Contribution for the book
"Nonlinear Deterministic Modelling and Forecasting
of Economic and Financial Time Series"
by A. Soofi and L. Cao
8 November 2000

Surrogate Data Test on Time Series

D. Kugiumtzis
Department of Statistics, University of Glasgow,
Glasgow G12 8QW, UK

Key word index: *time series, nonlinearity, surrogate data test*

Abstract

Given a real random-like time series, the first question to answer is whether the data carry any information over time, i.e. whether the successive samples are correlated. Using standard statistical testing, the least interesting null hypothesis of white noise has to be rejected if the analysis of the time series should be of any use at all. Further, if nonlinear methods are to be used, e.g. a sophisticated nonlinear prediction method instead of a linear autoregressive model, the null hypothesis to be rejected is that the data involve only temporal linear correlations and are otherwise random. A statistically rigorous framework for such tests is provided by the method of surrogate data. The surrogate data, generated to represent the null hypothesis, are compared to the original data under a nonlinear discriminating statistic in order to reject or approve the null hypothesis.

The surrogate data test for nonlinearity has become popular in the last years, especially with regard to the null hypothesis that the examined time series is generated by a Gaussian (linear) process undergoing a possibly nonlinear static transform. Properly designed surrogate data for this null hypothesis should possess the same autocorrelation and amplitude distribution as the original data and be otherwise random. However, the algorithms do not always provide surrogate data that preserve the original linear correlations, and this can lead to false rejections. The rejection of the null hypothesis may also depend on the applied nonlinear method and the choice of the method's parameters. Also, different observed time series from the same system may give different test results.

This chapter will describe the surrogate data test for the two hypotheses, i.e. white noise data and linear stochastic data. For the latter, some of the limitations and caveats of the test will be discussed and techniques to improve the robustness and reliability of the test will be reviewed. Finally, the tests will be applied to some financial data sets.

1 Introduction

Before building a model for the data, e.g. for prediction purposes, it is advisable to check whether the data alone suggest this type of modelling. Why put effort

on modelling economic relationships at all if the pertaining data cannot be distinguished from white noise? Or why use advanced nonlinear prediction models if there is no evidence of nonlinear structure in the data? New methods based on chaos and nonlinear dynamics theory have been used in finance in recent years with dubious success, and the validation of the results requires at least the rejection of statistical tests with less appealing null hypotheses for the observed data [1]. Here, we deal with two fundamental null hypotheses, first that the time series is simply white noise, and second, that it involves only temporal linear correlations and is otherwise random.

When an alternative hypothesis is not specified it is often hard to find test statistics with known analytic distributions. This is true particularly for the second hypothesis test, referred to as the *test for nonlinearity* . Then Monte Carlo simulation is used to build up the empirical distribution of the selected statistic, formed by the calculated values of the statistic on an ensemble of data sets representing the null hypothesis, the so-called *surrogate data* [2].

Surrogate data testing has become an indispensable tool of nonlinear time series analysis in many fields, such as physiology and geophysics, but it has not yet made its way into finance (e.g. see [3, 4, 5]). The application of the test is not without problems and some pitfalls have been reported in the last years with regard to the generation of the surrogate data, the nonlinear discriminating statistics, and certain characteristics of the time series, such as stationarity [5, 6, 7].

While the surrogate data test for white noise is not of any particular interest in many real applications (because the data almost always involve some kind of correlations), it is appropriate for many financial time series, which after removing the trend may look like white noise, e.g. the first differences of exchange rates. For financial time series with evident correlations, such as the volatility data, the surrogate data test for nonlinearity can be applied to assess the presence of nonlinear correlations and subsequently model the nonlinear structure and attempt better forecasts.

The two surrogate data tests are presented in Section 2. An extended discussion on the test for nonlinearity follows in Section 3 including the algorithms for the generation of surrogate data, the discriminating statistics, as well as the limitations and pitfalls in the implementation of the test. Then the application of the two tests on some financial data is presented in Section 4.

2 The Surrogate Data Test

Statistical hypothesis testing essentially consists of the null hypothesis H_0 and the discriminating statistic q. For time series, H_0 assigns a well-defined process to the observed data. We deal here with composite null hypotheses meaning that a family of processes rather than a single process is considered (i.e. exact values for the parameters of the process are not specified). The interest lies in finding the process inadequate for the given data, and then concluding that it is unlikely that

the observed data are generated by such a process.

The discriminating statistic q is a single number estimate of a characteristic of the data and its variation is such that it allows us to decide whether the time series is consistent with H_0 or not. When the distribution of q under H_0 is known analytically, the rejection region is at the tails of the distribution according to a given significance level α [8]. Otherwise, the distribution is formed numerically (through Monte Carlo simulation) from the values $q^1, \ldots, q^M$ of the statistic computed on an ensemble of M surrogate data consistent with H_0. The surrogate data are generated either from a model for the process in H_0 extracted from the observed data, termed the *typical realisation* approach, or as random data that preserve certain structures determined by H_0, called as the *constrained realisation* approach.

