Testing for nonlinearity in high-dimensional time series from continuous dynamics

A. Galka [a,*], T. Ozaki [b]

[a] *Institute of Experimental and Applied Physics, University of Kiel, 24098 Kiel, Germany*
[b] *Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan*

Abstract

We address the issue of testing for nonlinearity in time series from continuous dynamics and propose a quantitative measure for nonlinearity which is based on discrete parametric modelling. The well-known problems of modelling continuous dynamical systems by discrete models are addressed by a subsampling approach. This measure should preferably be combined with conventional surrogate data testing. The performance of the test is demonstrated by application to simulated, heavily noise-contaminated time series from high-dimensional Lorenz systems, and to experimental time series from a high-dimensional mode of Taylor–Couette flow. We also discuss the discrimination power of the test under surrogate data testing, when compared with other well-tried test statistics. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Time series analysis; Nonlinearity; Autoregressive modelling; Surrogate data testing

1. Introduction

In the last two decades the analysis of time series data obtained from experiments or field observations has been a field of active research. If such time series are generated by complicated systems for which it is impossible to solve or even set up the equations governing the dynamics, it is quite commonly assumed without further proof that such time series display significant nonlinearity. Consequently the analysis is carried out by advanced numerical algorithms borrowed from nonlinear dynamics, whereas the repertory of more traditional linear tools is largely neglected.

While for many systems the assumption of nonlinearity may be correct in principle, it has for specific cases to be shown explicitly that employing nonlinear tools and models is justified and useful. If an experimental time series of limited length and finite precision is given, it may be impossible to distinguish between nonlinear dynamics and linear dynamics involving stochastic components. This has been demonstrated by Casdagli et al. [1] for the closely related case of high-dimensional determinism.

It is for this reason that tests for nonlinearity are important tools in time series analysis. Currently the technique of surrogate data testing [2] is one of the most popular approaches to nonlinearity testing. Being motivated by statistical hypothesis testing, this technique presents an indirect way of detecting nonlinearity; as a consequence of this a failure to detect nonlinear-

* Corresponding author.
E-mail address: andreas@ang-physik.uni-kiel.de (A. Galka).

ity does not disprove nonlinearity, but may also result from an inappropriate choice of the test statistic. Recently more weak points of surrogate data testing were observed, first and foremost the effect that a rejection of the null hypothesis does not necessarily prove nonlinearity, but still may be a consequence of other properties of the time series such as non-stationarity [3]. There are also considerable problems with artefacts occurring in the process of generating the surrogate data sets [4,5].

For these reasons it would be desirable to design alternative, possibly more direct, tests for nonlinearity, which avoid the generation of surrogate data sets. Such tests exist, e.g. based on the bispectrum [6], but when applied to nonlinear deterministic systems they tend to display rather poor performance.

In this paper we design a simple but efficient test statistic for nonlinearity based on discrete parametric modelling, which provides in itself a meaningful measure of nonlinearity. The technique of surrogate data testing is retained only as a tool for investigating the significance of the results of the test statistic, and for excluding the possibility of static nonlinear transforms. The new test is applied both to simulated time series from the Lorenz system and to experimental time series from Taylor–Couette flow.

We also present results of comparing three different test statistics to high-dimensional time series, contaminated by strong noise components, under conventional surrogate data testing. It is shown that depending on the dimension of the underlying dynamical system and the amount and spectral properties of the noise, different test statistics provide the highest power of discrimination.

2. Surrogate data testing

In this section we give a very brief summary of surrogate data testing; for more detailed presentations see [2,7] and references therein.

In conventional surrogate data testing a null hypothesis is stated, such as: the time series in question was generated by a stationary linear gaussian stochastic process; in the process of measurement it was possibly distorted by a monotonic static nonlinear transform (i.e. a filter changing the distribution of amplitudes, but not involving previous values of the time series). Then a set of *surrogate* time series is generated which corresponds to this null hypothesis. In this paper we use the iterative amplitude-adjusted phase randomisation (IAAPR) algorithm proposed by Schreiber and Schmitz [8] for the generation of surrogate time series. This algorithm employs the Fourier expansion as a non-parametric linear model of the original time series, and aims at simultaneously preserving the power spectrum and the amplitude distribution of the original time series, while nonlinear properties are destroyed by phase randomisation. Following the results of Theiler and Prichard [9] the periodogram is employed as an estimator of the power spectrum (*constrained-realisation approach*) instead of a consistent estimator.

