stochastic model from the null distribution, and the distribution of the test statistic is approximated by the distribution of the artificial data set. It is proved in [5] that surrogate data resampling leads to valid tests for the case of circular stationary Gaussian processes: for all choices of test statistics one gets a test with exact level α. This result has been used to argue that for (noncircular) stationary processes one achieves approximately correct levels for surrogate tests if the test statistic does not heavily depend on boundary values of the time series (see [5]). We will give a short outline of the basic arguments in Sec. III. However, we will see that these results do not imply that surrogate data consistently estimate the critical values of the test statistic. A surrogate data test strongly differs from the test that rejects if the test statistics exceeds its critical value. A detailed description and discussion of this fact can be found in Sec. IV. This will be done by

*Email address: emammen@rumms.uni-mannheim.de
†Email address: sw_nandi02@yahoo.co.in

simulations and some asymptotic calculations. In Sec. V, a detailed technical discussion on two test statistics is given. Concluding remarks will be made in Sec. VI.

II. SURROGATE DATA

The method of surrogate data was first suggested by [1]. They proposed this method to check the statistical significance of a test statistic. Two distinct approaches, the typical realization approach and the method of constrained realization, were compared in [6]. The typical realization approach is similar to traditional bootstrap methods, where explicit model equations are fitted from the data and are used for the generation of resamples. The constrained realization method was proposed for the implementation of test procedures. In this approach the fitting of model equations is avoided. The principal idea is to generate data that are consistent with the hypothesis under consideration.

We now shortly review the method of generating surrogates based on Fourier transforms. Suppose a data vector $\mathbf{X_N}=(X_1,\ldots,X_N)$ is generated by a stationary Gaussian process. In this paper, following statistical language, we always understand a stationary Gaussian process as a stationary *linear* Gaussian process. A process is called Gaussian if its full dimensional distribution (and not only the marginal distribution of an amplitude at one fixed time point) is Gaussian. Then, stationarity and Gaussianity imply linearity, see Sec. 5.7 in Brockwell and Davis [7]. So, linearity need not be explicitly mentioned.

The periodogram function $I_{\mathbf{X}}(\omega)$, in terms of DFT, is defined as follows:

$$I_{\mathbf{X}}(\omega) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} X_t \exp(-i\omega t)\right|^2 = |\zeta_{\mathbf{X_N}}(\omega)|^2. \qquad (1)$$

Given the sample size N, the DFT at the Fourier frequency $\omega_j=2\pi j/N, j=1,\ldots,N$ can be written as (in polar form) $\zeta_{\mathbf{X_N}}(\omega_j)=\sqrt{I_{\mathbf{X}}(\omega_j)}\,\exp(i\theta_j)$, where θ_j is the phase. Then, using inverse FT and some simple algebra, X_t's can be recovered from DFT's as

$$X_t = \overline{X} + \sqrt{\frac{2\pi}{N}}\sum_{j=1}^{m} 2\sqrt{I_{\mathbf{X}}(\omega_j)}\,\cos(\omega_j t + \theta_j), \quad t=1,\ldots,N, \qquad (2)$$

for odd N and when N is even

$$X_t = \overline{X} + \sqrt{\frac{2\pi}{N}}\sum_{j=1}^{m} 2\sqrt{I_{\mathbf{X}}(\omega_j)}\cos(\omega_j t + \theta_j) + \sqrt{\frac{2\pi}{N}I_{\mathbf{X}}(\omega_{N/2})}\cos(\pi t + \theta_{N/2}), \quad t=1,\ldots,N. \qquad (3)$$

Here we define $m=(N-1)/2$ when N is odd and $m=(N-2)/2$ when N is even.

Surrogate data are generated by replacing the phases $\theta_1,\theta_2,\ldots$ by random values $\theta_1^*,\theta_2^*,\ldots$ in (2) and (3). Here $\theta_1^*,\ldots,\theta_m^*$ are independent and identically distributed $U[0,2\pi]$ and independent of $\theta_{N/2}^*$ (when N is even), which is 0 or π with a probability of 0.5.

By construction, the surrogate data $\mathbf{X}^*$ preserves the observed sample mean and periodogram values, that is

$$\overline{X^*} = \overline{X}, \quad I_{\mathbf{X}^*}(\omega_j) = I_{\mathbf{X}}(\omega_j), \quad j=1,\ldots,N.$$

It also preserves the circular autocovariances given by

$$\frac{1}{N}\sum_{t=1}^{N}(X_t^* - \overline{X^*})(X_{t+k}^* - \overline{X^*}) = \frac{1}{N}\sum_{t=1}^{N}(X_t - \overline{X})(X_{t+k} - \overline{X}) = r_{k,c},$$

where $X_{t+N}=X_t,\ X_{t+N}^*=X_t^*$.

The surrogate data are conditionally circular stationary [given the original sample $\mathbf{X_N}=(X_1,\ldots,X_N)$]. A process $\{X_t\}$ is circular of index N, if $X_{j+kN}=X_j$ for each $1\leqslant j\leqslant N$ and positive integer k. This implies that

$$E^*(X_t^*) = \overline{X}$$

and for $1\leqslant q\leqslant p\leqslant N$,

$$\mathrm{Cov}^*(X_p,X_q) = E^*(X_p^* - \overline{X})(X_q^* - \overline{X}) = r_{p-q,c}.$$

Here E^* denotes the conditional expectation given the original sample. Similarly, Cov^* denotes the conditional covariance.

The finite-dimensional conditional distributions of FT-based surrogates are asymptotically Gaussian (with autocovariances equal to the observed circular autocovariances and with mean equal to the empirical mean of the observed process). This follows under mild conditions on the spectral density function and by standard asymptotic arguments using (2) and (3), with θ_j replaced by θ_j^*. The surrogates have by definition a symmetric distribution. This implies that all expected odd-order central moments of the surrogate series vanish, as is the case for the original data set. However, higher-order even central and noncentral moments are not preserved and they do not match with the corresponding empirical moments of the observed series. For more results on higher-order moments, see [8].

III. ACCURACY OF LEVELS OF SURROGATE TESTS

There were some discussions on the validity of surrogate data tests. Asymptotics for the distribution of surrogate data were studied by [9]. In [10], (FT-based) surrogate data were compared with some other resampling schemes for time series. Higher-order moments and cumulants of surrogate data for a wide range of time series models were compared in [8]. It was done for the standard method of phase randomization [1] and for the rescaling method [10]. These results were applied to develop diagnostic tests to check convergence of Markov-Chain Monte Carlo algorithms. The power of surrogate data tests was discussed in [11] and [12], and it was argued that rejections may be caused by the only reason that the assumption of stationarity is violated. Our paper is an attempt to present detailed discussions of these findings. The validity of the method of (FT-based) surrogate data testing was considered in [5]. There, it was shown that surrogate data tests achieve exact levels for circular stationary Gaussian processes. In this section we expose his approach and discuss some conclusions. We argue that the surrogate data method is the only valid method for the full model of circular stationary Gaussian processes and that there exist alternative resampling schemes only for more restrictive models. We conjecture that under suitable conditions these findings carry over asymptotically to noncircular models. A circular process is Gaussian if $\mathbf{X_N}=(X_1,\ldots,X_N)$ is a multivariate Gaussian. A Gaussian circular process of index N is stationary if the X_t's have an identical mean and the covariance matrix of $\mathbf{X_N}$ is a circular matrix.