Using surrogate data, rejection of H_0 is determined using either the rank ordering (checking whether q^0 on the original data is first or last in the ordered concatenated list of $q^0, q^1, \ldots, q^M$) or the significance S (provided that $q^1, \ldots, q^M$ are fairly normally distributed) defined as

$$S = \frac{|q^0 - \langle q \rangle|}{\sigma_q}, \tag{1}$$

where $\langle q \rangle$ is the average and σ_q the standard deviation (SD) of $q^1, \ldots, q^M$. Significance of about 2 suggests the rejection of H_0 at the 95% level of confidence when, roughly, $M > 30$.

2.1 Null hypothesis of white noise

Modelling and prediction is based on statistical dependences of a present point of the time series on its past. When the data appear to be random, it is advisable to start with a test for the existence of statistical dependences, or correlations as are commonly called in the context of time series. The working null hypothesis H_0 is that the observed time series $\{x_i\}$, $i = 1, \ldots, N$, is uncorrelated, i.e. it is white noise with an unspecified distribution. Many standard independence tests, including the Brock, Dechert, and Scheinkman (BDS) test, make use of test statistics with known analytic distributions [8], while for the surrogate data testing any statistic that measures correlations can be used in principle.

The surrogate data for this H_0 are simply permutations of the original data, the so–called *scrambled* surrogates. This is the constrained realisation approach, generating random data having exactly the same amplitude distribution as the original. A typical realisation approach would estimate a model for the distribution of the original data and draw the samples of the surrogate time series from this distribution model. Obviously, the typical realisation approach is far more tedious for this H_0.

The surrogate data test can be combined with the standard analytic test in order to gain stronger evidence for rejecting or approving H_0, by double checking the test statistic q^0 with regard both to the analytic and empirical bounds. This can

be advantageous when q does not lie within the expected analytic distribution for reasons other than inconsistency with H_0, e.g. due to small data size.

As an example, say we want to investigate whether a linear autoregressive (AR) model is sufficient for modelling the chaotic Lorenz system corrupted with substantial observational noise [9]. Certainly, the residuals of the linear fit would contain the unexplained nonlinear structure, but in the presence of noise this information may not be detectable, so that a linear model may be equivalent (and simpler) to a nonlinear one. To test this, normal white noise is added to 1000 samples of the x-variable of the Lorenz system with SD of 10% and 20% of the SD of the data (the sampling time is 0.1 time units). The order of the AR model is set to 5, being the best trade-off between Akaike's information criterion and Rissanen's minimum description length for different levels of noise [10]. The two residual time series are then tested for independence using two test statistics, the turning point statistic q_{TUP}, restricted to distinguish white noise only from linearly correlated data, and the BDS statistic q_{BDS}, which is believed to have good power for any type of correlation (for a description of these tests, see e.g. [8]). Both test statistics are designed to have the standard normal distribution under H_0 and thus the rejection region at $\alpha = 0.05$ is $|q^0| > 1.96$. We compute the two statistics also on $M = 40$ scrambled surrogates and the results for the absolute values of the statistics are shown in Fig. 1a and Fig. 1b. The 10% noise level does not seem to mask the nonlinear structure contained in the residuals, and H_0 is clearly rejected by q_{BDS}, but not by q_{TUP}. When the noise level rises to 20%, H_0 is not rejected by either q_{BDS} or q_{TUP}. In this case, q^0_{BDS} and q^0_{TUP} are within their analytic and empirical distributions under H_0. Note that the empirical distribution, formed by the statistics on the surrogate data, does not align with the analytic distribution, e.g. for BDS the spread depends on the embedding dimension m while for the analytic distribution it is fixed for all m (in [3] critical values other than the standard normal are provided when $N/m < 200$).

Thus the surrogate data test seems to be more rigorous than the analytic test in that it takes care of aspects of the data other than the tested ones, such as sensitivity to certain parameters when the data size is small.

2.2 Null hypothesis of linear stochastic data

In the example above, rather than testing the residuals of a linear fit for correlations, it is more suitable to test directly the noisy time series for nonlinear correlations. The simplest H_0 for such a test is that the time series $\{x_i\}$ is generated by a normal (and thus linear) stochastic process. The surrogate data of the constrained realisation type for this H_0 are generated by phase randomisation and are often called *Fourier transform* (FT) *surrogates* [2]. For the typical realisation approach the surrogate data would be realisations of a linear model with normal input noise, such as an AR model, extracted from the original data. It was shown in [11] that the latter approach gives less powerful tests. In any case, a surrogate data set $\{z_i\}$ should preserve the original linear correlations, and in terms of the autocorrelation r this

reads $r_z(\tau) = r_x(\tau)$ for a sufficiently large range of lags τ. Proper discriminating statistics for this H_0 are in principle any nonlinear statistics, e.g. a nonlinear polynomial fit.