Then a suitable numerical algorithm (the *test statistic*) is applied to each time series from the set of all surrogate time series and the original time series; see [10,11] for discussions on the issue of choosing a suitable test statistic. It should be stressed that the test statistic needs not be designed specifically for detecting nonlinearity, in general any numerical measure which can be computed from time series data could be employed. If the result for the original time series differs significantly from the distribution of results for the surrogate time series, the null hypothesis can be rejected. We will usually evaluate the results by simple ranking, i.e. we will reject the null hypothesis, if the result for the original is larger than any result from the surrogates. This will work since we employ only one-sided tests [12]. The more surrogates are employed, the smaller will be the size (i.e. the probability of wrong rejection of the null) of the test.

3. Nonlinear autoregressive modelling

In order to design a quantitative measure for nonlinearity we start from the assumption that *nonlinear time series can be modelled better by nonlinear models than by linear models*. Now the problem arises that nonlinearity itself is not a property, but rather the

absence of a property. Hence we have to be more specific as to our interpretation of nonlinearity, i.e. we have to specify a well-defined class of nonlinear models. Such classes of models can be based on reconstructed state spaces [10,11], but here we prefer to explore direct parametric modelling of the dynamics, i.e., given the (scalar) time series x_i, $i = 1, \ldots, N$ (which we assume to have zero mean and unit variance), we look for an autoregressive model,

$$x_i = f(x_{i-1}, \ldots, x_{i-p}) + \epsilon_i, \tag{1}$$

where p is the model order and ϵ_i represents dynamical noise. If $f(\cdot)$ is to be chosen as a linear function, we obtain a classical AR(p) model,

$$x_i = \sum_{j=1}^{p} a_j x_{i-j} + \epsilon_i =: \hat{x}_i + \epsilon_i, \tag{2}$$

where $\hat{x}_i$ denotes the prediction (or *conditional mean*) of x_i. As an example of nonlinear choices for $f(\cdot)$ polynomial models have been used for the detection of nonlinearity [5,13]. However, here we choose to employ exponential autoregressive (ExpAR) models as introduced by Ozaki and Oda [14]:

$$x_i = \sum_{j=1}^{q} \left(a_j + b_j \exp\left(\frac{-x_{i-1}^2}{h} \right) \right) x_{i-j} + \epsilon_i$$
$$=: \hat{x}_i + \epsilon_i, \tag{3}$$

where h is a suitably chosen constant bandwidth parameter. In comparison to polynomial models these models have the advantages of a much simpler model structure (which facilitates model fitting) and provable stability for appropriate choice of parameters.

The estimation of an optimum set of parameters a_j for AR models can easily be implemented by the standard approach of solving the set of (linear) equations resulting from a least-squares approach:

$$\frac{\partial}{\partial a_k} \sum_{i=p+1}^{N} (x_i - \hat{x}_i)^2 = 0, \quad k = 1, \ldots p, \tag{4}$$

where $\hat{x}_i$ is inserted from Eq. (2). For ExpAR models the same approach remains valid, if $\hat{x}_i$ is inserted from Eq. (3); in this case one set of equations for the a_k and another one for the b_k results.

The bandwidth parameter h should be chosen such that $\exp(-x_{i-1}^2/h)$ approaches zero for the largest occurring values of x_{i-1}; therefore for each time series h is estimated by

$$h = -\frac{\max(x_{i-1}^2)}{\log c}, \tag{5}$$

where c is a small number selected in advance. It has been demonstrated that this approach yields consistently good estimates for h [15]. We choose $\log c = -30.0$ (which is a reasonable choice for unit variance time series).

4. Design of the test statistic

If for given model orders p and q the optimal AR(p) or ExpAR(q) models are estimated there will still remain prediction errors giving rise to a non-zero *residual variance*

$$v_p = \frac{1}{N - p} \sum_{i=p+1}^{N} (x_i - \hat{x}_i)^2, \tag{6}$$

and similarly for ExpAR(q). Based on this residual variance, the Akaike Information Criterion (AIC) [16,17] is defined as

$$\text{AIC} = N \log v_p + 2(P + 1), \tag{7}$$

where P denotes the number of the data-adaptive parameters a_j and b_j in the model (excluding the mean). Obviously, P becomes simply p for AR(p) models, whereas for ExpAR(q) models P becomes $2q$, i.e. P has to be even for ExpAR(q) models. AIC, being an *unbiased* estimate of the expected log-likelihood (multiplied by -2) of the model, serves as a measure of the quality of the model. By minimising AIC we could estimate the optimal model order: increasing the model order will always reduce the residual variance, but it will finally begin to increase AIC, since the second term in Eq. (7) increases. Here we choose to normalise AIC by dividing by the length of the time series N.