For circular stationary Gaussian processes surrogate data tests achieve the correct level. This holds for all test statistics. The rejection probability is constant and equal to the level. Such tests are called similar. On the other hand, the following controversial statement holds: If for a fixed test statistic the critical values should be chosen, such that the test becomes similar (i.e., has constant rejection probability), then the only way to achieve this aim is by calculating the critical values by the method of surrogate data. These two results are due to [5]. The main step to prove these results is the following fact: For a Gaussian circular stationary process, sample mean and circular sample autocovariances (or equivalently sample periodogram) are sufficient statistics. This means that the conditional distribution of the process given these statistics is fixed and does not depend on the parameters of the process. Furthermore, this conditional distribution is the distribution of surrogate data. For an explanation on why these two statements imply the results of [5], we briefly recall some facts from the theory of similar tests.

Let α be the size of a test $\phi(X)$, where X is the vector of observations. The test is similar if $E_\theta\,\phi(X)=\alpha$ for all $\theta\in\Theta_0$, where Θ_0 is the set of parameters on the null hypotheses. Similar tests can be easily constructed if a sufficient statistic S is available. Let $\mathcal{P}^X=\{P_\theta,\theta\in\Theta_0\}$ be the family of distributions of X on the hypotheses. Then the conditional distribution of X given S does not depend on the underlying parameter $\theta\in\Theta_0$ because S is sufficient. In particular, $E[\phi(X)|S=s]$ does not depend on θ. Then any test satisfying

$$E[\phi(X)|S=s]=\alpha \tag{4}$$

(except on a set of probability measure zero) is similar on $\mathcal{P}^X$. This immediately follows from

$$E[\phi(X)]=EE[\phi(X)|S]=\alpha.$$

A test satisfying (4) is said to have Neyman structure with respect to S.

Let us consider a test statistic $T=f(X)$ where now, on the null hypotheses, X is a circular stationary process. The basic idea of surrogate tests is to compare T with $T^*=f(X^*)$, where X^* are surrogate data. We now choose S as the tuple of sample mean and sample circular autocovariances. On the hypothesis of circular stationary processes this is a sufficient statistic and, given S, the statistics T and T^* have the same conditional distribution, see above. For a given $S=s$, choose now $k_\alpha(S)$ such that

$$P[T^*\geqslant k_\alpha(S)|S=s]=\alpha.$$

Then

$$\begin{aligned}P_\theta[T\geqslant k_\alpha(S)]&=E_\theta\{P_\theta[T\geqslant k_\alpha(S)|S]\}\\&=E_\theta\{P[T\geqslant k_\alpha(S)|S]\}\\&=E_\theta\{P[T^*\geqslant k_\alpha(S)|S]\}\\&=\alpha.\end{aligned}$$

Thus, surrogate tests achieve a correct level for all test statistics.

Suppose now that one wants to have a test with constant level α on a subset $\Theta_0^*\subset\Theta_0$ of the null hypothesis, i.e.,

$$P_\theta(T>k_\alpha)=\alpha\quad\text{for all }\theta\in\Theta_0^*.$$

Then

$$E_\theta\{P_\theta[T>k_\alpha|S]\}=P_\theta(T>k_\alpha)=\alpha\quad\text{for all }\theta\in\Theta_0^*.$$

Write now $u(S)=P_\theta[T>k_\alpha|S]$. Note that $u(S)$ does not depend on θ because S is a sufficient statistic. Thus we have that

$$E_\theta u(S)=\alpha\quad\text{for all }\theta\in\Theta_0^*.$$

If the family of distributions of S (for $\theta\in\Theta_0^*$) is "rich enough," this implies that the function u is constant and equal to α. We now give a sufficient condition on Θ_0^* for this implication. Suppose that the parameter θ is given by the mean μ of X_t and by the autocovariances $\gamma(k)=E[(X_t-\mu)\times(X_{t+k}-\mu)]$ for $0\leqslant k\leqslant N/2$. Then it can be shown that the implication holds if Θ_0^* contains a nondegenerate rectangle. Suppose that this holds. Then $P(T^*>k_\alpha|S)=P(T>k_\alpha|S)=u(S)=\alpha$. Thus the only way to calculate exact critical values of a test statistic with a constant level is given by the use of surrogate data. So, at first sight, there seems to be no alternative to surrogate data. However, there are alternatives if we relax our assumption on the level of accuracy. If we only require that the level is asymptotically equal to α and that this asymptotic relation only applies for a subclass of short-range-dependent processes we conjecture that a much more rich class of resampling methods has asymptotically correct levels. Note that for short-range-dependent processes $\gamma(k)$ converges to 0 exponentially for $k\to\infty$. This violates the condition that Θ_0^* contains a nondegenerate rectangle.

In this paper we discuss if the randomness of the surrogate data quantile $k_\alpha(S)$ changes the nature of a test. We will give examples where this is the case and where this does not happen. We conjecture that a discrimination between these two cases could be based on the check if T is (asymptotically) pivotal. A test statistic is called pivotal if its distribution does not depend on the underlying model parameter. By construction, surrogate data have the same unconditional distribution as the original sample. Therefore, also the pivotal test statistic calculated for the surrogate data has the same unconditional distribution (not depending on the model parameter). The distribution of the sufficient statistic S depends on the parameter θ. Let us denote this distribution by P_θ^S. Furthermore, we write P_θ^T for the distribution of T. If T is pivotal, $P^T=P_\theta^T$ does not depend on θ. The conditional distribution of T given S defines a Markov kernel K that does not depend on θ because S is sufficient. We can write $KP_\theta^S=P^T$ for parameter θ. Suppose now that the family P_θ^S is "rich enough." Then all elements of the family are mapped onto the same measure P^T. Typically, this only holds if K is degenerated, i.e., the conditional distribution of T given $S=s$ does not depend on s. Then $k_\alpha(S)$ also does not depend on S, i.e., $k_\alpha(S)$ is nonrandom. On the other side if T is nonpivotal, P_θ^T depends on θ and it cannot be that K is degenerated. Then at least for some α, the quantile $k_\alpha(S)$ must depend on S. These considerations motivate the conjecture that the change of the nature of a test by surrogate data is moderate in case of approximately pivotal test statistics.

IV. SOME TESTS AND THEIR SURROGATE VERSIONS

This section contains our major findings. We will discuss how the nature of a test changes due to the application of surrogate data critical values. We will do this for a class of test statistics. We start by considering the following class of circular processes:

$$X_t = A + c\sum_{j=1}^{m}\sqrt{B_j^2 + C_j^2}\,\cos(\omega_j t + \theta_j), \quad t = 1,\ldots,N, \quad (5)$$

where $A \sim N(0,\sigma^2)$, $B_j, C_j \sim N(0,\sigma_j^2)$, $\theta_j \sim U[0,2\pi]$ and A, B_j, C_j, and θ_j (with $j=1,\ldots,m$) are independent. Furthermore, $\omega_j = 2\pi j/N$, $c=\sqrt{2\pi/N}$, $m=(N-1)/2$, if N is odd and $m=(N-2)/2$, if N is even. Note that for N odd all circular stationary Gaussian processes with mean zero can be represented in the form (5) [see Eq. (2)]. For even N the last additive term in (3) is put equal to zero. Typically, this would result in an asymptotically negligible change for most circular stationary Gaussian processes with mean zero.

For X_t defined in (5) we have

$$I_X(\omega_k) = \frac{(B_k^2 + C_k^2)}{4}, \quad k = 1,\ldots,m \quad \text{and} \quad I_X(\omega_{N-k}) = I_X(\omega_k),$$

$$E(X_t) = 0 \quad \text{and} \quad \mathrm{Cov}(X_p, X_q) = r_{p-q} + r_{N-(p-q)},$$

$$1 \leqslant q \leqslant p \leqslant N.$$

Here

$$r_k = \frac{1}{N}\sum_{t=1}^{N-k}(X_t - \overline{X})(X_{t+k} - \overline{X})$$

is the k-lag sample autocovariance.