The H_0 as formulated above assumes that the amplitude distribution of the original data is normal, and the surrogates possess a normal amplitude distribution by construction. This may give rise to erroneous implementation of the test if the original amplitude distribution deviates from normality and the selected statistic is sensitive to the spatial distribution of the data. A more appropriate H_0 is that the time series is generated by a normal stochastic process undergoing a static, possibly nonlinear, transform, allowing in this way an arbitrary amplitude distribution. The surrogate data for this H_0 should fulfill $r_z(\tau) = r_x(\tau)$, as before, and preserve also the original amplitude distribution, that is $F_z(z) = F_x(x)$ in terms of the marginal cumulative density function (cdf) F.

Let us illustrate the test for nonlinearity on the noisy Lorenz data. To generate the surrogate data for the general H_0 we use the algorithm of AAFT (to be discussed later). For the discriminating statistic we choose the fit with local average maps (LAM) using 10 neighbour points [12], $q_{\mathrm{LAM}} = \mathrm{NRMSE}$, where NRMSE is the normalised root mean squared error of the one time step in-sample prediction (NRMSE=0 suggests perfect fit and NRMSE=1 means that the fit is as good as the mean value).

From the simulations it turns out that q^0_{LAM} is significantly smaller than q^i_{LAM}, $i = 1, \ldots, M$, for a long range of embedding dimensions m even for very high noise levels (see Fig. 1c for 80% and 20% noise levels, the latter to be compared to the BDS statistic in Fig. 1b). All the time series corrupted with up to 80% noise levels pass the test with either the significance or the rank ordering criterion (e.g. for the 80% noise level, S increases from 5 at $m = 1$ to 30 at $m = 15$, and for all m, $q^0_{\mathrm{LAM}} < q^i_{\mathrm{LAM}}$, $i = 1, \ldots, M$). For even higher noise levels the data cannot confidently be distinguished from stochastic linear data.

This simple example shows that at least when the LAM statistic is used, the discriminative power of the test for nonlinearity is better than that of the BDS test for independence applied to the residuals from a linear fit.

3 Implementation of the Nonlinearity Test

Two main aspects of the surrogate data test for nonlinearity are the algorithms for the generation of the surrogate data and the nonlinear statistics. If we can assume that the examined time series has a normal marginal cdf, then FT-surrogates can be used to match almost perfectly the original linear correlations (for special cases where FT-surrogates fail, see [13, 14, 11, 15, 16]). If normality cannot be assumed, one can transform the original data to have a normal marginal cdf, and then use the test with FT-surrogates on the transformed data [12, 14]. The so-called "gaussianisation" transform is done by rank ordering a normal white noise series to match the rank order of the original time series. However, it is not clear

why the results of this test are valid for the original time series. Moreover, the "gaussianisation" transform is monotonic whilst the general H_0 refers to any static transform. Actually, the assumption of monotonicity of the transform has been a key issue for the construction of algorithms for the generation of surrogate data.

3.1 Algorithms for the generation of surrogate data

It is in practice difficult to generate a random time series $\{z_i\}$ possessing both a given marginal cdf $F_x(x)$ and a given autocorrelation $r_x(\tau)$ as required by the general H_0. A perfect match of one condition results in possible deviation for the other. It is computationally easier to obtain $F_z(z) = F_x(x)$ and bear small discrepancies in the linear correlations, i.e. $r_z(\tau) \simeq r_x(\tau)$. This approach is preferred in all the following algorithms.

The first and most prominent algorithm is the amplitude adjusted Fourier transform (AAFT). This algorithm uses the "gaussianisation" transform to make the marginal cdf normal, then applies phase randomisation to remove any nonlinear structure without altering the linear correlations, and finally makes the "inverse-gaussianisation" transform to regain the original marginal cdf [2, 17]. AAFT has been the algorithm of choice in almost all real applications so far. However, it does not perform properly in general and fails to match well the linear correlations. This was already observed in the early stage of its use [14]. The reason for this is the inherent assumption in AAFT that the static transform in H_0 is monotonic, which cannot be assumed when dealing with real data [18]. Bias in the linear correlations can arise when the original data are consistent with H_0 (but the transform is non-monotonic), but also when they are not (nonlinear dynamics are present), and then it favours the rejection of H_0 when a nonlinear statistic sensitive to the linear correlations is applied [19].

