

# *Studies in Nonlinear Dynamics & Econometrics*

---

*Volume 12, Issue 1*

2008

*Article 4*

NONLINEAR DYNAMICAL METHODS AND TIME SERIES  
ANALYSIS

---

## Evaluation of Surrogate and Bootstrap Tests for Nonlinearity in Time Series

Dimitris Kugiumtzis\*

\* Aristotle University of Thessaloniki, dkugiu@gen.auth.gr

# Evaluation of Surrogate and Bootstrap Tests for Nonlinearity in Time Series\*

Dimitris Kugiumtzis

## Abstract

The validity of any test for nonlinearity based on resampling techniques depends heavily on the consistency of the generated resampled data to the null hypothesis of linear stochastic process. The surrogate data generating algorithms AAFT, IAAFT and STAP, as well as a residual-based bootstrap algorithm, all used for the randomization or bootstrap test for nonlinearity, are reviewed and their performance is compared using different nonlinear statistics for the test. The simulations on linear and nonlinear stochastic systems, as well as chaotic systems, reveals a variation in the test outcome with the algorithm and statistic. Overall, the bootstrap algorithm led to smallest test power whereas the STAP algorithm gave consistently good results in terms of size and power of the test. The performance of the nonlinearity test with the resampling techniques is evaluated on volume and return time series of international stock exchange indices.

---

\*I would like to thank the reviewers for their valuable comments.

# 1 Introduction

An important question when analyzing time series is about the presence of nonlinearity. In the setting of time series analysis based on dynamical system theory, also referred to as nonlinear or dynamic analysis of time series, the term nonlinearity refers to the nonlinear dynamics of the underlying to the time series system. The dynamics is assumed to be mainly deterministic and possibly corrupted by dynamic or observational noise (Diks, 2000, Kantz and Schreiber, 1997). Nonlinearity is established by ruling out that the generating mechanism of a random-like stationary time series is a linear stochastic process. In this setting, the dynamics of the stochastic component is downweighted and there is no distinguishing of nonlinearity in the mean or in the variance. On the other hand, in the setting of stochastic time series analysis, there are other alternatives to a linear stochastic process, involving the stochastic component, its moments and interaction with the deterministic component. Thus the alternative hypothesis may regard nonlinearity in the mean, e.g. a nonlinear autoregressive process (NAR), or in the variance, e.g. an autoregressive process with conditional heteroscedasticity (ARCH) (Fan and Yao, 2003, Granger and Teräsvirta, 1993, Tong, 1990). Such tests are referred to as linearity tests, though they are essentially the same as the test for nonlinearity used in the dynamical analysis setting (Cromwell, Labys, and Terazza, 1994, Patterson and Ashley, 2000).

Linearity tests abound in time series analysis. Most common are the tests for the null hypothesis of linearity in the conditional mean,  $H_0: P[E(x_t|A_t)=A_t'\theta]=1$ , or simply

$$H_0 : x_t = A_t'\theta + \epsilon_t, \quad (1)$$

where  $\{x_t\}_{t=1}^n$  is the time series,  $A_t$  is the information on  $x_t$  at time  $t - 1$  (in autoregression of order  $p$ , this is  $A_t = [x_{t-1}, \dots, x_{t-p}]$ ) and  $\{\epsilon_t\}_{t=1}^n$  denotes a series of independent and identically distributed (iid) variables. For a specific alternative hypothesis of the form  $H_1: x_t = A_t'\theta + f(A_t) + \epsilon_t$ , where  $f$  is a nonlinear function, a number of test statistics have been proposed, such as the RESET test that assumes a polynomial form of  $A_t'\theta$  (Ramsey, 1969), the neural network tests of White (1989) and Teräsvirta, Lin, and Granger (1993), other specific forms for  $f$ , such as the smooth transition autoregressive model (STAR) (Luukkonen, Saikkonen, and Teräsvirta, 1988a), the logistic STAR and the exponential STAR (Teräsvirta, 1994), the exponential autoregressive model (EXPAR) and self excited threshold autoregressive model (SETAR) (Tong, 1990), and the Hamilton test that uses a non-specific nonlinear form for  $f$  (Hamilton, 2001). Tests for a non-specific  $H_1$  include the bicorrelation and bispectrum tests (Hinich, 1982, 1996), as well as the generalized spectrum test (Hong and Lee, 2005), the McLeod test using squared residuals

from a linear fit (McLeod and Li, 1983), the Keenan and Tsay tests using residuals of a linear fit of  $x_t$  and  $x_t^2$  (Keenan, 1985, Tsay, 1986), and the BDS test that measures the density structure of residuals from a linear fit in an embedded space (Brock, Dechert, and Scheinkman, 1996). All these tests are implemented using the asymptotic distribution of the test statistic under  $H_0$ , called also null distribution, and they are referred to as *asymptotic tests*. Though asymptotic tests have been used up to-date (e.g. see for BDS in (Jašić and Wood, 2006), for Teräsvirta's neural network test in (Dagum and Giannerini, 2006) and for White's neural network test in (Kyrtsou, 2005)) the analytic null distribution may not always be accurate, altering the size and power of the tests (Brooks and Henry, 2000, Chan and Ng, 2004, Davies and Petrucci, 1986, Hjellvik and Tjøstheim, 1996, Lee, Kim, and Newbold, 2005, Yuan, 2000).

In dynamic analysis of time series a number of nonlinear measures have been developed that have been used or can potentially be used as test statistics for the linearity test, or as we call it here test for nonlinearity. The BDS statistic, which has been widely used in econometrics, is actually based on a measure of the density of points within a given distance (a basic measure used also to derive estimates of fractal dimension and entropy). Other nonlinear measures that have been used in the test for nonlinearity include entropy measures, such as the mutual information and the conditional mutual information (Diks and Manzan, 2002, Kugiumtzis, 2001), the largest Lyapunov exponent (e.g. used in Brzozowska-Rup and Orłowski (2004)) and the local linear fit, found to have the largest power among a number of nonlinear test statistics in Schreiber and Schmitz (1997). The null distribution of these statistics is unknown and randomization or bootstrap tests are called. Actually the focus has been on randomization tests making use of different schemes for the generation of the so-called surrogate data (Kugiumtzis, 2000, 2002a,b, Schreiber and Schmitz, 1996, 2000, Theiler, Eubank, Longtin, and Galdrikian, 1992). There are plenty of bootstrap approaches for correlated time series (e.g. see Politis (2003), Wu (2006)), but it appears that bootstrap approaches have been less used in conjunction with statistics from dynamical system theory (Brzozowska-Rup and Orłowski, 2004, Fernández-Rodríguez, Sosvilla-Rivero, and Andrada-Félix, 2005, Wolff, Yao, and Tong, 2004, Yao and Tong, 1998, Ziehmann, Smith, and Kurths, 1999).

Little work has also been done on comparing bootstrap and surrogate data in the test for nonlinearity, and in a single work we are aware of in Hinich, Mendes, and Stone (2005) the comparison is limited to surrogate data for Gaussian time series, the so-called Fourier transform (FT) surrogates (Theiler et al., 1992). For the null hypothesis of underlying linear stochastic process the time series does not have to be Gaussian, and the surrogate data have to preserve both the linear structure and the marginal distribution of the time series.

In this work, we consider three prominent algorithms for the generation of such

surrogate data, namely the Amplitude Adjusted Fourier Transform (AAFT) (Theiler et al., 1992), the Iterative AAFT (IAAFT) (Schreiber and Schmitz, 1996), and the Statically Transformed Autoregressive Process (STAP) (Kugiumtzis, 2002a). Also, we include in the simulation study a residual-based bootstrap algorithm. Representative linear and nonlinear stochastic systems, including a chaotic system, are tested for nonlinearity employing these resampling techniques. For the tests, we use three nonlinear statistics for non-specific alternative hypothesis representing different aspects of the data, the bicorrelation as a measure of nonlinear autocorrelation, the mutual information as a measure of entropy and information, and the local average fit as a non-specific nonlinear model, which is however built mainly for modeling nonlinear dynamics.