Now we estimate AR(p) and ExpAR(q) models for $p = 2, \ldots, 40$ and $q = 2, \ldots, 20$ (i.e. $P \leq 40$) for a time series of $N = 8192$ points length sampled from
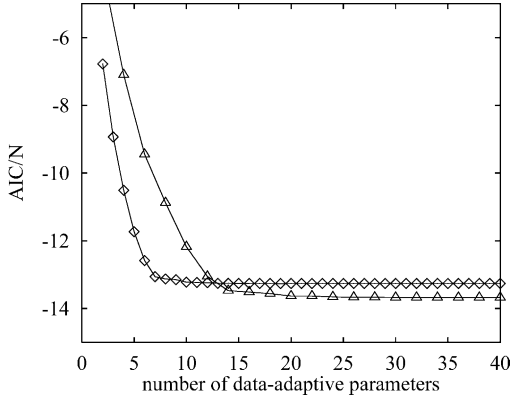
Fig. 1. Normalised AIC versus model order (i.e. number of data-adaptive parameters) for linear autoregressive (diamonds) and ExpAR (triangles) models of a time series of 8192 points length sampled from the *x*-coordinate of the Lorenz system. For ExpAR models the number of data-adaptive parameters always increases by 2.

the *x*-coordinate of the well-known Lorenz system $\dot{x}(t) = \sigma(y(t) - x(t))$, $\dot{y}(t) = rx(t) - y(t) - x(t)z(t)$, $\dot{z}(t) = x(t)y(t) - bz(t)$ at standard parameters $\sigma = 10.0$, $r = 28.0$ and $b = \frac{8}{3}$. The integration step size is 10^{-3} and the sampling step size is 25 integration steps (which is a reasonable choice for providing a good, but not too dense sampling of the trajectories). The result is shown in Fig. 1.

It can be seen that AIC/*N* decreases for increasing *P*; once a sufficiently high model order has been reached, AIC/*N* remains approximately constant (it will slowly increase again, thereby punishing the overfitting due to too high model order). For small $P \equiv p = 2q$ the linear AR(p) models achieve smaller values (i.e. better prediction), because they consider twice as long an interval of previous values for their prediction; then the nonlinear ExpAR(q) models catch up and surpass the linear models. This behaviour is to be expected, since the Lorenz time series clearly contains nonlinearity. But we note that the improvement in reduction of AIC/*N* due to the choice of a nonlinear model is not very striking, as compared to the reduction due to increasing the model order.

From Fig. 1 we can presume that the difference between the values of AIC/*N* for AR(p) and ExpAR(q) (where $q = \frac{1}{2}p$) provides a useful quantitative mea-

sure of nonlinearity, provided the model order p is sufficiently large:

$$\delta(p) := \frac{1}{N} \left(\text{AIC}(\text{AR}(p)) - \text{AIC}\left(\text{ExpAR}\left(\frac{p}{2}\right)\right)\right).$$
$$(8)$$

This test statistic bears strong similarity to those employed in more formal statistical tests for nonlinearity, namely likelihood ratio tests [18]; however, these tests cannot directly be applied to ExpAR(q) models, since they require that the linear model follows from the nonlinear model *in a unique way* by imposing suitable constraints on the parameters. But in order to transform Eq. (3) into Eq. (2) we could either demand $b_j = 0$ or $h = 0$ [19].

5. Subsampling approach

It is not difficult to find the reason for the apparently "weak" evidence for nonlinearity in the Lorenz system, as displayed in Fig. 1. The autoregressive models of Eqs. (2) and (3) implicitly assume discrete dynamics (with respect to time), whereas the Lorenz system is continuous and should rather be described by a differential equation. We could have avoided this problem by choosing a discrete system, such as the tent map or the Hénon map; in fact, most literature on detecting nonlinearity in time series employs simulated data from such discrete systems [9,11]. Certainly much can be learned from numerical experiments on time series generated by discrete systems; however, since most interesting time series from reality are generated by continuous systems, we regard such discrete maps as less relevant for practical applications.