A better approximation of the original linear correlations is achieved by an iterative scheme made in two steps, possessing the original linear correlations at the first step and the original marginal cdf at the second step [20]. This rather empirical algorithm, called iterative or improved AAFT (IAAFT), approximates well the original linear correlations. However, this approximation always has the same direction (i.e. the surrogate data are less correlated than the original data), and it is equally good for each surrogate data generation, resulting in small bias and variance. Even when the bias in the linear correlations is very small, when combined with small variance, it may turn out to be significant. Another iterative algorithm, suitable also for more specific constraints, makes use of simulation annealing and provides surrogates of about the same quality as IAAFT, but requires long computation time [21].

The last algorithm is the corrected AAFT (CAAFT), which corrects for the bias in the autocorrelation of AAFT using a typical realisation approach. A time series is generated by an appropriate AR model and is transformed through "inverse gaussianisation" to possess the original cdf exactly and the original linear correlations on average [19]. This approach involves two free parameters, the order of the AR

model and the number of trials needed in order to find the "best" model (the last parameter is simply set equal to the number of surrogates to be generated). The CAAFT surrogates estimate the original linear correlations without bias but usually with a larger variance than for the AAFT surrogates. As a result, the nonlinear statistics using CAAFT are often more spread than when using AAFT (the least spread is obtained with IAAFT). This makes the test with CAAFT generally more conservative than that with the other two algorithms.

3.2 Discriminating Statistics

A number of nonlinear measures has been proposed in the literature and most of them stem from the theory of dynamical systems and rely on state space reconstruction (making use of two parameters, the embedding dimension m and the delay τ). Such methods are the correlation dimension estimate, the largest Lyapunov exponent, the mutual information and the embedding dimension estimate by means of false nearest neighbours [22]. Each measure is sensitive to a nonlinear characteristic of the data. For example, the fit with LAM, used in Section 2, detects local spatio-temporal structures that support an enhanced global fit when nonlinear dynamics are present. Certainly, a local linear model would be equivalent to LAM, if not better because of better local modelling, but at the cost of longer computation time. It has been shown that the in-sample prediction error of a local map constitutes a more powerful statistic than the out-of-sample prediction error [11].

For the use of prediction measures as statistics, there is a large collection of other nonlinear models to choose from, such as neural networks and radial basis functions, but since the objective is discrimination and not best fitting, simple alternatives, such as global nonlinear polynomials, would do equally well in general. In particular, the series of Volterra polynomials has been proposed in the literature [23, 18], also because it includes linear polynomials and provides in this way a straightforward check for the preservation of the original linear correlations by the surrogates. Typically, second order terms in the polynomial form are sufficient to resolve nonlinearities, but there might be cases where cubic terms also have to be included. From all examples we have encountered, cubic terms were needed only for the Lorenz data.

Simple statistics, such as the three point autocorrelation and the time reversibility have been proposed as well [5].

There are few guidelines in the literature for the preference of a particular statistic and it seems that the properties and power of all these nonlinear statistics are not yet fully explored [24, 7]. This is mainly due to the varying performance of the statistics in different applications. Some statistics, such as the correlation dimension estimate and the false nearest neighbours, are found to be of limited use [7], while others, such as LAM, are found to have generally good discriminative power [24].

3.3 Limitations and pitfalls of the test

An important source for spurious test results is the imperfection of the algorithms for the generation of the surrogate data. Unfortunately, it turns out that the popular algorithm of AAFT is the least accurate, and the bias in the linear correlations often yields false rejections. For IAAFT, the small variance may amplify either the tiny bias in the linear correlations or the tiny insignificant differences in some nonlinear characteristics, resulting in significant discrepancies and unjustified rejection of H_0. So, the test with IAAFT surrogates is "too powerful" to such an extent that it becomes liable to spurious rejections. On the other hand, the test with the CAAFT surrogates is more conservative under the same reasoning. The unbiased estimation of the linear correlations is attained at the cost of a larger variance, and as a result, significant differences due to nonlinear characteristics may not be detected, failing to reject H_0. It should be stressed that, for CAAFT, the variance decreases as sample size increases, so it is more difficult for short time series to pass the test than it is for a longer one of the same origin. This property of CAAFT makes the uncertainty of inference increase in the presence of large sample variability, which is appropriate for a statistical test but not shared by the other two algorithms. Another effect of sample variability is the variation of the test results across different time series of the same system, as shown in [7] for subsequent electroencephalogram (EEG) recordings and different exchange rates.

When conducting the test one should be cautious about aspects in the time series irrelevant to nonlinear dynamics that can give rise to rejections, such as nonstationarity (see e.g. [6, 7]) and long coherent times (see [13, 14]).

Finally, it should be stressed that one should not attempt to optimise the parameters of the method on the original data and then use these values on the surrogate data because this would favour rejection of H_0. Also, the results of the test may be subject to small variations of the parameters of the methods [7].

4 Application to Financial Data

We apply here the two statistical tests discussed earlier to selected financial time series.