The algorithms for the generation of surrogate and bootstrap data and the test statistics are briefly presented in Sec. 2. Then the simulation setup is given and the results are discussed in Sec. 3. The test is then applied to volume and return time series of international stock exchange indices in Sec. 4. Finally, the pros and cons of the different approaches in the test for nonlinearity are discussed in Sec. 5.

## 2 Randomization and Bootstrap Tests for Nonlinearity

Randomization and bootstrap are both resampling approaches that generate random samples from the original data under given conditions. For the test for nonlinearity, the conditions for the resampled time series  $\{z_t\}_{t=1}^n$  are that it preserves the linear correlation and marginal distribution of the original time series  $\{x_t\}_{t=1}^n$ .

### 2.1 Surrogate data for the test for nonlinearity

In the surrogate data test, the null hypothesis of stochastic linear process is postulated in terms of a Gaussian process and reads that the time series  $\{x_t\}_{t=1}^n$  is generated by a standard Gaussian process  $\{s_t\}$  under a static (instantaneous) transform  $h$ ,

$$\mathbf{H}_0 : x_t = h(s_t), \quad \{s_t\} \sim \mathbf{N}(0, 1, \rho_s), \quad (2)$$

where  $\rho_s$  is the autocorrelation of  $\{s_t\}$ . The transform  $h$  may be linear or nonlinear, and monotonic or non-monotonic. The underlying Gaussian process accounts for the presence of only linear dynamics in the observed time series and the transform  $h$  allows for deviations from the Gaussian marginal distribution.

The algorithms AAFT, IAAFT and STAP that generate surrogate data for the randomization test preserve exactly the condition for the marginal distribution,

$f_x(x_t) = f_z(z_t)$ , where  $f_x(x_t)$  is the marginal probability density function (pdf) of  $\{x_t\}_{t=1}^n$  (we refer to the pdf rather than the cumulative density function (cdf)  $F_x(x_t)$ , as done elsewhere, in order to facilitate presentation of histogram estimates later on). The three algorithms approximate the condition for the linear correlation in different ways: AAFT and IAAFT approximate the sample power spectrum,  $S_x(f) \simeq S_z(f)$ , where  $S_x(f)$  is the periodogram of  $\{x_t\}_{t=1}^n$ , whereas STAP approximates the sample autocorrelation,  $r_x(\tau) \simeq r_z(\tau)$ , for a sufficient range of lags  $\tau$ . The AAFT and IAAFT algorithms follow the constrained realization approach directly attempting to generate data that fulfill the two conditions, whereas the STAP algorithm uses a typical realization approach and attempts to build a proper autoregressive model in order to generate data that match the two conditions (see Theiler and Prichard (1996) for comparisons of the two types of approaches but for the hypothesis of Gaussian time series).

The AAFT algorithm was built under the assumption of monotonic  $h$  (Theiler et al., 1992). In Kugiumtzis (1999) it was shown that when  $\{s_t\}$  is Gaussian and  $h$  is non-monotonic, AAFT cannot match the linear structure of  $\{x_t\}_{t=1}^n$ . Discrepancies in the linear structure may also occur with IAAFT because it approximates  $S_x(f)$  iteratively starting from a flat spectrum and the algorithm terminates at about the same accuracy of approximation for each surrogate data generation. In some cases, the discrepancy in approximation in conjunction with the small variance of  $S_z(f)$  causes a bias in the linear structure approximation and thus favors rejection of  $H_0$ . Such problems in the application of AAFT and IAAFT on chaotic systems have been reported using Monte Carlo simulations with different test statistics in Kugiumtzis (2001). Problems with the two Fourier-based algorithms were reported also in Mammen and Nandi (2004).

The STAP algorithm uses a typical realization approach and estimates an AR model from the so-called Gaussian autocorrelation  $r_u(\tau)$ . To find  $r_u(\tau)$ , the autocorrelation transform  $\psi$  is estimated,  $r_x = \psi(r_u)$ , from the sample static transform  $g$  of a Gaussian time series  $\{u_t\}_{t=1}^n$  to  $\{x_t\}_{t=1}^n$  given as  $x_t = g(u_t) = F_x^{-1}(\Phi(u_t))$ , where  $\Phi(u)$  is the standard normal cdf. The use of  $\Phi$  here instead of the sample cdf of normal iid (used initially in Kugiumtzis (2002a)), gives unique solution for the AR model and better accuracy of the STAP algorithm. Then Gaussian time series  $\{u_t\}_{t=1}^n$  are generated by the AR model and are rescaled to the final surrogate time series  $\{z_t\}_{t=1}^n$  under  $g$  transform in order to attain  $f_x(x_t)$  (or  $F_x(x_t)$ ) (Kugiumtzis, 2002a). Thus the sample autocorrelation of  $\{z_t\}_{t=1}^n$  may vary a lot when  $n$  is small and this affects the variance of the nonlinear statistic as well. In this way, subtle nonlinearities in the original time series may be masked and the test may become conservative and have small power, depending always on the chosen test statistic.

One could argue that the test be applied to the back transformed data to Gaussian cdf, denoted  $\{w_t\}_{t=1}^n$ , by applying the inverse  $g$  transform. This would work

in the case of a monotonic  $h$  in  $H_0$  and in fact it would render the statistics pivotal using FT surrogates, as pointed in Theiler and Prichard (1996). However, for a non-monotonic  $h$ ,  $\{w_t\}_{t=1}^n$  may not be a Gaussian time series though it has always marginal Gaussian cdf. Thus a non-Gaussian  $\{w_t\}_{t=1}^n$  may be derived either from a time series  $\{x_t\}_{t=1}^n$  consistent to  $H_0$  (linear stochastic but non-monotonic  $h$ ) or a time series  $\{x_t\}_{t=1}^n$  inconsistent to  $H_0$  (not linear stochastic), as pointed in Kugiumtzis (2000). Therefore the analysis cannot be done on  $\{w_t\}_{t=1}^n$  as it can lead to false rejection (for the first case) at the cost of having to deal with non-pivotal (or less pivotal) statistics and thus reaching lower power levels for the test.

## 2.2 Bootstrap data for the test for nonlinearity

The standard approach in nonparametric testing is rather bootstrap than randomization. When the original sample contains correlations, as for time series, people use either block bootstrap (joining together randomly chosen blocks) or residual-based (called also model-based) bootstrap approaches (resampling the residuals of a model and feeding them into the model to generate bootstrapped time series) (Politis, 2003). For the latter, the simplest is the naive approach, drawing the bootstrap residuals  $e_t^*$  randomly by replacement from the model residuals  $e_t$ . There are also variations of this approach, such as the wild bootstrap (Liu, 1998). In the time series literature most common is the naive approach that we also use here (e.g. see Chen and Liu (2001), Hjellvik and Tjøstheim (1995)).

In residual-bootstrap, an AR model is typically fitted to  $\{x_t\}_{t=1}^n$  where its order is estimated by an order selection criterion, such as the Akaike information criterion (AIC). We denote this residual-based bootstrap algorithm as ARboot. The use of AR model in ARboot aims at preserving the original linear structure. We observed that for some systems a better match of autocorrelation could be attained with a higher order of AR than the one estimated by AIC, and we followed this whenever it was necessary. Redrawing from the residuals does not assure the preservation of the marginal cdf, which is preserved exactly by all surrogate algorithms. Large discrepancies in the marginal cdf may cause false rejections for the test as we show in the next Section. STAP and ARboot are both typical realization approaches in that they both use an AR model to preserve the linear structure of the original time series, but STAP preserves exactly the original marginal cdf, sharing partly an attribute of the constrained realization approaches.

## 2.3 Test statistics

We use test statistics for non-specific alternatives as the focus is on the quality of the resampled data rather than the power of the test for a specific alternative hypothesis.

We consider measures applied directly on the time series without removing first the linear correlations, as is done in BDS. We use three statistics regarding different aspects of the data: the bicorrelation, extending the standard autocorrelation to the third order autocorrelation, the mutual information, constructed under information and entropy ideas, and the fit with a local average mapping, referring to a simple nonparametric and nonlinear model.