For the process defined in (5) we consider different test statistics, namely first-lag sample autocorrelation, measures of asymmetry, higher-order central moments, higher-order joint central moments, and higher-order cumulants. Other measures have been proposed for checking the nonlinear chaotic behavior of the generator of a time series. In particular, correlation dimension and maximum Lyapunov exponent are widely used. But these statistics cannot be calculated by an automatic procedure, and for this reason, it is difficult to implement them in a simulation study. As an alternative we have considered correlation sums. They are defined as samples analogous of correlation integrals and can be computed by an automatic scheme. The following test statistics have been used in our simulation study:

$$T_1 = \frac{1}{N}\sum_{t=1}^{N-1}(X_t - \overline{X})(X_{t+1} - \overline{X})/\hat{\sigma}^2, \quad T_2 = \sum_{t=1}^{N}(X_t - \overline{X})^3/\hat{\sigma}^3,$$

$$T_3 = \frac{\mathcal{N}\{X_t > X_{t+1}\}}{N}, \quad T_4 = \frac{1}{N}\sum_{t=1}^{N}(X_t - \overline{X})^4,$$

$$T_5 = \frac{1}{N}\sum_{t=1}^{N}(X_t - \overline{X})^5, \quad T_6 = \frac{1}{N}\sum_{t=1}^{N}(X_t - \overline{X})^6,$$

$$T_7 = \frac{1}{N}\sum_{t=1}^{N}(X_t - \overline{X})^7, \quad T_8 = \max_\tau Q(\tau),$$

$$Q(\tau) = \frac{\sum_{t=\tau+1}^{N}(X_{t-\tau} - X_t)^3}{\left[\sum_{t=\tau+1}^{N}(X_{t-\tau} - X_t)^2\right]^{3/2}},$$

$$T_9 = \frac{1}{N}\sum_{t=1}^{N-2}\prod_{k=0}^{2}(X_{t+k} - \overline{X}), \quad T_{10} = \frac{1}{N}\sum_{t=1}^{N-4}\prod_{k=0}^{4}(X_{t+k} - \overline{X}),$$

$$T_{11} = C_N(r), \quad T_{12} = \log[C_N(r)]/\log(r),$$

where $\hat{\sigma}^2 = N^{-1}\Sigma_{t=1}^{N}(X_t - \overline{X})^2$ and

$$C_N(r) = \frac{\sum_{i=2}^{N}\sum_{j=1}^{i} I(\|\mathbf{X}_i^\nu - \mathbf{X}_j^\nu\| < r)}{N(N-1)/2}$$

denotes the correlation sum. Here I is the indicator function and $\|\mathbf{X}\| = \max_k |X_k|$. The vector $\mathbf{X}_i^\nu = (X_{i-(\nu-1)d}, X_{i-(\nu-2)d}, \ldots, X_i)^T$ belongs to the phase space with embedding dimension ν and the delay time d. We use delay time $d=2$. Simulations were done for different embedding dimensions. The results turned out to be similar and we report the results only for embedding dimension $\nu=4$.

The test statistic T_1 has been added for theoretical reasons. Its use would only make sense for testing if the one-lag autocorrelation exceeds a certain level. For our hypothesis that contains processes with autocorrelations of all values between −1 and 1, this test statistic makes no sense. But we will see how the method of surrogate data transforms this test into a meaningful test. The test statistics T_3 and T_8 have been proposed as measures of time asymmetry. It has been argued that time asymmetry gives a strong indication for nonlinearity. The statistics T_2 and $T_4, \ldots, T_7$ have been proposed as test statistics for normality. $T_4, \ldots, T_7$ could be replaced by studentized versions, e.g., $T_4/\hat{\sigma}^4$. This would not change the surrogate test because $\hat{\sigma}^2$ has an identical value for the original data and for the surrogate data, i.e, $N^{-1}\Sigma_{t=1}^{N}(X_t - \overline{X})^2 = N^{-1}\Sigma_{t=1}^{N}(X_t^* - \overline{X^*})^2$. The test statistics T_9 and T_{10} are joint higher-order central moments and they are proposed to test the nonlinearity of the dynamics. We have also considered other higher-order cumulants as in [8], but they are not reported here.

In our simulation study we generated data from model (5) for different choices of σ_j^2: $\sigma_{j(1)}^2 = \exp(-j/m)$, $\sigma_{j(2)}^2 = \exp(-j)$, $\sigma_{j(3)}^2 = I_Y(\omega_j)$, where $I_Y(\omega_j)$ is the periodogram function at ω_j

of one realization of an autoregressive (AR) process Y_t of order 2. Here $j=1,\ldots,m$ and $m=(N-1)/2$ if N is odd and $m=(N-2)/2$, if N is even. We always choose $\sigma^2=1.0$. The Gaussian random variables and uniformly distributed random variables are generated by using the routines given in [13]. Note that for the processes generated by using $\sigma^2_{j(2)}=\exp(-j)$, the autocorrelation very slowly decreases, e.g., the autocorrelation between X_1 and X_m is approximately equal to -0.4. The time series generated by $\sigma^2_{j(3)}=I_Y(\omega_j)$ is circular, but at the same time, the autocorrelation function matches that of the underlying AR(2) process. We have considered several stationary Gaussian AR(2) processes, but we will report here only for $Y_t=1.5Y_{t-1}-.55Y_{t-2}+e_t$. For generating the surrogates, we computed the periodogram values (or the DFT), at $\omega_j=2\pi j/N$, $j=1,\ldots,m$ with the fast Fourier transform algorithm. We used sample size $N=256$. For each simulated $\mathbf{X_N}$, 1000 surrogate data vectors $\mathbf{X_N^*}$ were generated and for each of these 1000 surrogate data vectors we calculate the test statistics $T_j(\mathbf{X_N^*})$ for $j=1,\ldots,12$. The $(1-\alpha)$-th quantile of $T_j(\mathbf{X_N^*})$ is denoted by $\hat{k}_{j\alpha}$. This whole procedure is replicated 1000 times. This gives 1000 values of the critical values $\hat{k}_{j\alpha}$ and 1000 values of the test statistic $T_j(\mathbf{X_N})$. The $(1-\alpha)$-th quantile of the 1000 values of the test statistic is denoted by $k_{j\alpha}$. This is an approximation for the $(1-\alpha)$-th quantile of the distribution of $T_j(\mathbf{X_N})$ and is calculated by Monte Carlo calculations. In the following discussions, we will neglect the (random) inaccuracy in the Monte Carlo calculation of $k_{j\alpha}$. We compare two tests: the surrogate data test that rejects if $T_j(\mathbf{X_N})>\hat{k}_{j\alpha}$ and the test that rejects if $T_j(\mathbf{X_N})>k_{j\alpha}$. Clearly, for real data the second test is not available because it requires knowledge of the distribution of the test statistic. We have included this test for theoretical reasons. One may conjecture that these two tests are asymptotically equivalent: the probability that one test rejects and the other one accepts converges to zero. If this would be true (for the hypothesis and for the alternative) it would allow for a very simple understanding of the surrogate data test. In particular, it would give simple asymptotic formulas for the power function of the surrogate data test. We will see that in general the asymptotic equivalence of the two tests does not hold. For many cases the surrogate data estimate $\hat{k}_{j\alpha}$ of the critical value significantly differs from the true critical value $k_{j\alpha}$. In this respect surrogate data tests show a distinct behavior as compared to other resampling tests. For a large class of models, and for other resampling methods, one has that $\hat{k}_{j\alpha}$ and $k_{j\alpha}$ are asymptotically equivalent. Histograms of the test statistics $T_j(\mathbf{X_N})$ and the corresponding surrogate critical values $\hat{k}_{j\alpha}$ for $\alpha=.05$ are plotted in Figs. 1–7 for test statistics T_3, T_4, T_8, T_9, T_{10}, T_{11}, and T_{12}. This is done for $\sigma^2_{j(1)}=\exp(-j)$. Similar plots with other test statistics are not included to avoid repetitions. For some test statistics, the critical values concentrate around a fixed value c, for example. Thus in this case, the surrogate test will show a similar behavior as the test that rejects if the test statistic exceeds the value c. On the other hand, for T_1, T_4, T_6, T_{11}, and T_{12} the range of the values of the test statistic is of the same order as the range of the surrogate data estimate of their critical values. Thus $\hat{k}_{j\alpha}$ is not asymptotically equivalent to the true nonrandom quantile $k_{j\alpha}$ and the surrogate data test $T_j(\mathbf{X_N})>\hat{k}_{j\alpha}$ may be quite different from the theoretical test that rejects if $T_j(\mathbf{X_N})>k_{j\alpha}$. So in this case, for an understanding of the surrogate test it does not suffice to study the distribution of the test statistic alone. Here we have to look at the joint distribution of the test statistic and of the surrogate critical values. Only an understanding of the joint distribution enables us to study when the test statistic is larger than the surrogate critical value. In particular, it is clear that in this case the surrogate test strongly differs from the test that rejects if the test statistic exceeds its critical value.