4.1 Search for correlations

Exchange rates, as well as many financial time series, are considered to be effectively random walks, and their first differences white noise. To assess this hypothesis, a test of independence can be conducted. We applied this test on the first differences of the monthly exchange rates USD/GBP from June 1978 to December 1993. In this period, the time series appears fairly stationary. We could reject the H_0 of white noise with both the turning point statistic ($|q^0_{\mathrm{TUP}}| = 3.5$, the largest $|q_{\mathrm{TUP}}|$ on the surrogates being 2.8) and the BDS statistic ($|q^0_{\mathrm{BDS}}| > 2$ and increasing for $2 < m < 10$, it is also larger than the $|q_{\mathrm{BDS}}|$ on the surrogates). The same results

were obtained when the test was applied to the weekly data in the same period, indicating that exchange rate differences at the monthly and weekly time scales contain some kind of correlations. However, when we tested the first differences of daily exchange rates, the H_0 of white noise could not be rejected, neither with q_{TUP} nor with q_{BDS}, suggesting that the day to day exchange rates are uncorrelated.

4.2 Search for nonlinear dynamics

The analysis and prediction of volatility data is of particular interest in some financial areas, e.g. risk management. In mathematical terms, the volatility is a moving average smoothing of the standardised first differences (first returns) of a financial time series. The interest here is whether the apparent correlations in the volatility data are artificial, i.e. due to smoothing, or whether they reflect inherent structures that may give rise to interesting nonlinear dynamics.

We use the same time series of weekly exchange rates and derive the volatility time series using a time window of 12 weeks. We generate AAFT, IAAFT and CAAFT surrogates and apply the test for nonlinearity using the fit with a series of Volterra polynomials up to second order and for $m = 10$. We repeat the test for different delay parameters τ and prediction times T. The general feature is that there is little evidence of nonlinearity, but the results of the test vary with the algorithm for the generation of the surrogate data and the parameters τ and T. We show the results for the three algorithms and for three combinations of τ and T in Fig. 2. The AAFT, IAAFT, and CAAFT algorithms gave large, small and no bias in the linear correlations, respectively. This can be seen from the discrepancies in NRMSE for the linear polynomials (i.e. up to the first 10 polynomial terms). If the data would have contained nonlinear correlations, one would expect that with the inclusion of nonlinear terms, the NRMSE for the original data would have dropped more than for the surrogates, but this does not seem to be the case for these data. However, only the NRMSE for the CAAFT surrogates always contain the NRMSE for the original data, resulting in a small S and no rejection of H_0 (see Fig. 2d). The statistics using the AAFT surrogates suggest marginal rejection of H_0 for all τ and T, erroneously, as the rejection holds for both linear and nonlinear polynomials. The IAAFT surrogates give more varying results, and small fluctuations in the NRMSE of the surrogates seem to change the test result as shown in Fig. 2d in terms of S. Particularly, H_0 is not rejected for any polynomial when $\tau = 5$ and $T = 1$; it is rejected paradoxically with linear polynomials but not with nonlinear polynomials of many interaction terms when $\tau = 1$ and $T = 1$; and it is confidently rejected only with nonlinear polynomials (as if nonlinear dynamics would be present) when $\tau = 1$ and $T = 3$. The same conclusions are drawn on the basis of rank ordering. Using LAM and mutual information (not shown here) the test results for AAFT and CAAFT are as for the Volterra polynomials and for IAAFT no rejections are obtained.

For all the statistics in this example, the variance using CAAFT is as large as when using AAFT, so that the lack of rejection with CAAFT cannot be attributed to

excessive variance. The results above warn about the practical complications of the surrogate data test and caution the risk of getting false rejections. The overall conclusion would be that these volatility time series is not likely to contain nonlinear dynamics.

5 Discussion

In financial time series the interest is focused on predictions. Increasing the complexity of the model one can always achieve good predictions on the data on which the parameters are optimised. This fact alone by no means suggests that this model is better than a simpler one, and independent validation requires genuine predictions on another sufficiently long time series, often not available in finance. In this context, hypothesis testing is a useful and objective indicator for the limitations of modelling and prediction.

The use of a more sophisticated or interesting model cannot be justified if the hypothesis that the data are consistent with a simpler model is not rejected. Thus, failure to reject the null hypothesis for independence questions any achieved predictions and assigns them to data over-fitting. Accordingly, failure to reject the null hypothesis of a linear stochastic model for a given time series invalidates any nonlinear modelling and prediction on this time series.

On the other hand, the rejection of the hypothesis of a simple uninteresting model does not validate *per se* the model we have in mind as the only alternative. For the surrogate data test for nonlinearity on an econometric time series, a confident rejection does not necessarily imply chaos in the observed market, appointing nonlinear deterministic modelling as the only appropriate approach. There might be other econometric models of stochastic processes able to explain this market's behaviour, which do not belong to the rejected class of univariate linear stochastic processes. However, one should not be left with the impression that rejection of this null hypothesis is not worth the effort because it opens the way for nonlinear analysis and validates the use of promising techniques to pursue predictions.