**Bicorrelation (BIC)** The bicorrelation test is the equivalent of the bispectrum test in time domain (Brooks and Hinich, 2001, Hinich, 1982, 1996). The bicorrelation function (called also third-order moment or three-point autocorrelation) at two positive lags  $\tau$  and  $s$  ( $\tau < s$ ) is defined as

$$G(\tau, s) = (n - s)^{-1} \sum_{t=1}^{n-s} x(t)x(t + \tau)x(t + s). \quad (3)$$

The portmanteau statistic of bicorrelation is

$$H = \sum_{s=2}^L \sum_{\tau=1}^{s-1} (n - s)G^2(\tau, s), \quad (4)$$

where the number of lags  $L$  is a free-parameter set by the user. For the null hypothesis of independence it was shown that  $H \sim \chi_{(L-1)L/2}^2$  (Hinich, 1996). However, this asymptotic result is not useful for the test for nonlinearity (unless the linear correlation is first removed).

In Barnett and Wolff (2005), Schreiber and Schmitz (1997), the bicorrelation in (3) was used to detect nonlinear correlation at specific lags. In the same way, we use the bicorrelation at specific small  $\tau$  as test statistics, denoted BIC, and we always set  $s=2\tau$ .

**Mutual Information (MUT)** The mutual information is an entropy-based measure that estimates the general correlation (linear and nonlinear) between  $x_t$  and  $x_{t-\tau}$  for different lags  $\tau$ . It is defined as (e.g. see Kantz and Schreiber (1997))

$$I(\tau) = \sum_{i,j} p_i \log \frac{p_{i,j}}{p_i p_j}. \quad (5)$$

Here the histogram-based estimate is used and in the above expression the summation is over the bins of the partition of the data,  $p_i$  is the estimated probability that a data point  $x_t$  is in bin  $i$ ,  $p_j$  is the estimated probability that a data point  $x_{t-\tau}$  is in bin  $j$ , and  $p_{i,j}$  is the estimated joint probability that  $x_t$  is in bin  $i$  and  $x_{t-\tau}$  is in bin  $j$ . The bins are equidistant and the number of bins is set to  $\sqrt{n/5}$ . The test statistic is  $I(\tau)$  at specific small lags  $\tau$ , denoted MUT.

**Local Average Mapping (LAM)** The local average mapping defines the one-step ahead prediction  $\hat{x}_{t+1}$  of a target point  $\mathbf{x}_t$ , where  $\mathbf{x}_t = [x_t, x_{t-1}, \dots, x_{t-m+1}]'$ , as the average of the one step ahead mappings of the  $k$  nearest points to  $\mathbf{x}_t$  (Kantz and Schreiber, 1997). The goodness of fit is measured with the normalized root mean square error (NRMSE). The test statistic is the NRMSE at different embedding dimensions  $m$ , denoted LAM. Among a number of statistics including BIC and MUT, LAM was found to give larger power to the surrogate data test for nonlinearity when applied to chaotic time series (Kugiumtzis, 2001, Schreiber and Schmitz, 1997).

## 2.4 Implementation of the test with resampled data

The null distribution for a statistic  $q$  is formed by the values of  $q$  computed on an ensemble of  $M$  resampled time series,  $q_1, q_2, \dots, q_M$ . Then if the statistic computed on the original time series, denoted  $q_0$ , is in the tails of the empirical null distribution,  $H_0$  is rejected.

Often in surrogate data testing a parametric approach is followed for the test decision. Here, we use a distribution-free approach and reject  $H_0$  if  $q_0$  is smaller than the  $\alpha/2$  quantile or larger than the  $1-\alpha/2$  quantile of the set  $\{q_0, q_1, q_2, \dots, q_M\}$  (assuming a two-sided test). In all simulations we use  $M=1000$  and for  $\alpha=0.05$  and a two-sided test,  $H_0$  is rejected if  $q_0$  is in the first or last 25 positions of the ordered sequence of  $q_0, q_1, q_2, \dots, q_{1000}$ .

## 3 Monte Carlo Simulations

The simulation study is focused mainly on the appropriateness of the resampled approaches to form the null distribution of the test statistics. Moreover, the size and power of the test with the three statistics is assessed for each resampling technique.

### 3.1 The systems

To assess the resampled techniques and test statistics we applied the test with each one of them to time series from different systems listed below.

1. A power transform of an AR(1) process with normal input noise, consistent to  $H_0$ ,

$$x_t = s_t^a, \quad s_t = 0.3 + 0.8s_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathbf{N}(0, 1), \quad (6)$$

and  $\{\epsilon_t\}$  is an iid process. We consider a monotonic transform for  $a=3$  and a nonmonotonic transform for  $a=2$ .

2. An AR(1) process with conditional heteroscedasticity, ARCH(1,1), used in Luukkonen, Saikkonen, and Teräsvirta (1988b) and Hjellvik and Tjøstheim (1995)

$$x_t = 0.6x_{t-1} + \epsilon_t, \quad \epsilon_t = \eta_t \sqrt{0.2 + 0.8\epsilon_{t-1}^2}, \quad \eta_t \sim \text{N}(0, 1), \quad (7)$$

and  $\{\eta_t\}$  is an iid process. This process is nonlinear in the conditional variance.

3. The bilinear autoregressive process, BL(1,1), used also in Luukkonen et al. (1988b) and Hjellvik and Tjøstheim (1995)

$$x_t = \phi_0 + \phi_1 x_{t-1} + \psi_1 x_{t-1} \epsilon_{t-1} + \epsilon_t, \quad \epsilon_t \sim \text{N}(0, 1). \quad (8)$$

Two parameter settings are used. The first, for  $\phi_0=2.0$ ,  $\phi_1=-0.9$ ,  $\psi_1=-0.1$ , gives strong alternating autocorrelation, and the second, for  $\phi_0=1.0$ ,  $\phi_1=0.3$ ,  $\psi_1=-0.2$ , gives weak autocorrelation. The two BL systems are denoted BL1 and BL2, respectively. The nonlinearity source for the BL systems is the interaction of the stochastic component with the state variable.

4. The chaotic Henon map corrupted by observational additive iid noise

$$x_t = s_t + \epsilon_t, \quad s_t = 1 - 1.4s_{t-1}^2 + 0.3s_{t-2}, \quad \epsilon_t \sim \text{N}(0, \sigma_\epsilon^2). \quad (9)$$

Three noise levels are used, given as  $\sigma_\epsilon = b\sigma_s$ , for  $b=0.05, 0.4, 0.8$  and  $\sigma_s$  the standard deviation of the noise-free data. The nonlinearity here is in the deterministic dynamics, which is masked by the added noise at a degree depending on  $b$ .

### 3.2 The simulation setup

For each system, 1000 Monte Carlo realizations of size  $n=128$  and  $n=512$  were generated and 400 realizations of size  $n=1024$  (for the large data size a smaller number of realizations was found to give stable results at a manageable computational load). The three statistics BIC, MUT and LAM, were computed on each time series and on an ensemble of  $M=1000$  surrogate time series generated by AAFT, IAAFT and STAP, and bootstrap time series generated by ARboot. Thus three randomization tests and one bootstrap test were carried out for each test statistic. Actually, the test statistics were 9 as each one was computed for a varying parameter:  $\tau=1, 2, 3$  for BIC and MUT, and  $m=1, 2, 3$  for LAM. Thus for each Monte Carlo realization, 36 tests were performed (for 9 test statistics and 4 resampling approaches).