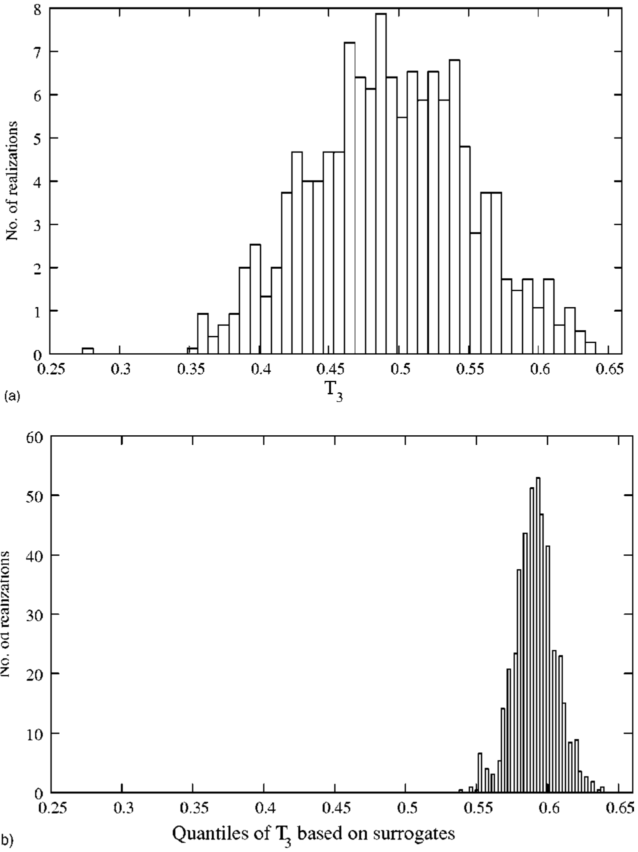

FIG. 1. Plot of the histogram of the test statistic T_3 (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

Table I gives a detailed overview on the quantitative differences between the two tests for $\alpha=0.05$ and for test statistics T_1–T_{12}. The differences are measured by p, p_1, and p_2, which are defined as follows:

$$p=\frac{\mathcal{N}\{T_j(\mathbf{X_N})>\hat{k}_{j\alpha}\}}{1000},$$

$$p_1=\frac{\mathcal{N}\{T_j(\mathbf{X_N})>\hat{k}_{j\alpha},T_j(\mathbf{X_N})<k_{j\alpha}\}}{1000},$$

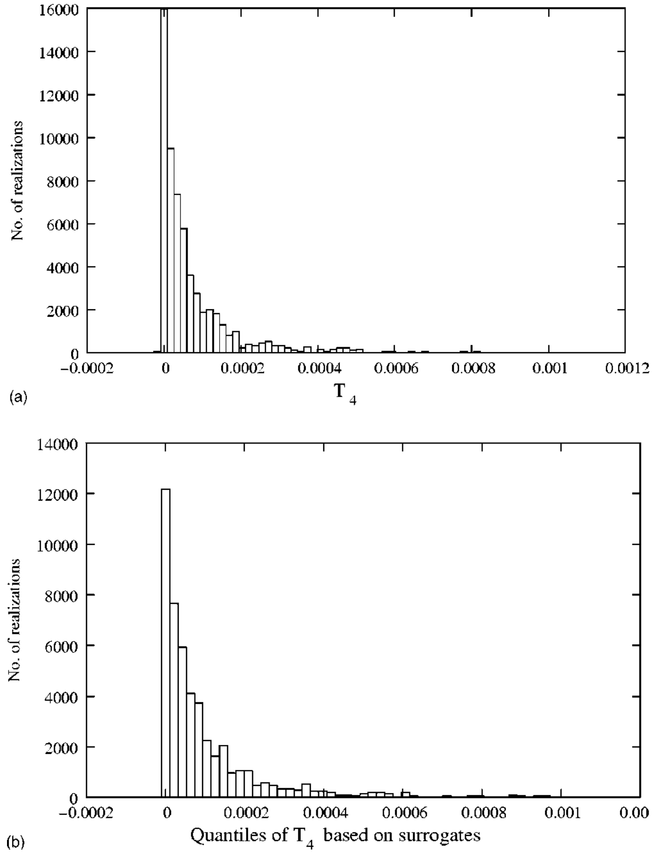

FIG. 2. Plot of the histogram of the test statistic T_4 (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

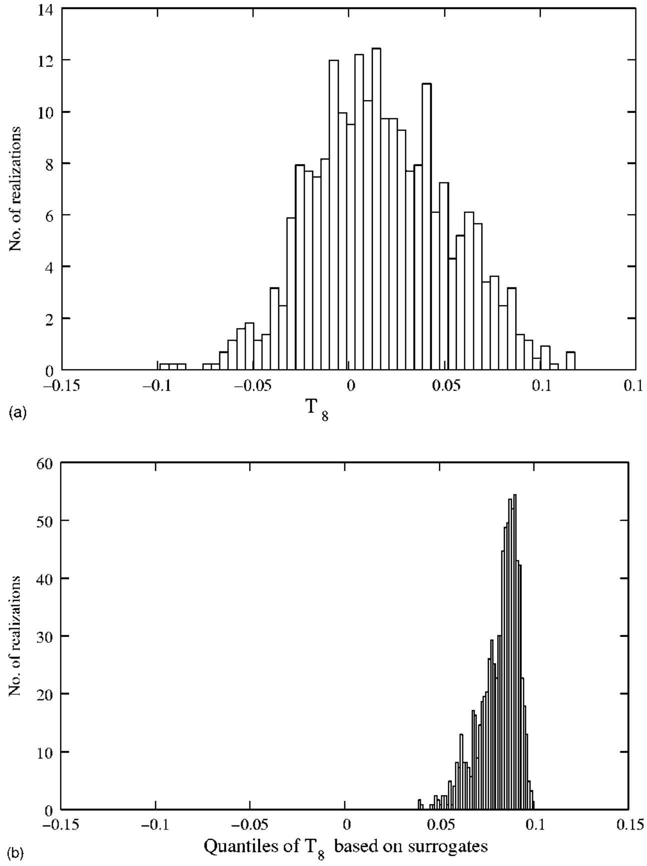

FIG. 3. Plot of the histogram of the test statistic T_8 (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

$$p_2 = \frac{\mathcal{N}\{T_j(\mathbf{X_N}) < \hat{k}_{j\alpha}, T_j(\mathbf{X_N}) > k_{j\alpha}\}}{1000}. \qquad (6)$$