It is not an obvious fact that the dynamics of continuous systems can be modelled by discrete models, since such models invariably will pretend that there was no trajectory of the system between the sampled points at all; nevertheless recent work of Aguirre and Billings [20] has shown that it is possible to capture all essential characteristics of continuous dynamics by well-chosen discrete models. As an example, if we choose too small a sampling step size, each prediction from Eqs. (2) and (3) will be based on only a short

time series segment, and the dynamics of such segments will essentially be linear. Consequently we have to include a longer time series segment, if we want to find clearer evidence of nonlinearity. This could be done either by choosing considerably larger model orders or by larger sampling step sizes, i.e. for given data by subsampling. Since large model orders yield unwieldy models at the cost of overfitting and high computational expense, we prefer the latter approach.

Let S denote the subsampling factor, then we estimate $AR(p)$ and $ExpAR(\frac{1}{2}p)$ models from the subsampled time series x_i, $i = S, 2S, \ldots, (\text{floor}(N/S)S)$ (where $\text{floor}(x)$ is the largest integer smaller or equal to x), such that $S = 1$ corresponds to the case without subsampling. As a disadvantage of subsampling, large amounts of data are discarded, especially for high S, which results in low statistical accuracy of the models. In order to overcome this disadvantage we propose to form more subsampled time series from the discarded values, the first of which is given by x_i, $i = 1, S+1, 2S+1, \ldots, (\text{floor}(N/S)S-S+1)$, such that finally a total of S nested time series are available, which are formally treated as independent. In a quite similar manner time-delay embedding with a delay time larger than one sampling step achieves a subsampling without any loss of data. The standard pro-

cedure for estimating $AR(p)$ and $ExpAR(\frac{1}{2}p)$ models is then slightly generalised to the case of estimating one common model for S different time series. In this case a summation over all nested time series turns up in the argument of the derivation operator in Eq. (4), and the x_i have to be relabelled according to their membership to the subsampled time series. The further implementation of the least-squares approach is straightforward.

6. Application to simulated time series

We now calculate the nonlinearity test statistic δ according to Eq. (8) for $S = 1, \ldots, 30$ and $p = 2, 4, \ldots, 30$ for the same Lorenz time series as before. The results are shown in the left panel of Fig. 2. Each curve corresponds to one value of p. It can be seen that for increasing p the curves $\delta(p, S)$ rapidly converge to a common curve, indicating that the model order is sufficiently large; this convergence corresponds to the flat regions of the curves shown in Fig. 1. Indeed we see that for increasing S the nonlinearity test statistic $\delta(p, S)$ also increases, up to more than twice its value at $S = 1$. This maximum is reached at $S = 13$, indicating that we need such heavy subsampling in order
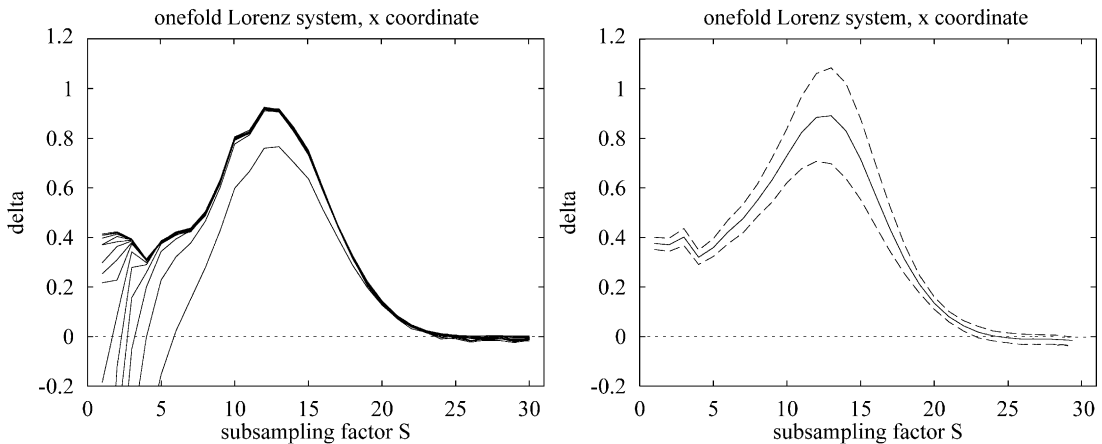