### 3.3 Consistency of resampled time series

The consistency of the surrogate and bootstrap time series  $\{z_t\}_{t=1}^n$  to  $H_0$  is determined by the preservation of  $f_x(x_t)$  and  $r_x(\tau)$  of  $\{x_t\}_{t=1}^n$ . The time series generated by ARboot do not match  $f_x(x_t)$  and actually the shape of  $f_z(z_t)$  can be very different. This is shown in Fig. 1a for a time series from the cubic power of the AR(1) process. Note that for all surrogate time series it is  $f_z(z_t) = f_x(x_t)$  by construction. On the other hand, the AAFT and IAAFT surrogate time series may have bias in the

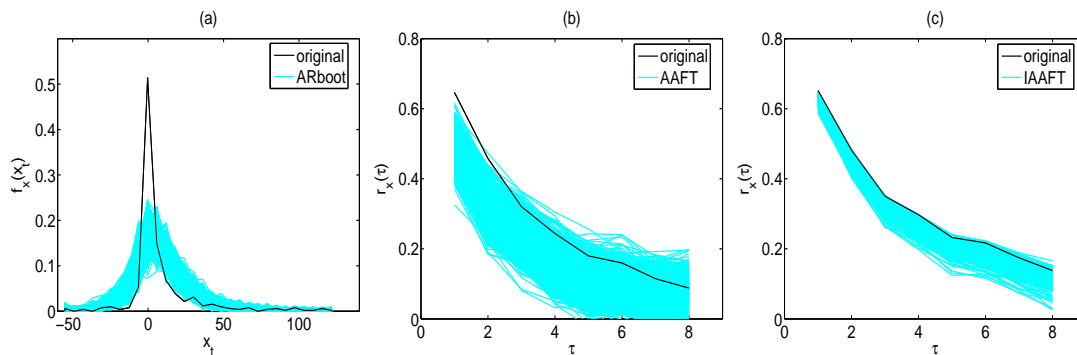


Figure 1: (a) Histogram-based estimate of the distribution of  $\{x_t\}_{t=1}^{512}$  from the cubic power of the AR(1) process and the same for 1000 ARboot time series. (b) Autocorrelation for  $\{x_t\}_{t=1}^{512}$  from the square power of the AR(1) process and for 1000 AAFT time series. (c) Autocorrelation for  $\{x_t\}_{t=1}^{512}$  from the cubic power of the AR(1) process and for 1000 IAAFT time series.

estimation of  $r_x(\tau)$ , as shown for AAFT and the square power of the AR(1) process in Fig. 1b and for IAAFT and the cubic power of the AR(1) process in Fig. 1c. Note that for IAAFT the variance of  $r_z(\tau)$  is much smaller and makes the small bias to be significant. The STAP time series and the ARboot time series estimate  $r_x(\tau)$  without bias, provided that the order of the AR generating process is appropriately set. The discrepancies in  $f_x(x_t)$  for ARboot time series and in  $r_x(\tau)$  for the AAFT and IAAFT time series may favor rejection of  $H_0$  when the test statistic is sensitive to these features. In the following results on size and power of the test we include  $r_x(\tau)$  for  $\tau=1, 2, 3$  as test statistics, denoted AUT.

### 3.4 Size and power of the test

Even for simple systems consistent to  $H_0$  (such as the square and cubic power of an AR(1) process), AAFT, IAAFT and ARboot generate time series that are not consistent to  $H_0$  (see Fig. 1). This results in false rejections and large test size,

as shown in Fig. 2 for the cubic power of an AR(1) process and  $n=128$ , and for the three nonlinear statistics and AUT ( $\tau=1$  for AUT, MUT and BIC, and  $m=3$  for LAM). The small mismatch in autocorrelation of IAAFT (see also Fig. 1c) is

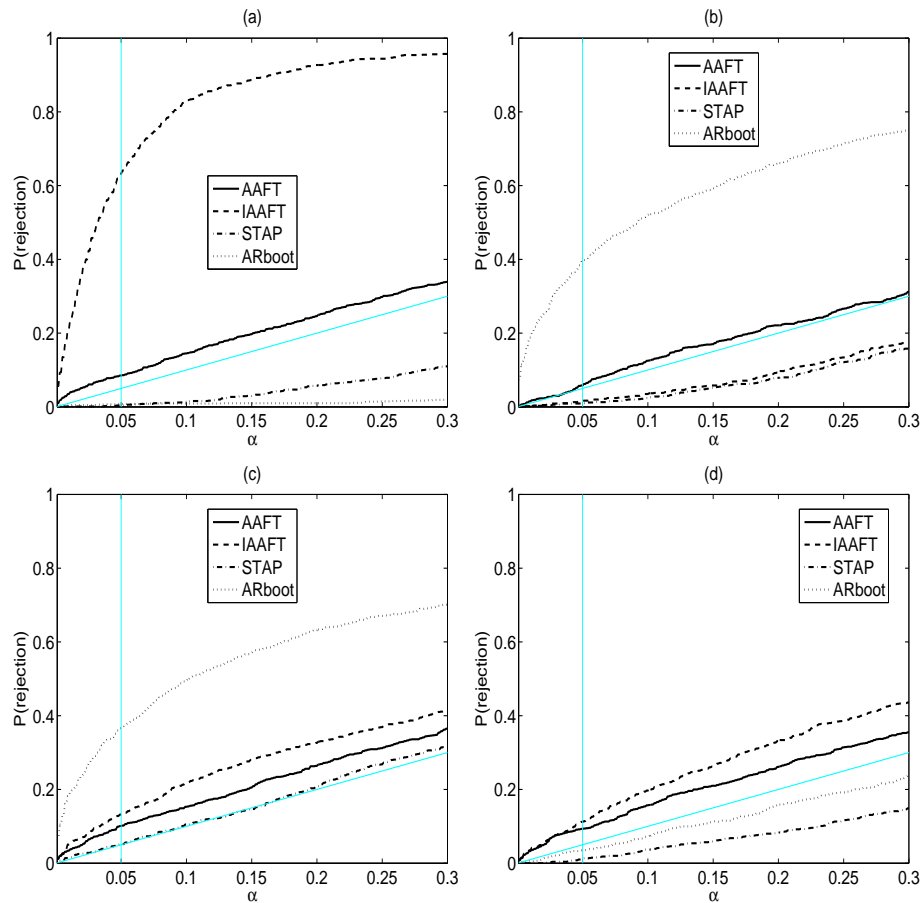


Figure 2: Probability of rejection of  $H_0$  as a function of significance level  $\alpha$  estimated by the relative frequency of rejection over 1000 realizations of size  $n=128$  from the cubic power of the AR(1) process. The line types for the randomization and bootstrap tests are denoted in the legend. The statistics are AUT ( $\tau=1$ ) in (a), BIC ( $\tau=1$ ) in (b), MUT ( $\tau=1$ ) in (c) and LAM ( $m=3$ ) in (d). The diagonal line and the  $\alpha=0.05$  line are shown in grey color to facilitate comparisons.

actually very significant as it gives large size of the test with AUT (see Fig. 2a). This results in large actual test size also with MUT and LAM, but not BIC. The mismatch of  $f_x(x_t)$  with the ARboot time series gives rise to very large test size with BIC and MUT, but not LAM. The AAFt and STAP perform properly here with AAFt showing somehow larger actual size than the nominal size and the opposite

is observed for STAP. These features of the resampled techniques and statistics hold also for larger data sizes and for other parameter values of the test statistics.

Summary results are shown in Table 1 for all systems, statistics, resampled techniques, and data sizes. The columns 3 to 9 display the estimated probability of rejection of  $H_0$  at  $\alpha=0.05$  for each system, where in each cell the values from top to bottom are for  $n=128, 512, 1024$ . The estimated probabilities of rejection are calculated from 1000 Monte Carlo simulations for  $n=128, 512$ , and 400 simulations for  $n=1024$ . Note that the probabilities of rejection in the third column of Table 1 for the monotonic transform of the AR(1) process correspond to the intersection points of the curves with the vertical line at  $\alpha=0.05$  in Fig. 1.

For the non-monotonic transform of the AR(1) process (see column 4 in Table 1), both AAFT and IAAFT show a mismatch in AUT, but only AAFT carries this mismatch to the nonlinear statistics giving large actual size of the test, which actually increases with  $n$ . ARboot gives also large actual size (with BIC and MUT) due to the mismatch in  $f_x(x_t)$  as for the monotonic transform. Note that ARboot along with STAP match perfectly the autocorrelation and as a result there are no rejections with AUT ( $\tau=1$ ) for all systems (the same holds for  $\tau=2$  and  $\tau=3$ ). So, only the test with STAP does not involve type I error, as shown for the two linear stochastic processes, and the actual test size tends to be somehow smaller than the nominal size, suggesting that the test using STAP may be conservative.