The fraction p gives an estimate of the level of the surrogate data test. Because the test achieves the correct level for all test statistics, p should be equal to α. So the different values of p are only caused by simulation errors. They are reported here for better interpretation of p_1 and p_2. The fractions p_1 and p_2 count the relative number of cases where one test rejects and the other one accepts. When p_1 and p_2 are large and almost equal to the size α of the test, then this implies that the sets $\{T_j(\mathbf{X_N}) > \hat{k}_{j\alpha}\}$ and $\{T_j(\mathbf{X_N}) > k_{j\alpha}\}$ are almost nonoverlapping. When the two probabilities are small the above two sets overlap in a large area. Figures 1–7 and Table I show that the considered tests behave quite differently. For T_1, T_4, T_6, T_{11}, and T_{12} the variance of the surrogate quantiles is of the same order as the variance of the test statistic itself. This explains the large values of p_1 and p_2 in Table I. For the other test statistics, p_1 and p_2 are smaller, but not negligible. For the test statistic T_3 with $\sigma^2_{j(1)}$, the quantiles based on surrogates concentrate at one value (plot is not provided here). Then $\hat{k}_{j\alpha} = k_{j\alpha}$ and it holds that $p_1 = p_2 = 0$. But for T_3, in the case of long-range-dependent processes the values of p_1 and p_2 are nonnegligible. We conclude that in many cases the tests $T_j(\mathbf{X_N}) > \hat{k}_{j\alpha}$ and $T_j(\mathbf{X_N}) > k_{j\alpha}$ behave quite differently. Our simulations only show this for the hypothesis. By a standard asymptotic argument this can be extended to points of the alternative. Neighbored points of the alternative, i.e,. points for which the Neyman-Pearson test has nontrivial power, are also called contiguous. If a statistic S_N converges to zero under a contiguous point of the alternative, then it must also converge to zero on the hypothesis. This is a central argument often used in asymptotic test theory. Application with $S_N = (\hat{k}_{j\alpha} - k_{j\alpha}) / [\mathrm{var}\{T_j(\mathbf{X_N})\}]^{1/2}$ shows that this quantity cannot converge to zero on contiguous alternative points (because otherwise it must also converge to zero for points of the hypothesis). This implies that the two tests behave differently also on contiguous points of the alternative.

All realizations of the processes, so far discussed, are of length $N = 256$. For such a small data size one may expect large variations of the test statistics. It may be argued that the described phenomenon is an effect due to small sample size. In many applications, one comes across very large data sets. Due to this reason, we have repeated simulations for longer time series. We consider circular process (5) with $\sigma_j^2 = \exp(-j)$ and $N = 2048$. The results are reported in Table III. The values of p_1 and p_2 are not close to zero. This implies that surrogate data tests do not consistently estimate their critical values. We conclude that the same findings, as above, also apply for large data sets. The tests based on surrogate resamples may behave quite differently from the theoretical

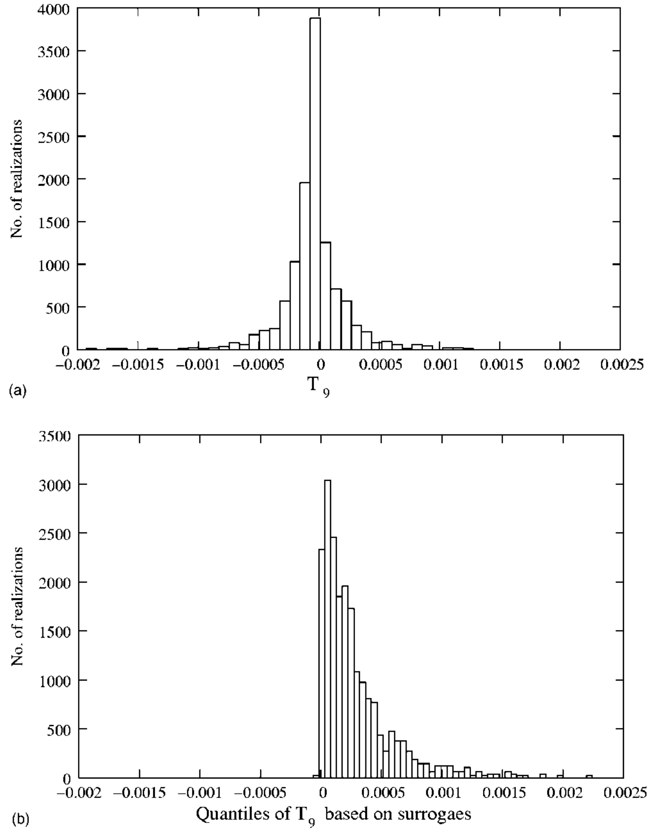

FIG. 4. Plot of the histogram of the test statistic T_9 (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

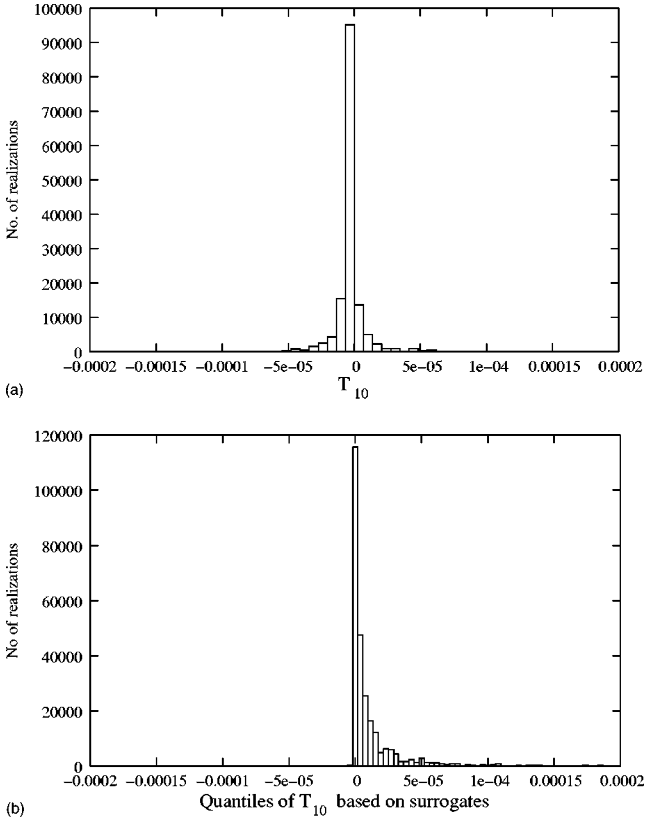

FIG. 5. Plot of the histogram of the test statistic T_{10} (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

tests that reject if test statistics exceed their critical values.