Fig. 2. Left panel: Nonlinearity test statistic δ versus subsampling factor S for model orders $p = 2, 4, \ldots, 30$ (curves in ascending order) for a time series of 8192 points length sampled from the x-coordinate of the Lorenz system (same time series as in Fig. 1). Right panel: Nonlinearity test statistic δ versus subsampling factor S for model order $p = 20$; the figure shows average values (solid line) and 95%-confidence profiles (two standard deviations; dashed lines) for the results of 100 independent realisations of the x-coordinate of the Lorenz system, each consisting of 8192 points. For both panels only values at integer S correspond to actual numerical results.

to detect the full extent of the nonlinearity in this system. This is remarkable, since the original time series did not have an unreasonably small sampling step size. A delay time of 13 sampling steps corresponds quite well to the average distance between consecutive extrema (maximum and minimum or vice versa) of the Lorenz x-coordinate. For higher subsampling $\delta(p, S)$ decreases again and soon reaches zero, indicating that linear and nonlinear models perform equally well. This is to be expected for time series from chaotic systems, since the permanent loss of information about the system state due to a positive Lyapunov exponent will render the task of capturing the dynamics in a simple parametric model impossible for too sparsely sampled time series segments. As a consequence of the increasing complexity of the prediction task, both AR(p) and ExpAR($\frac{1}{2}p$) models treat the time series as a linear stochastic process, whence the nonlinear model can no longer outperform the linear model.

In the right panel of Fig. 2 we show the result of repeating this analysis for 100 different realisations of the Lorenz x-coordinate, each of 8192 points length. The model order was chosen as $p = 20$. The individual results were averaged for each S; the figure also shows the corresponding 95%-confidence profiles. It can be seen that the curve $\delta(S)$ is quite smooth for this system (for the moment ignoring the fact that it is defined only for integer S) and that the distribution of this statistic is not very broad. Clearly the width of this distribution depends on the length of the time series.

7. Comparison with surrogate time series

In the case of the noise-free onefold Lorenz system we expect clear evidence for nonlinearity, and indeed δ reaches almost unity at its maximum (for time series normalised to unit variance). When analysing real-world time series we have to face much less favourable conditions in terms of the quality of the data and the complexity of the dynamics, and even for undeniably nonlinear time series δ will be confined to values much closer to zero. But also for purely linear time series δ may attain positive values due to statistical fluctuations, therefore we will employ the

technique of surrogate data testing in order to test for significance of positive δ.

It should be stressed that the paradigm which led to the definition of AIC has no inherent connections with the concept of statistical hypothesis testing [17], although this is frequently not realised. Therefore it should be possible to provide confidence intervals for δ without resorting to the surrogate data approach. We intend to address this issue in future work.

There is another reason for the need to perform a surrogate data test in addition to estimating the new measure for nonlinearity: if the amplitude distribution of a purely linear time series was distorted by a monotonic nonlinear, but static transform, the test statistic δ may also attain weakly positive values. Although we intend to reduce the role of the surrogate data approach in nonlinearity testing, we do not yet have a convincing solution for this case, so we need surrogate data testing due to its ability to test against a more general null hypothesis.

For the same Lorenz x-coordinate time series as used in Sections 4 and 6 we generate 100 surrogate time series by the IAAPR algorithm [8]. Already the original time series had been chosen such that the endpoint mismatch of the time series itself and of its first derivative (i.e. the endpoint phase mismatch) are minimised, in order to avoid artefacts from poor spectrum estimation due to "wraparound artefact" effects [21]. We calculate the curves $\delta(S)$ for $p = 20$ for these surrogate time series; the results are shown, together with the corresponding result for the original time series, in the upper left panel of Fig. 3. It can be seen that the curves $\delta(S)$ stay close to zero for the surrogate time series, which is the expected behaviour, since surrogate time series are linear by construction. It is obvious that the null hypothesis can be clearly rejected even for $S > 13$.