For the ARCH process, AAFT and IAAFT again cannot match the autocorrelation and the rejection rate for AUT increases with  $n$  (see column 5 in Table 1). Therefore the high rejection rates obtained with AAFT and IAAFT and the three nonlinear statistics cannot be assigned to high power of these tests. Note that similarly high rejection rates were obtained for linear stochastic systems solely due to the mismatch in autocorrelation. The ARboot time series preserve the original  $f_x(x_t)$  of the ARCH process (the simulations showed only small discrepancies at the peak) and are appropriate for the test. Indeed ARboot and STAP tend to give about the same power for the test, which is about the same for BIC and MUT and very low for LAM, even for  $n=1024$ . The latter shows the inappropriateness of LAM to detect nonlinearities in the variance.

The mismatch of  $r_x$  with AAFT and IAAFT occurs when the bilinear (BL) process has strong  $r_x$  but not when it has weak  $r_x$  (BL1 and BL2 in columns 5 and 6 of Table 1). This explains that the probability of rejection with AAFT and IAAFT is higher than for STAP in the first case but only slightly higher than for STAP in the second case. ARboot preserves well the original  $f_x(x_t)$  for both BL processes and the power of the test with ARboot is at the same level as or smaller than for STAP. For BL1, the test has no power with MUT and LAM statistics and increases the power with  $n$  only with the BIC statistic and at a larger rate with STAP. To the contrary, for BL2, MUT and LAM give larger power than BIC when  $n$  increases.

statistic	algorithm	AR(1) $x = s^3$	AR(1) $x = s^2$	ARCH	BL1	BL2	Henon 40%	Henon 80%	
AUT ( $\tau=1$ )	AAFT	0.09	0.64	0.48	0.89	0.04	0.06	0.01	
		0.07	0.98	0.66	0.95	0.10	0.62	0.05	
		0.06	1.00	0.71	0.97	0.14	0.92	0.21	
	IAAFT	0.63	0.64	0.49	0.67	0.05	0.07	0.05	
		0.99	0.95	0.89	0.72	0.04	0.10	0.06	
		0.99	0.97	0.95	0.71	0.05	0.15	0.07	
	STAP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ARboot	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BIC ( $\tau=1$ )	AAFT	0.06	0.23	0.30	0.16	0.15	0.78	0.29	
		0.05	0.43	0.40	0.82	0.21	1.00	0.78	
		0.05	0.69	0.42	0.98	0.28	1.00	0.96	
	IAAFT	0.02	0.04	0.27	0.23	0.14	0.82	0.30	
		0.08	0.07	0.45	0.83	0.21	1.00	0.79	
		0.14	0.09	0.51	0.98	0.27	1.00	0.96	
	STAP	0.01	0.01	0.22	0.10	0.13	0.81	0.29	
		0.00	0.01	0.32	0.79	0.21	1.00	0.79	
		0.01	0.01	0.34	0.98	0.27	1.00	0.96	
	ARboot	0.39	0.46	0.32	0.07	0.15	0.71	0.18	
		0.76	0.92	0.48	0.14	0.21	1.00	0.77	
		0.94	1.00	0.51	0.25	0.28	1.00	0.96	
MUT ( $\tau=1$ )	AAFT	0.10	0.43	0.41	0.37	0.18	1.00	0.34	
		0.08	0.97	0.89	0.43	0.72	1.00	0.91	
		0.06	1.00	0.99	0.34	0.96	1.00	1.00	
	IAAFT	0.13	0.06	0.18	0.14	0.17	1.00	0.35	
		0.37	0.09	0.41	0.08	0.70	1.00	0.93	
		0.56	0.10	0.49	0.09	0.96	1.00	1.00	
	STAP	0.05	0.01	0.10	0.01	0.13	0.96	0.18	
		0.10	0.00	0.29	0.00	0.67	1.00	0.83	
		0.11	0.01	0.39	0.00	0.94	1.00	0.99	
	ARboot	0.37	0.20	0.08	0.00	0.11	0.96	0.18	
		0.86	0.87	0.32	0.01	0.52	1.00	0.79	
		0.99	1.00	0.37	0.01	0.82	1.00	0.99	
LAM ( $m=3$ )	AAFT	0.09	0.48	0.24	0.69	0.16	1.00	0.60	
		0.07	0.92	0.48	0.84	0.46	1.00	0.98	
		0.06	0.98	0.64	0.92	0.80	1.00	1.00	
	IAAFT	0.11	0.07	0.10	0.27	0.15	1.00	0.60	
		0.10	0.05	0.21	0.28	0.45	1.00	0.98	
		0.09	0.08	0.28	0.39	0.79	1.00	1.00	
	STAP	0.01	0.00	0.04	0.00	0.06	1.00	0.28	
		0.00	0.00	0.09	0.00	0.36	1.00	0.91	
		0.00	0.00	0.13	0.00	0.74	1.00	1.00	
	ARboot	0.04	0.00	0.02	0.00	0.08	1.00	0.46	
		0.04	0.00	0.07	0.00	0.39	1.00	0.92	
		0.06	0.00	0.10	0.00	0.76	1.00	1.00	

Table 1: Summary results for the nonlinearity test from Monte Carlo simulations.

For the chaotic Henon map, the noise amplitude affects the power of the test, as expected, but also the quality of the AAFT and IAAFT surrogate time series with AAFT giving significant discrepancies in  $r_x$  when the noise amplitude is 40% (see column 8 of Table 1). For this noise level, the power of all statistics is very

high, ranking LAM first (being 1 even for  $n=128$ ) and BIC third (0.7-0.8 for  $n=128$ ). All resampling techniques give about the same power with STAP and ARboot performing similarly. This is not true when the noise level increases to 80%, where for  $n=128$  AAFT and IAAFT give about double power than STAP and ARboot for MUT and LAM. This difference reduces for  $n=512$  and they all approach one at  $n=1204$ . For BIC, the power is the same with all resampling techniques and generally smaller than for MUT and LAM. The LAM statistic turns out to give the highest power of the test independent of the resampling technique, which agrees with the conclusion in Schreiber and Schmitz (1997).

Among the three statistics, LAM gives highest test power for the chaotic system, BIC for ARCH and BL with strong  $r_x$ , and MUT for BL with weak  $r_x$ . The estimation of mutual information is data demanding and this is reflected in the increasing power of MUT with  $n$ , which is faster compared to BIC and LAM in many cases. LAM turns out to have smaller power than MUT when the nonlinearity regards the stochastic component (mostly for ARCH but also for BL2). Overall, MUT seems to be the statistic of choice when the time series is large, otherwise one should use LAM if the alternative aims at nonlinearity in the system dynamics and BIC if it regards stochastic nonlinearity.

Similar results to these presented in Table 1 were obtained with other parameter values of the statistics, but for larger lags the autocorrelation levels off for all systems and the nonlinear effects vanish as well. For LAM, better discrimination is actually obtained for  $m=1$  in the case of BL2 and the probability of rejection for both STAP and ARboot increases to about 0.15, 0.56 and 0.88 for data sizes 128, 512 and 1024, respectively. The statistics MUT and LAM involve other method-specific parameters. For MUT, the choice of the number of bins is critical. Our simulations showed that adjusting the number of bins to the sample size as  $\sqrt{n/5}$  gives better results than using a fixed number of bins. For example, using 16 bins the probability of rejection reduces to half of this displayed in Table 1 for the systems of BL2 and Henon with 40% and 80% noise and STAP surrogates. For LAM, we used a fixed number of neighboring points  $k=5$  for the results in Table 1, but other simulations have showed that the performance of LAM gets better with an increasing  $k$  with  $n$ .

## 4 Application to financial time series

The performance of the resampling techniques and test statistics is evaluated on time series from 6 international stock exchange indices, namely the US stock exchange indices Dow Jones, NASDAQ and S&P500, the Athens Stock Exchange index (ASE), the index of the 100 largest companies listed at the London Stock

Exchange (FTSE100), and the Hang Seng index (HSI) of the Hong Kong Stock Exchange. For each index, the randomization and bootstrap tests for nonlinearity are applied using  $M=1000$  resampled time series and the statistics MUT, BIC and LAM, as well as AUT, for a range of the free parameter for each statistic ( $\tau$  and  $m$ ).