Before a final rejection or approval of the null hypothesis is asserted, a number of aspects has to be taken care of with regard to the set-up of the surrogate data test. For the test of independence, surrogate data are useful because they provide empirical rejection regions for the statistic, which are often more accurate than the analytic ones. The effectiveness of this test relies mainly on the discriminative power of the chosen test statistic and it turns out that the BDS statistic performs the best. However, for the hypothesis of a linear normal stochastic process, the BDS statistic applied to the residuals from a fitted linear model is substantially less powerful than a nonlinear statistic applied directly to the data, such as the local average map.

For the general hypothesis of a staticly distorted normal stochastic process, the quality of the surrogate data turns out to be of immense importance. A bias in the linear correlations of the generated surrogate data can often give rise to artificial

differences in the statistic on the original and surrogate data, leading to false rejections. The prominent algorithm of amplitude adjusted Fourier transform (AAFT) turns out to be the one that gives the largest bias. The test with the iterated AAFT (IAAFT) tends generally to have more power but at the risk of being biased towards rejection because of a small, but possibly significant, bias in the linear correlations. On the other hand, the corrected AAFT (CAAFT) gives no bias but larger variance, which makes the test less powerful but the rejection more reliable, when obtained.

The power of the test is also determined by the choice of the nonlinear method that gives the statistic. It is difficult to provide conclusive results for the power of each statistic because it changes with different data types. Actually, for the same data set the outcome of the test may vary with the method and also with the parameters of the method. So, one should accumulate enough statistical evidence from different methods and parameters to confidently accept or reject the hypothesis. In this respect, a single test with an arbitrary statistic should be regarded insufficient. Note that the lack of robustness of the test with different parameters and statistics may often be due to the inaccuracies in the algorithms generating the surrogate data. In general, we believe it is better to have a conservative test (using CAAFT) rather than a possibly biased test (using AAFT). Also, the statistics from the fit with Volterra polynomials can identify these inaccuracies and therefore the use of this statistic is strongly recommended.

The quality of the data could be another source of variation in the test results. One should first question whether the time series is representative of the system under study. This question is particularly relevant in financial applications where there is a mass of time series regarding the same economic system. Apart from noise, the data may contain certain characteristics that can be mistaken for dynamic nonlinearities. For example a couple of bursts in an econometric time series attributed to exogenous sudden events (such as strikes, earthquakes or currency devaluations) may give grounds for rejections and further claims for nonlinear dynamics.

Finally, it should be stressed that the surrogate data approach has the potential for devising other hypotheses that would be of interest in finance, e.g. involving Markov chain processes as attempted in [4].