In the definition of δ the same data is used for fitting the models and for evaluating and comparing their performance. In order to check whether this "in-sample" approach introduces any bias into the nonlinearity test, we have repeated the analysis using an "out-of-sample" variant of the proposed algorithm, such that each time series is divided into two non-overlapping time series segments, one for
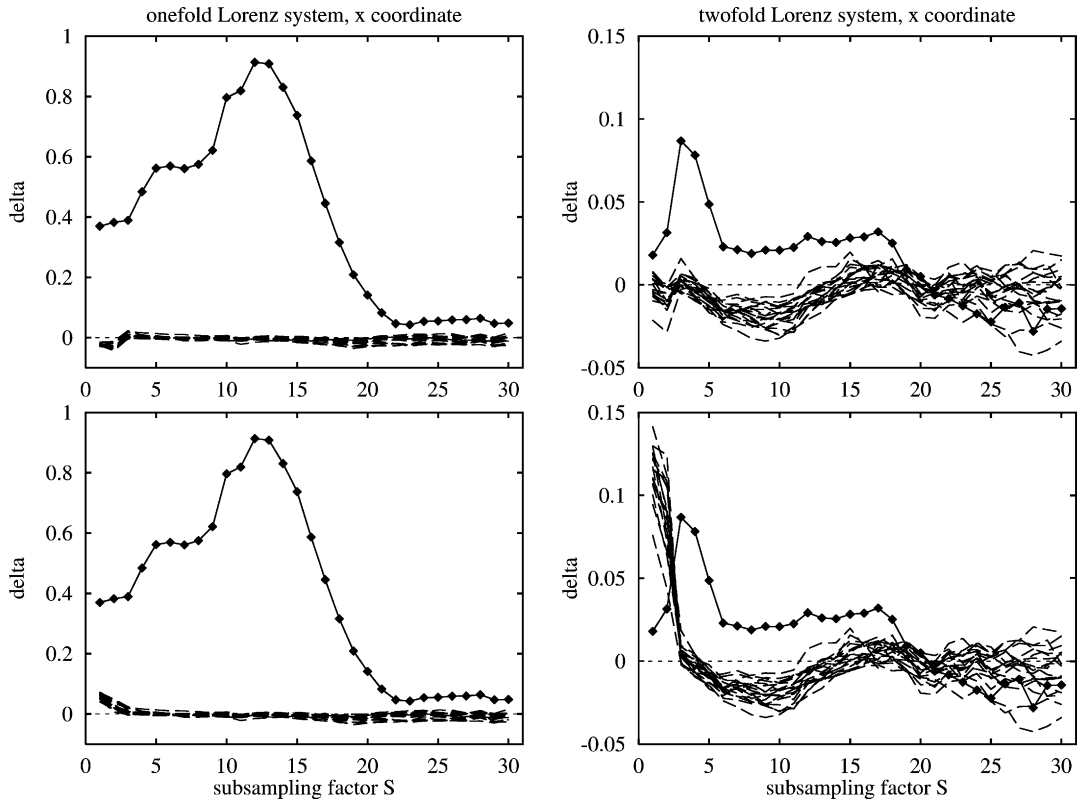
Fig. 3. Nonlinearity test statistic δ versus subsampling factor S for model order $p = 20$ for time series of 8192 points length sampled from the x-coordinate of the Lorenz system (right panels), and from the sum of the x-coordinates of the twofold Lorenz system (left panels). Note the different scale on the vertical axis of the left panels. All panels show results for the original time series (filled diamonds and solid lines) and for 100 surrogate time series (dashed lines). For the upper panels the surrogates were generated such that the power spectrum is exactly preserved, and for the lower panels the surrogates were generated such that the amplitude distribution is exactly preserved. Only values at integer S correspond to actual numerical results.

fitting the models and the other for calculating the test statistic $\delta(p, S)$. We found that the variance of the distribution of surrogate results is somewhat increased by this approach, but in all cases which were tested the conclusion about the validity of the null hypothesis remained the same. Therefore we conclude that an out-of-sample variant of the algorithm is unnecessary.

8. Application to time series with higher complexity

We now address a more difficult task by applying the nonlinearity test to time series of higher complexity as compared to the plain Lorenz x-coordinate. A

considerable increase in complexity results from considering two independent realisations of the Lorenz equations (*twofold Lorenz system*) and then adding up the two x-coordinates; thereby we obtain a scalar observable from a chaotic system which evolves in a six-dimensional state space and has two positive Lyapunov exponents. In a similar way Theiler and Prichard [9] analysed the sum of five independent realisations of the Hénon system (which was furthermore contaminated by large amounts of noise); again we are faced with a more difficult problem, since the twofold Lorenz system produces continuous time series.

One time series of $N = 8192$ points length is sampled from the sum of the x-coordinates of the twofold

Lorenz system, parameters of dynamics and of integration are the same as given in Section 4. Again the endpoint mismatch and endpoint phase mismatch are minimised as far as possible, and 100 surrogate time series are generated by the IAAPR algorithm. We calculate the curves $\delta(S)$ for $p = 20$ for the original time series and the surrogate time series; the results are shown in the upper right panel of Fig. 3. Even for the original time series the direct evidence of nonlinearity is much weaker now, we see a peak at $S = 3$ with $\delta \approx 0.09$. But still the discrimination between the original time series and the set of surrogate time series is obviously possible. Due to the presence of two non-interacting Lorenz systems the optimum sampling scale for the detection of nonlinearity is considerably reduced as compared to the onefold Lorenz system.