#### 4.1 Daily volumes of stock exchanges

First we consider the volume of the daily stock exchanges in the period 3/12/2002 – 28/9/2004. The selected time series contain slow drifts, but the augmented Dickey-Fuller test did not show evidence of unit-root type of non-stationarity for varying number of lags and in any of the six time series (Greene, 2007). However, one should be cautious about concluding for nonlinearity when rejecting  $H_0$  in the presence of drifts (though statistically insignificant).

In Fig. 3 the results of the distribution-free approach for the test are shown for ASE. The AAFT and IAAFT are inappropriate for the test as IAAFT underestimates and AAFT overestimate  $r_x(\tau)$  giving rejections with AUT for a long range of  $\tau$  values. ARboot also mismatches  $r_x(\tau)$  for  $\tau > 4$ . Rejection of  $H_0$  is obtained for a range of parameter values of MUT and LAM, also with STAP that is more conservative but more reliable for this case. BIC does not give significant rejection with STAP except when  $\tau=7$ . The well-established rejection of  $H_0$  here should not be directly interpreted as evidence of nonlinearity (and possibly nonlinear dynamics given the rejection with LAM) because it could be the effect of departures from stationarity (though not statistically significant). Indeed the ASE volume time series contains more prominent slow drifts than the other volume time series. Moreover, this time series is more spiky having skewed distribution and long right tail and this, in conjunction with significant autocorrelation across many lags, may cause the inability of AAFT and IAAFT to cope with the two conditions in  $H_0$ . The larger autocorrelation in AAFT surrogate time series explains their better local fit (smaller NRMSE) than for the ASE volume time series (see Fig. 3d).

The resampled time series generated by STAP and ARboot fulfill the conditions of  $H_0$  for the other volume time series as well but not the surrogate time series generated by AAFT and IAAFT. As shown in Table 2, the  $p$ -values regarding AUT for  $\tau=1$  are small for all but the S&P500 time series when using IAAFT and for ASE and FTSE100 when using AAFT. This significant difference in AUT may explain the discrimination with nonlinear statistics in the respective cases, particularly when it cannot be established also by STAP or ARboot. For example, the test on HSI using IAAFT and MUT ( $\tau=1$ ) gives  $p \simeq 0.01$  (as for AUT), but using STAP and ARboot the  $p$ -values are large.

The results with STAP and ARboot show some evidence for nonlinearity only in ASE when using MUT ( $p < 0.01$ ) and in FTSE100 when using BIC ( $p < 0.01$ ).

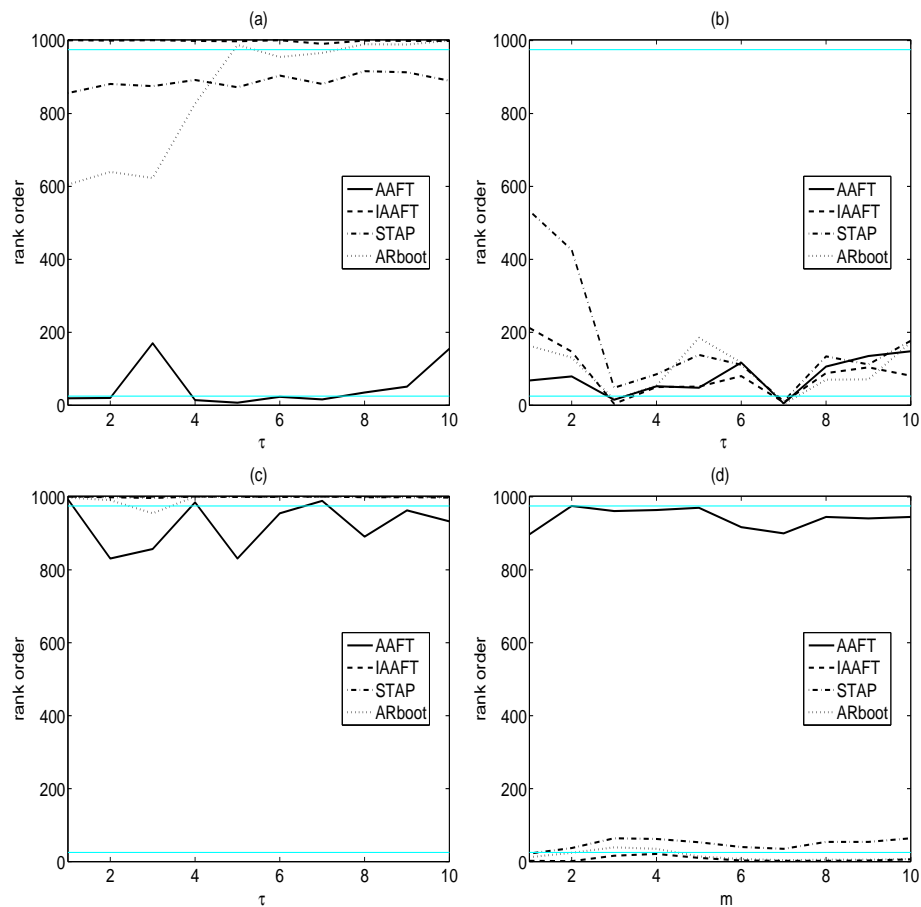


Figure 3: The rank of the statistic  $q_0$  on the volume time series of ASE in the ordered list of 1001 values, where  $M=1000$  resampled time series of different types are used, as given in the legend. In (a) the statistic is AUT for  $\tau=1, \dots, 10$ , in (b) BIC for  $\tau=1, \dots, 10$ , in (c) MUT for  $\tau=1, \dots, 10$ , and in (d) LAM for  $m=1, \dots, 10$ . The lower and upper thresholds for  $\alpha=0.05$  are shown with horizontal grey lines.

ARboot gives also marginal rejection for Dow Jones with BIC ( $p \simeq 0.06$ ) and for ASE with LAM ( $p \simeq 0.08$ ). For the latter, rejection is obtained for other values of  $m$  also with STAP (see Fig. 3d).

## 4.2 Monthly returns of stock exchanges

We investigate the existence of nonlinearity in monthly returns of the same 6 stock exchange indices in the period 1/1987 to 9/2004. All time series show no apparent drift but larger volatility from 1997 onwards that could be taken as evidence of non-

index	statistic	Resampling techniques			
		AAFT	IAAFT	STAP	ARboot
Dow Jones	AUT	0.148	0.026	0.703	0.855
	MUT	0.068	0.901	0.633	0.869
	BIC	0.158	0.076	0.160	0.058
	LAM	0.012	0.234	0.436	0.218
NASDAQ	AUT	0.098	0.030	0.458	0.701
	MUT	0.150	0.819	0.428	0.989
	BIC	0.498	0.513	0.525	0.334
	LAM	0.074	0.424	0.408	0.551
S&P500	AUT	0.917	0.114	0.372	0.795
	MUT	0.070	0.168	0.184	0.599
	BIC	0.861	0.793	0.811	0.971
	LAM	0.947	0.809	0.370	0.699
ASE	AUT	0.038	0.000	0.290	0.791
	MUT	0.016	0.000	0.000	0.006
	BIC	0.136	0.424	0.935	0.326
	LAM	0.080	0.032	0.128	0.078
FTSE100	AUT	0.000	0.028	0.781	0.837
	MUT	0.140	0.032	0.551	0.294
	BIC	0.002	0.002	0.004	0.002
	LAM	0.004	0.953	0.843	0.641
HSI	AUT	0.156	0.016	0.965	0.907
	MUT	0.789	0.008	0.176	0.178
	BIC	0.200	0.208	0.625	0.424
	LAM	0.102	0.665	0.969	0.196

Table 2: The  $p$ -values of the test for nonlinearity with different resampling techniques and statistics for the 6 volume time series of world markets as given in the first column. For the statistics AUT, MUT and BIC,  $\tau=1$  and for LAM  $m=3$ . The  $p$ -values are computed for the two-sided test based on 1000 resampled data (surrogate or bootstrap) and using the distribution-free approach.

stationarity in variance or as an effect of an ARCH underlying process. The results of the test for nonlinearity are shown in Table 3. Overall, there is little evidence for nonlinearity, regardless of the resampling technique and statistic used for the test.