Up to now we only considered circular processes. This was done because for these processes surrogate data tests achieve exact levels. We now present simulations for an (noncircular) AR(1) process:

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon \sim N(0,1), \tag{7}$$

with $\phi=0.5$ and 0.99. We have made simulations for the same test statistics as for the circular processes. The results for different AR(1) processes with surrogate resamples are given in Table II. We observe that also for such processes, p_1 and p_2 are not negligible. For noncircular processes, no theoretical result about asymptotic and finite-sample level accuracy (like circular processes) exists. In the simulations, we observe that levels are approximately correct for the process (7) with $\phi=0.5$, but level inaccuracies have been observed for most of the test statistics when $\phi=0.99$. This implies that some restrictions on the test statistics and/or models are required for noncircular processes. Also for $\phi=0.99$, we get that p_1 and p_2 are quite large. For longer time series with $N=2048$ (see Table III), this effect does not disappear. Thus, again, we conclude that surrogate data tests are quite different from the theoretical test that rejects if the test statistic exceeds its critical value.

Surrogate data tests achieve finite-sample level accuracies for all circular stationary Gaussian processes. As mentioned above, for other resampling methods a quite different behavior was claimed in theoretical studies; e.g., in a lot of papers on bootstrap for a wide range of applications it was shown that the bootstrap test is asymptotically equivalent to the theoretical test that rejects if the test statistic exceeds its (unknown) $(1-\alpha)$-th quantile. This would imply a performance of bootstrap tests that is qualitatively different from that observed here for surrogate data tests. We checked this by a small simulation. For autoregressive processes we implemented a parametric bootstrap method. We have excluded T_1, T_{11}, and T_{12}, as no proper bootstrap procedure was found for testing using these test statistics. Resamples are generated from the fitted autoregressive process and they are used to calculate test statistics denoted by T_j^*. Bootstrap tests work in a slightly different way than surrogate data tests. The bootstrap resamples $(X_1^*, \ldots, X_n^*)$ cannot directly be used for checking the significance of T_j because the bootstrap resamples do not fulfill the hypothesis of the test statistic; e.g., for $j=3$, we do not have that the conditional mean of $\{X_i^* > X_{i+1}^*\}$ is equal to 0.5. For this reason we follow the usual bootstrap approach based on prepivoting. The bootstrap test rejects if $[T_j - \mu_j]/\hat{\sigma}^l \geqslant \hat{k}_{j\alpha}^b$. Here $\mu_j = E[T_j]$ in case $E[T_j]$ is known, i.e., for $j=2,3,5,7,9$, and 10. For $j=4$, it is estimated as $\mu_j = 3\hat{\sigma}^4$ and for $j=6$, the term μ_j is put equal to $15\hat{\sigma}^6$. The norming $\hat{\sigma}^l$ substitute the usual studentization that is known to measure a higher-order accuracy of bootstrap. Naturally we put $l=0$ for $j=3$; $l=3$ for $j=2,9$; $l=j$ for $j=4,\ldots,7$; and $l=5$ for $j=10$. The quantile $\hat{k}_{j\alpha}^b$ is calculated

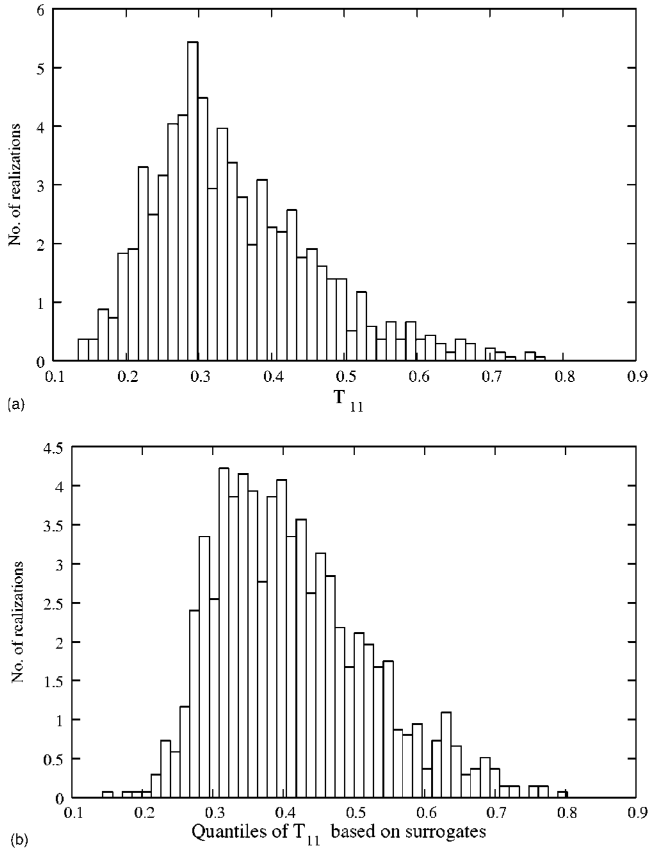

FIG. 6. Plot of the histogram of the test statistic T_{11} (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

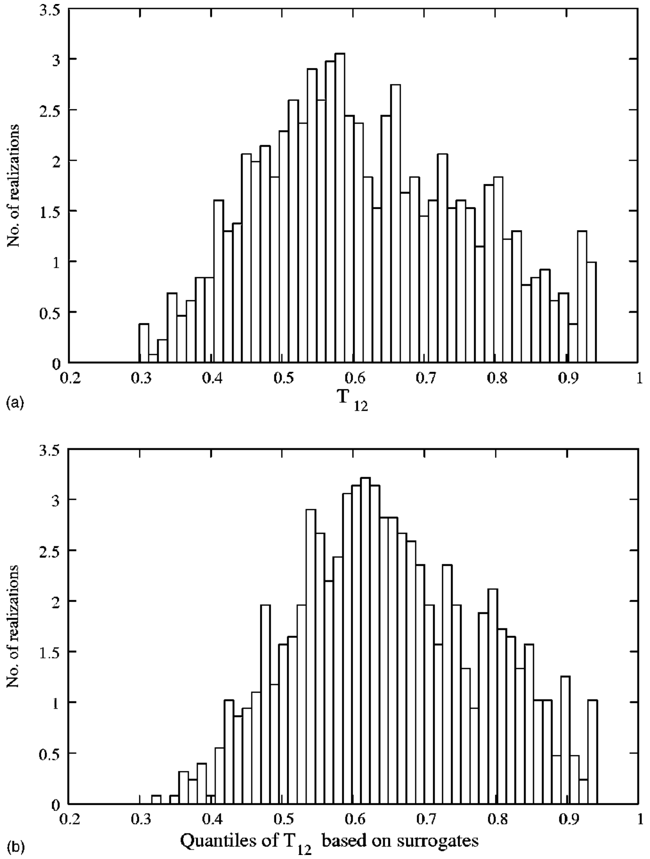

FIG. 7. Plot of the histogram of the test statistic T_{12} (upper plot) and plot of the histogram of 95% quantiles based on surrogate data (lower plot).

by bootstrap resampling. It is the conditional quantile of $[T_j^* - E^*(T_j^*)]/\hat{\sigma}^{*l}$, where E^* denotes the conditional expectation given the original sample and $\hat{\sigma}^{*l}$ is the empirical variance of a bootstrap sample. For $j=8$ we used the same test statistic as above. In the resampling we calculated $T_8^* = \max_\tau Q^*(\tau)$, where $Q^*(\tau) = Q'(\tau) - E^* Q'(\tau)$ and $Q'(\tau)$ is defined as $Q(\tau)$, but with the original sample replaced by a bootstrap sample. The results of the simulations with $N = 256$ and 2048 are summarized in Table IV. The level accuracy is slightly worse compared to the performance of surrogate data. The relative values of p_1 and p_2 are slightly smaller than for surrogate data if one corrects for level inaccuracies. (Note that $p + p_2 - p_1 = 0.05$ and for this reason p_1 and p_2 cannot vanish if $p - 0.05$ is large in absolute value.) Thus the simulations support the conjecture that surrogate data testing has more accurate level accuracies but may change the nature of the test. But the differences seem not so large as may be expected from the bootstrap literature.