Now we would like to consider a seemingly insignificant technical detail of the surrogate time series generation algorithm: in Section 2 we mentioned that both power spectrum and amplitude distribution of the original time series are to be preserved in the surrogate time series. But on the basis of a given original time series of finite length we cannot have both demands fulfilled at the same time (unless we return *precisely* to the original time series); whereas commonly the amplitude distribution is kept preserved exactly (in fact, the surrogate time series are permutations of the original time series then) and minor inaccuracies of the power spectrum (more precisely, the periodogram) are tolerated, Kugiumtzis [5] has demonstrated recently that for certain test statistics such tiny inaccuracies of the spectrum may indeed cause erroneous results of nonlinearity testing. Following his suggestion we have stopped the iteration of our surrogate time series generation algorithm at a step, where the power spectrum is exactly preserved. Now if we stop the same iteration after the same number of steps at the exactly preserved amplitude distribution, we obtain the results shown in the lower panels of Fig. 3. It can be seen that for the twofold Lorenz system at low S the surrogates apparently display more nonlinearity than the original time series! We do not yet have an explanation for this artefact, but it is obviously removed quite well by following Kugiumtzis's suggestion. This result is even more remarkable, since to the human eye the actual

differences between a surrogate time series with exact power spectrum and the same surrogate time series with exact amplitude distribution appear to be tiny and negligible.

Furthermore we apply the nonlinearity test to a time series obtained from a high-precision hydrodynamical experiment, the Taylor–Couette system [22]. This time series contains scalar measurements of the axial velocity component (obtained by Laser–Doppler–Anemometry) at a particular site within a 12-vortex-flow displaying the axisymmetric "very-low-frequency" (VLF) mode [23] at a Reynolds numbers of $Re \approx 530$; at this value the dynamics is described by a strange attractor with a finite correlation dimension of at least $d_2 \approx 5.5$. The time series consists of 16 384 points, sampled at 2 Hz; it was chosen out of a longer time series such that endpoint mismatch and endpoint phase mismatch are minimal. Since the experiment was running for several days under well-controlled constant conditions it can be assumed that the time series is stationary.

In the left panel of Fig. 4 the curves $\delta(p, S)$ are shown for $S = 1, \ldots, 30$ and $p = 2, 4, \ldots, 50$; it can be seen that for this time series the convergence for increasing p occurs much slower. Indeed the AIC does not reach a minimum even for model orders of up to $p = 100$. But since beyond $p \approx 30$ only minor changes take place, we choose $p = 30$ as model order. For $S = 1$ there is no evidence of nonlinearity at all in this time series (the linear model is performing even slightly better than the nonlinear), but around $S = 7$ we find a clear peak of nonlinearity. Still the absolute value of $\delta \approx 0.08$ is fairly small, which is not surprising, considering the high complexity of this time series, which is revealed by the rather high value of the correlation dimension estimate.

Only by comparison with surrogate time series it becomes evident that the nonlinearity in this time series is significant (right panel of Fig. 4); again 100 surrogate time series have been employed. Consequently the further analysis of time series from this experiment by tools from nonlinear time series analysis is justified. It can also be seen that there is still some spurious "nonlinearity" in the surrogate time series at small S, although in this case we have preserved the
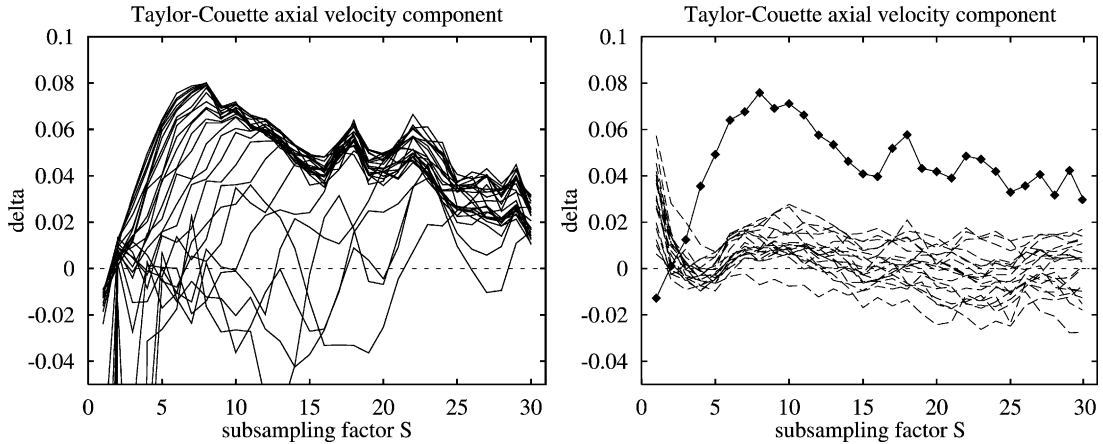
Fig. 4. Left panel: Nonlinearity test statistic δ versus subsampling factor S for model orders $p = 2, 4, \ldots, 50$ (curves roughly in ascending order) for a time series of $16\,384$ points length measured from a Taylor–Couette flow experiment (axial velocity component). Right panel: Nonlinearity test statistic δ versus subsampling factor S for model order $p = 30$ for the same time series as in the left panel. The panel shows results for the original time series (filled diamonds and solid line) and for 100 surrogate time series (with exactly preserved power spectrum; dashed lines). For both panels only values at integer S correspond to actual numerical results.