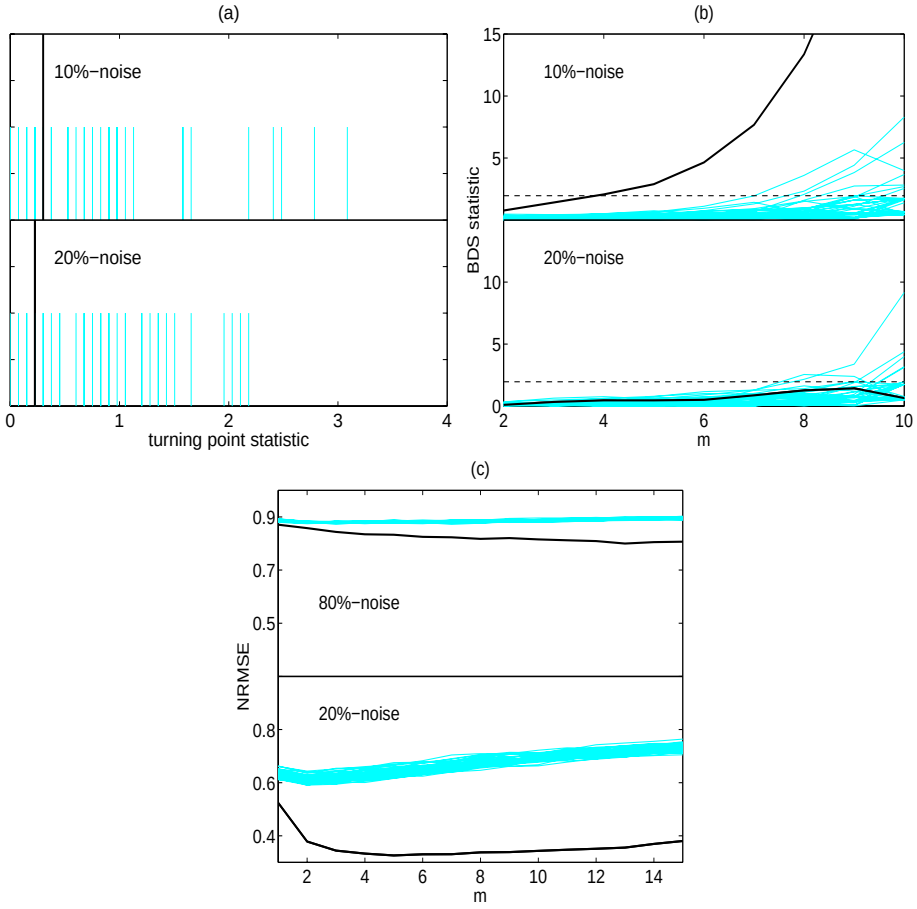

Figure 1: (a) The $|q_{\mathrm{TUP}}|$ statistic for the residuals of the linear fit of the Lorenz data, corrupted with 10% and 20% of white noise, at the upper and lower panel, respectively. The large black bar is for the original data and the grey bars for the 40 scrambled surrogates (some coincide). (b) The $|q_{\mathrm{BDS}}|$ statistic as a function of the embedding dimension m for the same data as in (a) (the distance in the computation of the correlation integral is set to half of the SD of the data). The thick black line is for the original data and the grey lines for the scrambled surrogates. The dashed horizontal line denotes the threshold for $\alpha = 0.05$. (c) The q_{LAM} statistic as a function of m for the Lorenz data corrupted with 80% and 20% of white noise at the upper and lower panel, respectively. The thick black line is for the original data and the grey lines for the 40 AAFT surrogates.