V. DETAILED DISCUSSION OF TEST STATISTICS T_1 AND T_4

In this section, we discuss why surrogate data tests based on test statistics T_1 and T_4 transform to tests for circular stationarity. T_1 was proposed to test whether first-lag autocorrelation exceeds a certain value, whereas T_4 was proposed to measure deviations from normality. But after application of surrogate data, tests based on T_1 and T_4 look for quite different types of alternatives. We have discussed for T_1 and T_4, but similar arguments also apply for other test statistics. Because the circular sample autocovariance is preserved for surrogate data we have that

$$T_1(\mathbf{X_N}) + \frac{1}{N}(X_N - \overline{X})(X_1 - \overline{X})\hat{\sigma}^{-2}$$

$$= T_1(\mathbf{X_N^*}) + \frac{1}{N}(X_N^* - \overline{X})(X_1^* - \overline{X})\hat{\sigma}^{-2}.$$

This gives

$$[T_1(\mathbf{X_N}) - T_1(\mathbf{X_N^*})]\hat{\sigma}^2 = \frac{c^2}{N}\sum_{j,k=1}^{m}(B_j^2 + C_j^2)^{1/2}(B_k^2 + C_k^2)^{1/2}[\cos(\theta_j^*)\cos(\omega_k + \theta_k^*) - \cos(\theta_j)\cos(\omega_k + \theta_k)].$$

The surrogate data test rejects if the $(1-\alpha)$-th quantile of the conditional distribution of this difference (given B_j, C_j, and θ_j for $j=1,\dots,m$) exceeds 0. This test has exact level α for the hypothesis that (conditionally given B_j and C_j for $j = 1,\dots,m$) the variables $\theta_1,\dots,\theta_m$ are conditionally independent with uniform distribution on $[0,2\pi]$. It could be argued that this is a test that measures deviations from the hypothesis of circular stationarity.

TABLE I. p, p_1, and p_2 for different test statistics and for different circular processes with $N=256$.

σ_j^2		T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
$e^{-j/m}$	p	0.048	0.044	0.037	0.043	0.044	0.047	0.047	0.062	0.057	0.042	0.044	0.045
	p_1	0.044	0.001	0.000	0.031	0.008	0.026	0.011	0.025	0.014	0.010	0.028	0.028
	p_2	0.047	0.008	0.000	0.039	0.015	0.030	0.015	0.014	0.009	0.019	0.035	0.022
e^{-j}	p	0.048	0.049	0.055	0.046	0.050	0.049	0.047	0.037	0.049	0.052	0.048	0.066
	p_1	0.032	0.018	0.009	0.043	0.040	0.044	0.038	0.004	0.036	0.040	0.044	0.063
	p_2	0.036	0.020	0.004	0.048	0.042	0.047	0.042	0.018	0.038	0.040	0.047	0.048
$I_Y(\omega_j)$	p	0.049	0.053	0.043	0.040	0.054	0.039	0.054	0.039	0.054	0.051	0.046	0.046
	p_1	0.044	0.014	0.000	0.036	0.029	0.035	0.032	0.004	0.027	0.027	0.027	0.026
	p_2	0.046	0.012	0.000	0.047	0.026	0.047	0.029	0.017	0.024	0.028	0.030	0.015

Similar arguments apply for the test statistic T_4. Again, by construction, the surrogate data test rejects with probability α if (conditionally given B_j and C_j for $j=1,\ldots,m$) the variables $\theta_1,\ldots,\theta_m$ are conditionally independent with uniform distribution on $[0,2\pi]$. No further restriction on the distributions of B_j and C_j for $j=1,\ldots,m$ is needed. In particular, they may have very heavy tailed distributions. That means that this surrogate data test does not look for deviations from normality. Again, it is a test for circular stationarity. This may become also clear by the following representation:

$$\begin{aligned}T_4(\mathbf{X_N}) - T_4(\mathbf{X_N^*}) = {} & \frac{c^4}{8}\sum_{j_1+j_2+j_3+j_4=N} d(j_1,j_2,j_3,j_4)[\cos(\theta_{j_1}+\theta_{j_2} \\ & + \theta_{j_3}+\theta_{j_4}) - \cos(\theta^*_{j_1}+\theta^*_{j_2}+\theta^*_{j_3}+\theta^*_{j_4})] \\ & + \frac{c^4}{2}\sum_{j_1+j_2+j_3-j_4\in\{0,N\}} d(j_1,j_2,j_3,j_4) \\ & \times[\cos(\theta_{j_1}+\theta_{j_2}+\theta_{j_3}-\theta_{j_4}) - \cos(\theta^*_{j_1} \\ & + \theta^*_{j_2}+\theta^*_{j_3}-\theta^*_{j_4})] \\ & + \frac{3c^4}{8}\sum_{j_1+j_2-j_3-j_4=0} d(j_1,j_2,j_3,j_4)[\cos(\theta_{j_1} \\ & + \theta_{j_2}-\theta_{j_3}-\theta_{j_4}) \\ & - \cos(\theta^*_{j_1}+\theta^*_{j_2}-\theta^*_{j_3}-\theta^*_{j_4})],\end{aligned}$$

where

$$\begin{aligned}d(j_1,j_2,j_3,j_4) = {} & [(B_{j_1}^2+C_{j_1}^2)(B_{j_2}^2+C_{j_2}^2)(B_{j_3}^2+C_{j_3}^2)(B_{j_4}^2 \\ & + C_{j_4}^2)]^{1/2}.\end{aligned}$$

So in particular this surrogate test checks if the distribution of $\theta_{j_1}\pm\theta_{j_2}\pm\theta_{j_3}\pm\theta_{j_4}$ is a fourfold convolution of uniform distributions, whereas the test based on T_1 checks the distribution of pairwise sums $\theta_j\pm\theta_k$. The test may not reject in the case of heavy tailed distributions of amplitudes; the same large values of B_j and C_j are used for the original sample and for the surrogates. We again would like to emphasize the point that for both test statistics the nature of the test drastically changes by use of surrogate data critical values. Tests for one-lag autocorrelation or deviations from normality are transformed to tests on circular stationarity. We now briefly want to explain why this change is more evident for tests using even moments than for tests using odd moments. It can be shown that

$$\begin{aligned}E^*[T_4(\mathbf{X_N^*})] = {} & \frac{3c^4}{8}\left[\sum_{j=1}^{m}(B_j^2+C_j^2)\right]^2 \\ & + \frac{3c^4}{4}\sum_{j\neq k}^{m}(B_j^2+C_j^2)(B_k^2+C_k^2).\end{aligned}$$

Thus the test statistic T_4 is corrected by a term that heavily depends on the values of B_j and C_j. On the other hand, odd empirical moments have conditional mean zero. This follows by symmetry of the distribution of surrogate data. Therefore, here randomness of surrogate critical values are only caused

TABLE II. p, p_1, and p_2 for different test statistics and for different AR(1) processes with $N=256$ and using surrogate resamples. $X_t=\phi X_{t-1}+e(t)$, $e(t)\sim N(0,1)$.