All algorithms generate proper resampled time series with only AAFT failing in the case of Dow Jones and S&P500 (the respective  $p$ -values of AUT in Table 3 are below 0.05). All algorithms give clear rejection of  $H_0$  with the BIC statistic ( $\tau=1$ ) for ASE and all but the AAFT algorithms give marginal rejection with LAM ( $m=3$ ) for FTSE100. The return time series have approximately normal distribution

index	statistic	Resampling techniques			
		AAFT	IAAFT	STAP	ARboot
Dow Jones	AUT	0.026	0.715	0.857	0.997
	MUT	0.859	0.567	0.719	0.392
	BIC	0.182	0.180	0.226	0.246
	LAM	0.072	0.244	0.318	0.212
NASDAQ	AUT	0.438	0.426	0.829	0.863
	MUT	0.252	0.440	0.472	0.733
	BIC	0.513	0.482	0.565	0.639
	LAM	0.801	0.989	0.865	0.899
S&P 500	AUT	0.042	0.937	0.863	0.923
	MUT	0.979	0.404	0.573	0.274
	BIC	0.268	0.312	0.356	0.388
	LAM	0.408	0.847	0.863	0.769
ASE	AUT	0.240	0.188	0.791	0.857
	MUT	0.102	0.228	0.288	0.751
	BIC	0.000	0.002	0.000	0.002
	LAM	0.154	0.246	0.330	0.368
FTSE100	AUT	0.226	0.232	0.905	0.889
	MUT	0.549	0.833	0.977	0.484
	BIC	0.757	0.717	0.795	0.985
	LAM	0.138	0.036	0.018	0.076
HSI	AUT	0.993	0.745	0.929	0.871
	MUT	0.258	0.208	0.410	0.831
	BIC	0.372	0.402	0.382	0.294
	LAM	0.360	0.312	0.603	0.386

Table 3: As in Table 2 but for the 6 monthly return time series of world markets.

and the autocorrelation decreases fast to the zero level. This may explain that the resampling techniques perform similarly.

## 5 Discussion

Randomization and bootstrap tests for nonlinearity are used to form the null distribution of the nonlinear statistics because the asymptotic null distribution is either insufficient, as for the bivariate correlation statistic, or unknown, as for the statistics of mutual information and local average mapping. We considered these three test statistics that measure different nonlinear features of the time series in order to evaluate the performance of the resampling techniques. The AAFT and IAAFT

algorithms (of constrained realization type making use of Fourier-transform) and the STAP algorithm (of typical realization type making use of an AR generating model) are used to generate the so-called surrogate data for the randomization test. The ARboot algorithm (residual-based bootstrap) is used to generate the bootstrap time series. All resampled data have to fulfill the two conditions underlined by the  $H_0$  of linear stochastic process, namely to preserve the linear structure and marginal distribution of the original time series.

It has been shown that AAFT, IAAFT and ARboot do not always fulfill the two conditions of  $H_0$ . AAFT and IAAFT may give bias in the estimation of the autocorrelation of the original time series and ARboot may give bias in the estimation of the original marginal distribution. However, the exact source of these shortcomings cannot always be determined and it seems that it depends on the system. The problem of ARboot tends to occur when the time series has heavily skewed distribution, probably due to the replacement principle of bootstrap when drawing from AR model residuals. Our simulations on different linear and nonlinear stochastic systems showed that the algorithms of AAFT, IAAFT and ARboot may be inappropriate in generating resampled data that can preserve both the marginal distribution and the linear structure of the original time series at different cases each. Mismatch of either of the two data attributes favors rejection of  $H_0$  and gives large actual test size and false test power. On the other hand, STAP matches both features but with a variance in the estimation of the original autocorrelation that is somehow larger than for AAFT and much larger than for IAAFT. This results in larger variance of the nonlinear statistics that decreases with the increase of the time series length. Consequently, the actual size of the test with STAP is smaller than the nominal size and the power is smaller than for AAFT and IAAFT (in the cases they fulfill the conditions of  $H_0$ ). The power of STAP is at the level of the power of ARboot (provided it fulfills the conditions of  $H_0$ ). This can be explained by the fact that both algorithms generate resampled data based on AR-models and therefore attain the same level of variance in the estimation of the autocorrelation.

The smaller power of typical realization approaches (STAP and ARboot) compared to the constrained realization approaches (AAFT and IAAFT) when the conditions of  $H_0$  are fulfilled may be attributed to the use of non-pivotal test statistics, as pointed out by Theiler and Prichard (1996) for the test for Gaussian time series. Indeed the variance of the nonlinear test statistics is at the level of the variance of the autocorrelation, which is larger for the typical realization approaches and results to smaller power. However, the variance decreases with the increase of sample size and the two approaches converge in terms of power. The findings of this work have shown that for the correct implementation of the resampling test for nonlinearity the consistency to the two conditions of  $H_0$  is far more important than the effect of the statistics being pivotal.

The statistics are derived from measures that are defined in terms of one or more parameters, so that the test results depend on the selection of the parameters. In our simulations, all the systems have fast decreasing autocorrelation, so that a small lag for the mutual information and bicorrelation gives suitable statistics. The same yields for the local average mapping when using a small embedding dimension. For other systems, the selection of the parameters may require an exploratory study first. Anyway, the parameters should not be optimized on the original time series, as that would favor rejection of  $H_0$ . This holds in particular for the mutual information and local average mapping that involve also other method-specific parameters (the number of bins for the data discretization and the number of neighboring points, respectively).

The test for nonlinearity was applied to the daily volume and monthly returns of six international stock exchange indices. For the first type of data the autocorrelation decreases slowly (this can be taken as evidence of non-stationarity but formal unit-root tests rejected it), so that the test statistics were computed for a larger range of parameters. The problems of AAFT and IAAFT in matching the autocorrelation were observed for many of the real time series, so that rejections of  $H_0$  with nonlinear statistics in these cases were not reliable. On the other hand, ARboot fulfilled always the conditions of  $H_0$  and had similar performance to STAP, as expected from the simulation study. STAP and ARboot gave rejection of  $H_0$  for the volume and return time series of ASE and FTSE100, but only with one of the three statistics in each case (but for a range of the parameter specific to each statistic). Visual inspection of these time series does not indicate that they possess different features that could indicate departure from linearity or stationarity.

The intension of the work on the real time series was rather to pinpoint the pros and cons of the resampling techniques than to draw conclusions about nonlinearity in these data. The overall conclusion is that AAFT, IAAFT and ARboot are not always proper for the test for nonlinearity and one has to assure first that they fulfill the conditions of  $H_0$ . On the other hand, STAP generates always proper resampled time series for the test, but the test with STAP may be at cases conservative. Regarding the three statistics used in this work, the mutual information seems to give generally the largest power when the time series length is sufficiently large (more than a couple of hundreds of samples). The bicorrelation performs at cases better than the mutual information, and the local average mapping gives largest power when there is deterministic nonlinear dynamics in the examined time series.