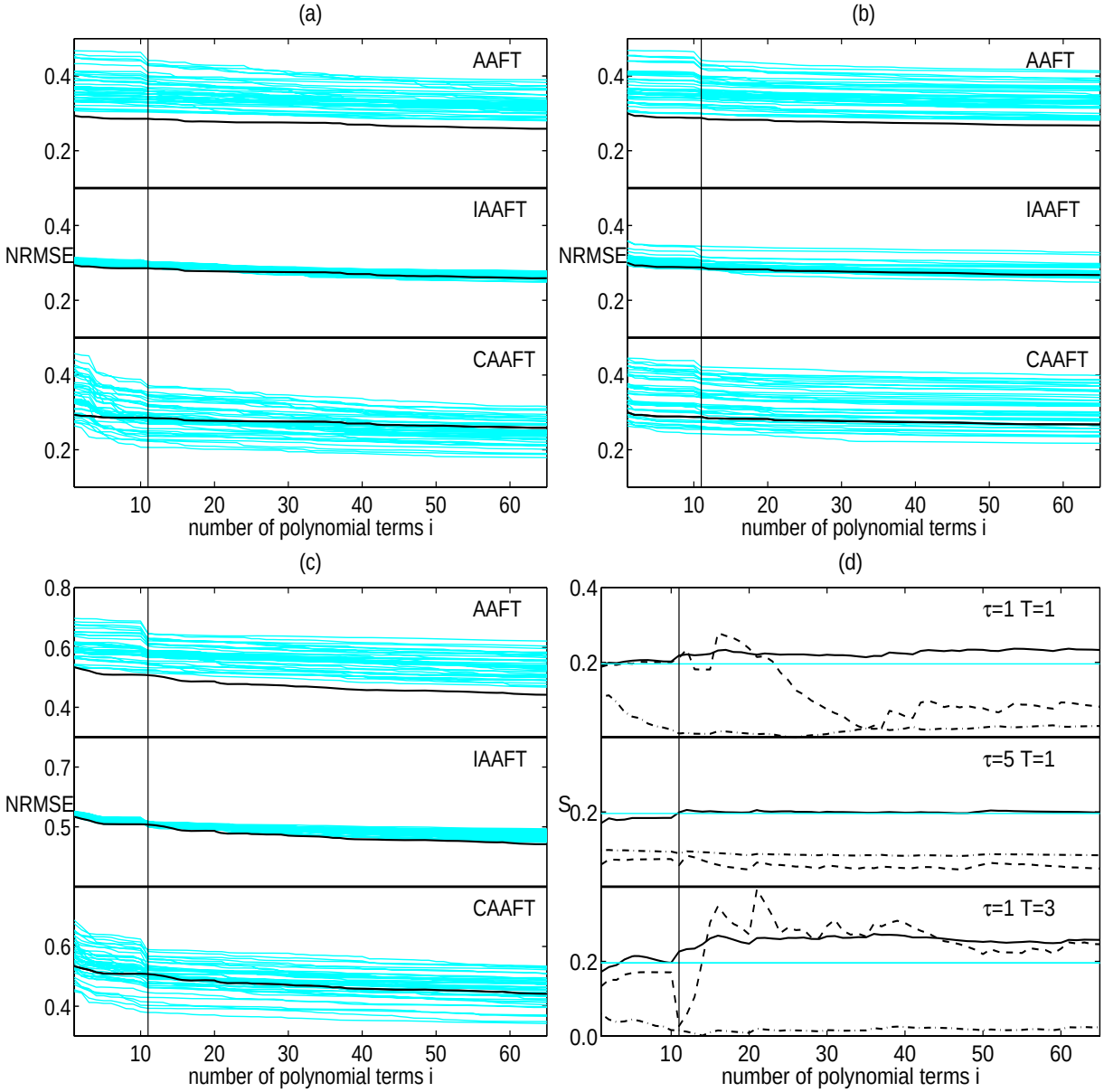

Figure 2: (a) The NRMSE of the fit with Volterra polynomials ($m = 10$) as a function of the polynomial terms i, for the volatility exchange rate data (black line) and its AAFT, IAAFT and CAAFT surrogates (grey lines in the upper, middle and lower panel, respectively). The free parameters are $\tau = 1$ and $T = 1$. (b) The same as in (a) but for $\tau = 5$ and $T = 1$. (c) The same as in (a) but for $\tau = 1$ and $T = 3$. (d) The significance of the NRMSE statistics in (a), (b) and (c) shown in the upper, middle and lower panel, respectively. The solid line is for the AAFT surrogates, the dashed line for the IAAFT surrogates and the dashed-dotted line for the CAAFT surrogates. In all panels, the vertical line distinguishes linear from nonlinear polynomial terms.

References

[1] R. M. A. Urbach. *Footprints of Chaos in the Markets: Analyzing Non-linear Time Series in Financial Markets and other Real Systems*. Prentice Hall Publishing, 2000.

[2] J. Theiler, S. Eubank, A. Longtin, and B. Galdrikian. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58:77 – 94, 1992.

[3] W. A. Brock, D. A. Hsieh, and B. LeBaron. *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence*. Massachusetts Institute of Technology, 1991.

[4] G. Deco, C. Schittenkopf, and B. Schürmann. Dynamical analysis of time series by statistical tests. *International Journal of Bifurcation and Chaos*, 7(12):2629 – 2652, 1997.

[5] T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142(3-4):346 – 382, 2000.

[6] J. Timmer. Power of surrogate data testing with respect to nonstationarity. *Physical Review E*, 58(4):5153 – 5156, 1998.

[7] D. Kugiumtzis. On the reliability of the surrogate data test for nonlinearity in the analysis of noisy time series. to appear in International Journal of Bifurcation and Chaos.

[8] J. B Cromwell, W. C. Labys, and M. Terazza. *Univariate Tests for Time Series Models*. Number 07–099 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage, Thousand Oaks, CA, 1994.

[9] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20:130, 1963.

[10] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag, New York, 1991.

[11] J. Theiler and D. Prichard. Constrained realization Monte-Carlo method for hypothesis testing. *Physica D*, 94:221 – 235, 1996.

[12] M. B. Kennel, R. Brown, and H. D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45:3403 – 3411, 1992.

[13] J. Theiler, P. S. Linsay, and D. M. Rubin. Detecting nonlinearity in data with long coherent times. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 429 – 455. Addison-Wesley Publishing Company, Reading, MA, 1994.

[14] M. Palŭs. Testing for nonlinearity using redundancies: quantitative and qualitative aspects. *Physica D*, 80:186 – 205, 1995.

[15] J. Theiler and D. Prichard. Using 'surrogate surrogate data' to calibrate the actual rate of false positives in tests for nonlinearity in time series. *Fields Institute Communications*, 11:99, 1997.

[16] C. J. Stam, J. P. M. Pijn, and W. S. Pritchard. Reliable detection of nonlinearity in experimental time series with strong periodic components. *Physica D*, 112:361 – 380, 1998.

[17] J. Theiler, B. Galdrikian, A. Longtin, S. Eubank, and J. D Farmer. Using surrogate data to detect nonlinearity in time series. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *SFI Studies in the Sciences of Complexity*, pages 163 – 188, Reading, MA, 1992. Addison – Wesley.

[18] D. Kugiumtzis. Test your surrogate data before you test for nonlinearity. *Physical Review E*, 60(3):2808 – 2816, 1999.

[19] D. Kugiumtzis. Surrogate data test for nonlinearity including non-monotonic transforms. *Physical Review E*, 62(1):1 – 2, 2000.

[20] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77(4):635 – 638, 1996.

[21] T. Schreiber. Constrained randomization of time series data. *Physical Review Letters*, 80(10):2105 – 2108, 1998.

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

[23] B. Barahona and Ch.-S. Poon. Detection of nonlinear dynamics in short, noisy time series. *Nature*, 381(6579):215 – 217, 1996.

[24] T. Schreiber and A. Schmitz. Discrimination power of measures for nonlinearity in a time series. *Physical Review E*, 55(5):5443 – 5447, 1997.