ϕ		T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
0.05	p	0.049	0.054	0.043	0.051	0.045	0.051	0.051	0.053	0.055	0.048	0.038	0.057
	p_1	0.046	0.006	0.000	0.042	0.009	0.030	0.013	0.012	0.015	0.020	0.020	0.036
	p_2	0.048	0.004	0.000	0.042	0.016	0.030	0.013	0.011	0.012	0.024	0.034	0.030
0.99	p	0.421	0.061	0.071	0.070	0.059	0.065	0.060	0.121	0.058	0.057	0.208	0.013
	p_1	0.392	0.015	0.019	0.068	0.044	0.061	0.048	0.069	0.040	0.042	0.174	0.006
	p_2	0.023	0.005	0.000	0.049	0.036	0.047	0.039	0.000	0.033	0.036	0.013	0.028

TABLE III. p, p_1, and p_2 for different test statistics for circular process with $\sigma_j^2=\exp(-j)$ and AR(1) process $X_t=.05X_{t-1}+e(t)$ with $N=2048$ using surrogate resamples.

					$\sigma_j^2=\exp(-j), \quad j=1,\ldots,m$					
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
p	0.054	0.054	0.053	0.049	0.058	0.050	0.061	0.063	0.055	0.057
p_1	0.018	0.022	0.015	0.045	0.044	0.044	0.046	0.014	0.043	0.042
p_2	0.013	0.020	0.013	0.047	0.037	0.045	0.037	0.002	0.039	0.037
					$X_t=0.5X_{t-1}+e(t), \quad e(t)\sim N(0,1)$					
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
p	0.077	0.050	0.052	0.061	0.062	0.060	0.070	0.061	0.054	0.059
p_1	0.074	0.002	0.002	0.046	0.014	0.028	0.018	0.010	0.009	0.014
p_2	0.048	0.004	0.000	0.036	0.004	0.019	0.000	0.001	0.007	0.007

by the random nature of the conditional variance and conditional higher-order moments of the surrogate data. This explains that these critical values are more stabilized. In general, we do not have a clear intuition for which types of tests the randomness of surrogate critical values are small and for which are not. This is an important problem and requires further research.

VI. CONCLUSIONS

In this paper we mainly concentrate on circular processes. This has been done because for this class of models surrogate data tests achieve exact levels. In simulations we have also included noncircular AR(1) processes. We have shown that surrogate data tests do not always consistently estimate their critical values. Thus surrogate data tests may differ essentially from the theoretical tests that reject if the test statistics exceed their critical values. This means that the nature of the test may change drastically by the use of surrogate data. After the application of surrogate data, a test that measures for a certain type of deviation from the null hypothesis may look for quite different types of alternatives. An example is the test statistic T_4. This test statistic measures for heavy tails of the amplitudes of the process. However, after using surrogate data, the test measures for deviations from stationarity. In this respect, surrogate data tests differ from bootstrap methods. For almost all bootstrap tests, under certain conditions, it has been shown that they consistently estimate their critical values. This implies that bootstrap tests achieve asymptotically exact levels. They are asymptotically equivalent to the tests that reject if the test statistics exceed their critical values. These findings are supported by simulations presented in this paper. But the difference between bootstrap and surrogate is not as drastic as expected.

Surrogate data tests achieve exact levels for all circular stationary Gaussian processes. We conjecture that consistent estimation of critical values may not be possible for all such processes. Note that this class also contains all types of long-range-dependent processes. In particular, we think that bootstrap methods will work only for more restrictive classes. This would imply that surrogate data can be applied for a

TABLE IV. p, p_1, and p_2 for different test statistics and for different AR(1) processes with $N=256$ and 2048 using bootstrap resamples. $X_t=\phi X_{t-1}+e(t), e(t)\sim N(0,1)$.

N	ϕ		T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
256	0.5	p	0.050	0.046	0.023	0.048	0.026	0.040	0.059	0.053	0.051
		p_1	0.007	0.001	0.001	0.013	0.004	0.013	0.014	0.006	0.014
		p_2	0.008	0.000	0.029	0.016	0.029	0.024	0.006	0.005	0.014
256	0.99	p	0.055	0.053	0.014	0.044	0.009	0.026	0.105	0.056	0.047
		p_1	0.009	0.003	0.000	0.005	0.000	0.002	0.069	0.009	0.003
		p_2	0.006	0.006	0.037	0.012	0.042	0.027	0.016	0.005	0.008
2048	0.5	p	0.046	0.059	0.053	0.051	0.052	0.050	0.060	0.055	0.055
		p_1	0.006	0.004	0.005	0.006	0.007	0.012	0.009	0.005	0.008
		p_2	0.002	0.000	0.003	0.007	0.007	0.013	0.001	0.001	0.005
2048	0.99	p	0.046	0.038	0.026	0.043	0.027	0.048	0.055	0.048	0.043
		p_1	0.005	0.000	0.000	0.002	0.000	0.002	0.013	0.007	0.002
		p_2	0.010	0.004	0.025	0.010	0.024	0.006	0.010	0.010	0.011

much richer class of models. On the other hand, if more restrictive models are appropriate it may be appropriate to switch from surrogate data tests to bootstrap methods. Further research is needed to specify when bootstrap is preferable and which bootstrap method should be chosen.

In our study, we observe that the difference between surrogate data tests and theoretical tests is present for some test statistics and does not appear for some others. It would be interesting to understand for which type of test statistics this different behavior appears and under which conditions this phenomenon disappears. This point requires further study.

Surrogate data can be used to obtain estimates of the mean and variance of a test statistic. It is a common approach to use these estimates to get a studentized version of the test statistic and to use quantiles of the standard normal or t distribution as critical values. Clearly this test does not have an exact level, but for large enough sample sizes under regularity conditions on the test statistic, one may expect approximately accurate levels. This may be shown by standard applications of the central limit theorem for the conditional distribution of the test statistic for surrogate data. These arguments imply that direct calculation of critical values by surrogate data or indirect calculation by use of surrogate data mean and variance will lead to asymptotic equivalent tests. The arguments could be extended to studies of the behavior of both tests on the alternative. By use of the same arguments we expect that both tests are asymptotically equivalent. A more detailed study of this point is deferred to another paper.

In this paper, we mainly discuss stationary circular Gaussian. This has been done because for this class of processes, exact finite sample accuracy of levels holds for surrogate data tests. In simulations we observe that for more general classes of stationary Gaussian processes that are noncircular, levels are only asymptotically correct. The classical method of phase randomization to generate surrogate data is discussed in this paper. There are several other methods proposed in the literature that are more appropriate for noncircular processes. It would be interesting to extend our discussion to these methods. In particular, it should be checked if the described phenomena remain present for these procedures.

ACKNOWLEDGMENTS

This study is supported by DFG project MA 1026/7-1 in the framework of priority program SPP-1114. We would also like to thank Jens Timmer for constructive comments on the manuscript.

[1] J. Theiler *et al.*, Physica D **58**, 77 (1992).

[2] S. Elgar, R. T. Guza, and R. J. Seymour, J. Geophys. Res. B **89**, 3623 (1984).

[3] A. R. Osborne, A. D. Kirwan, A. Provenzale, and L. Bergamasco, Physica D **23**, 75 (1986).

[4] P. Grassberger, Nature (London) **323**, 609 (1986).

[5] K. S. Chan, Fields Inst. Commun. **11**, 77 (1997).

[6] J. Theiler and D. Prichard, Physica D **110**, 221 (1996).

[7] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods* (Springer, New York, 1987).

[8] D. Nur, R. C. Wolff, and K. L. Mengersen, Comput. Stat. Data Anal. **37**, 487 (2001).

[9] W. J. Braun and R. J. Kulpurger, Commun. Stat: Theory Meth. **26(6)**, 1329 (1997).

[10] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications* (Cambridge University Press, Cambridge, 1997).

[11] J. Timmer, Phys. Rev. E **58**, 5153 (1998).

[12] J. Timmer, Phys. Rev. Lett. **85**, 2647 (2000).

[13] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran, The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1992).