## References

- Barnett, A. G. and R. C. Wolff (2005): "A time-domain test for some types of nonlinearity," *IEEE Transactions on Signal Processing*, 53, 26–33.
- Brock, W., W. Dechert, and J. Scheinkman (1996): "A test for independence based on the correlation dimension," *Econometric Reviews*, 15, 197–235.
- Brooks, C. and O. T. Henry (2000): "Can portmanteau nonlinearity tests serve as general mis-specification tests?: Evidence from symmetric and asymmetric GARCH models," *Economics Letters*, 67, 245–251.
- Brooks, C. and M. J. Hinich (2001): "Bicorrelations and cross-bicorrelations as non-linearity tests and tools for exchange rate forecasting," *Journal of Forecasting*, 20, 181–196.
- Brzozowska-Rup, K. and A. Orłowski (2004): "Application of bootstrap to detecting chaos in financial time series," *Physica A*, 344, 317–321.
- Chan, W.-S. and M.-W. Ng (2004): "Robustness of alternative non-linearity tests for SETAR models," *Journal of Forecasting*, 23, 215–231.
- Chen, R. and L. Liu (2001): "Functional coefficient autoregressive models: Estimation and tests of hypotheses," *Journal of Time Series Analysis*, 22, 151–174.
- Cromwell, J. B., W. C. Labys, and M. Terazza (1994): *Univariate Tests for Time Series Models*, number 07–099 in Sage University Paper Series on Quantitative Applications in the Social Sciences, Thousand Oaks, CA: Sage.
- Dagum, E. B. and S. Giannerini (2006): "A critical investigation on detrending procedures for non-linear processes," *Journal of Macroeconomics*, 28, 175–191.
- Davies, N. and J. D. Petrucci (1986): "Detecting non-linearity in time series," *The Statistician*, 35, 271–280.
- Diks, C. (2000): *Nonlinear Time Series Analysis: Methods and Applications*, World Scientific.
- Diks, C. and S. Manzan (2002): "Tests for serial independence and linearity based on correlation integrals," *Studies in Nonlinear Dynamics and Econometrics*, 6, 1–22.
- Fan, J. and Q. Yao (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer Verlag.

- Fernández-Rodríguez, F., S. Sosvilla-Rivero, and J. Andrada-Félix (2005): “Testing chaotic dynamics via Lyapunov exponents,” *Journal of Applied Econometrics*, 20, 911–930.
- Granger, C. W. J. and T. Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford: Oxford University Press.
- Greene, W. H. (2007): *Econometric Analysis*, New Jersey: Prentice Hall, 6th edition.
- Hamilton, J. D. (2001): “A parametric approach to flexible nonlinear inference,” *Econometrica*, 69, 537–573.
- Hinich, M. J. (1982): “Testing for gaussianity and linearity of a stationary time series,” *Journal of Time Series Analysis*, 3, 169–176.
- Hinich, M. J. (1996): “Testing for dependence in the input to a linear time series model,” *Journal of Nonparametric Statistics*, 6, 205–221.
- Hinich, M. J., E. M. Mendes, and L. Stone (2005): “Detecting nonlinearity in time series: Surrogate and bootstrap approaches,” *Studies in Nonlinear Dynamics & Econometrics*, 9, 1–13.
- Hjellvik, V. and D. Tjøstheim (1995): “Nonparametric tests of linearity for time series,” *Biometrika*, 82, 351–368.
- Hjellvik, V. and D. Tjøstheim (1996): “Nonparametric statistics for testing of linearity and serial independence,” *Journal of Nonparametric Statistics*, 6, 223–251.
- Hong, Y. and Y.-J. Lee (2005): “Generalized spectral tests for conditional mean models in time series with conditional heteroscedasticity of unknown form,” *Review of Economic Studies*, 72, 499–541.
- Jašić, T. and D. Wood (2006): “Testing for efficiency and non-linearity in market and natural time series,” *Journal of Applied Statistics*, 33, 113–138.
- Kantz, H. and T. Schreiber (1997): *Nonlinear Time Series Analysis*, Cambridge: Cambridge University Press.
- Keenan, D. M. (1985): “A Tukey non-additivity-type test for time series nonlinearity,” *Biometrika*, 72, 39–44.
- Kugiumtzis, D. (1999): “Test your surrogate data before you test for nonlinearity,” *Physical Review E*, 60, 2808 – 2816.

- Kugiumtzis, D. (2000): “Surrogate data test for nonlinearity including non-monotonic transforms,” *Physical Review E*, 62, 25 – 28.
- Kugiumtzis, D. (2001): “On the reliability of the surrogate data test for nonlinearity in the analysis of noisy time series,” *International Journal of Bifurcation and Chaos*, 11, 1881 – 1896.
- Kugiumtzis, D. (2002a): “Statically transformed autoregressive process and surrogate data test for nonlinearity,” *Physical Review E*, 66, 025201.
- Kugiumtzis, D. (2002b): “Surrogate data test on time series,” in A. Soofi and L. Cao, eds., *Modelling and Forecasting Financial Data, Techniques of Non-linear Dynamics*, Kluwer Academic Publishers, chapter 12, 267 – 282.
- Kyrtsou, C. (2005): “Evidence for neglected linearity in noisy chaotic models,” *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 15, 3391–3394.
- Lee, Y.-S., T.-H. Kim, and P. Newbold (2005): “Spurious nonlinear regression in econometrics,” *Economics Letters*, 87, 301–306.
- Liu, R. Y. (1998): “Bootstrap procedure under some non-i.i.d. models,” *Annals of Statistics*, 16, 1696–1708.
- Luukkonen, R., P. Saikkonen, and T. Teräsvirta (1988a): “Testing linearity against smooth transition autoregressive models,” *Biometrika*, 75, 491–499.
- Luukkonen, R., P. Saikkonen, and T. Teräsvirta (1988b): “Testing linearity in univariate time series,” *Scandinavian Journal of Statistics*, 15, 161–175.
- Mammen, E. and S. Nandi (2004): “Change of the nature of a test when surrogate data are applied,” *Physical Review E*, 70, 016121.
- McLeod, A. I. and W. K. Li (1983): “Diagnostic checking ARMA time series models using squared-residual autocorrelations,” *Journal of Time Series Analysis*, 4, 269–273.
- Patterson, D. M. and R. A. Ashley (2000): *A Nonlinear Time Series Workshop – A Toolkit for Detecting and Identifying Nonlinear Serial*, Norwell: Kluwer Academic Publishers.
- Politis, D. N. (2003): “The impact of bootstrap methods on time series analysis,” *Statistical Science*, 18, 219–230.

- Ramsey, J. B. (1969): “Tests for specification errors in classical linear least squares regression analysis,” *Journal of the Royal Statistical Society B*, 31, 350–371.
- Schreiber, T. and A. Schmitz (1996): “Improved surrogate data for nonlinearity tests,” *Physical Review Letters*, 77, 635 – 638.
- Schreiber, T. and A. Schmitz (1997): “Discrimination power of measures for nonlinearity in a time series,” *Physical Review E*, 55, 5443 – 5447.
- Schreiber, T. and A. Schmitz (2000): “Surrogate time series,” *Physica D*, 142, 346 – 382.
- Teräsvirta, T. (1994): “Specification, estimation, and evaluation of smooth transition autoregressive models,” *Journal of the American Statistical Association*, 89, 208–218.
- Teräsvirta, T., C. Lin, and C. Granger (1993): “Power of the neural network linearity test,” *Journal of Time Series Analysis*, 14, 209–220.
- Theiler, J., S. Eubank, A. Longtin, and B. Galdrikian (1992): “Testing for nonlinearity in time series: the method of surrogate data,” *Physica D*, 58, 77 – 94.
- Theiler, J. and D. Prichard (1996): “Constrained realization Monte-Carlo method for hypothesis testing,” *Physica D*, 94, 221 – 235.
- Tong, H. (1990): *Non-linear Time Series: A Dynamical System Approach*, New York: Oxford University Press.
- Tsay, R. S. (1986): “Nonlinearity tests for time series,” *Biometrika*, 73, 461–466.
- White, H. (1989): “An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks,” Proceedings of the International Joint Conference on Neural Networks, Washington, DC: IEEE Press, New York, NY.
- Wolff, R., Q. Yao, and H. Tong (2004): “Statistical tests for Lyapunov exponents of deterministic systems,” *Studies in Nonlinear Dynamics and Econometrics*, 8, 1–17.
- Wu, C. F. J. (2006): “Jackknife, bootstrap and other resampling methods in regression analysis,” *Annals of Statistics*, 14, 1261–1295.
- Yao, Q. and H. Tong (1998): “A bootstrap detection for operational determinism,” *Physica D*, 115, 49–55.

- Yuan, J. (2000): “Testing linearity for stationary time series using the sample interquartile range,” *Journal of Time Series Analysis*, 21, 713–722.
- Ziehmann, C., L. A. Smith, and J. Kurths (1999): “The bootstrap and Lyapunov exponents in deterministic chaos,” *Physica D*, 126, 49–59.