power spectrum exactly. Further work is needed in order to understand this effect.

9. Comparison with other test statistics

Finally we shall compare the performance of our quantitative measure of nonlinearity under conventional surrogate data testing with two other test statistics. For this investigation we fix the model order p to 20.

- According to results of Schreiber and Schmitz [11] the r.m.s. prediction error of a non-parametric locally constant predictor in time-delay reconstructed state space yields consistently good discrimination power. The algorithm is described in detail in [24]. This statistic bears some similarity to our measure of nonlinearity δ, which is based on parametric modelling; however, it is currently not possible to apply the concept of AIC to non-parametric models. On the other hand, the statistic has the advantage of operating directly on trajectories of continuous systems in state space. For the practical application of a locally constant predictor various parameters have to be chosen; time delay, embedding dimension and

neighbourhood size. We have adjusted these parameters in the way recommended by Schreiber and Schmitz [11], i.e. by fixing them through a preparatory investigation such that the predictive performance is optimised.

- Furthermore we employ the time irreversibility test proposed by Diks et al. [25] as test statistic. Note that the absence of reversibility of a time series under change of the direction of time is a "property" *closely related, but not equivalent* to nonlinearity. Nevertheless such a measure can be expected to yield good discrimination power under surrogate data testing, although so far this has not been tested. This approach is also based on a time-delay reconstructed state space and evaluates a fairly sophisticated test statistic on the set of all pairs of reconstructed vectors, just as it is also the case in the well-known Grassberger–Procaccia algorithm for correlation dimension estimation [26]. Parameters of the time-delay embedding and a bandwidth parameter of the test statistic are chosen such that a maximum value of this irreversibility test statistic is obtained. In order to reduce the effects of dynamical correlations among the reconstructed vectors and also to reduce the computational time expense, the test is evaluated only on a randomly

chosen subset of 10^6 pairs of vectors (for time series of $N = 8192$ points length, corresponding to a total of 33.55×10^6 pairs).

For this investigation we choose the twofold and threefold Lorenz systems, i.e. the sum of two or three x-coordinates of independent Lorenz systems, parameters are the same as in previous sections. In order to make the discrimination task more demanding, observational noise is added to the pure time series. The relative noise level ranges from 0 (no noise) to 1 (equal amounts of signal and noise); for the definition of the relative noise level and for the appropriate normalisation of the resulting noisy time series we follow Eq. (8) of Schreiber and Schmitz [11]. The noise is chosen either as white gaussian or as in-band noise, i.e. having the same power spectrum

(or, more precisely, periodogram) as the original data; in the latter case the noise realisation is in fact simply taken from a surrogate data set of a realisation of the noise-free process. Time series of $N = 8192$ points length (which is a rather small data set size, considering the high dimension of the twofold and threefold Lorenz systems) are selected from longer realisations of the processes, such that endpoint mismatch and endpoint phase mismatch are minimised; for obvious reasons, this is done for the case of white noise contribution prior to adding the noise realisation, and for in-band noise *after* adding a longer realisation of the noise. By this method 100 time series are created for each case; for each of these time series 19 surrogates are computed by IAAPR; consequently the surrogate data test has a size of 5%. The results are shown in Fig. 5.
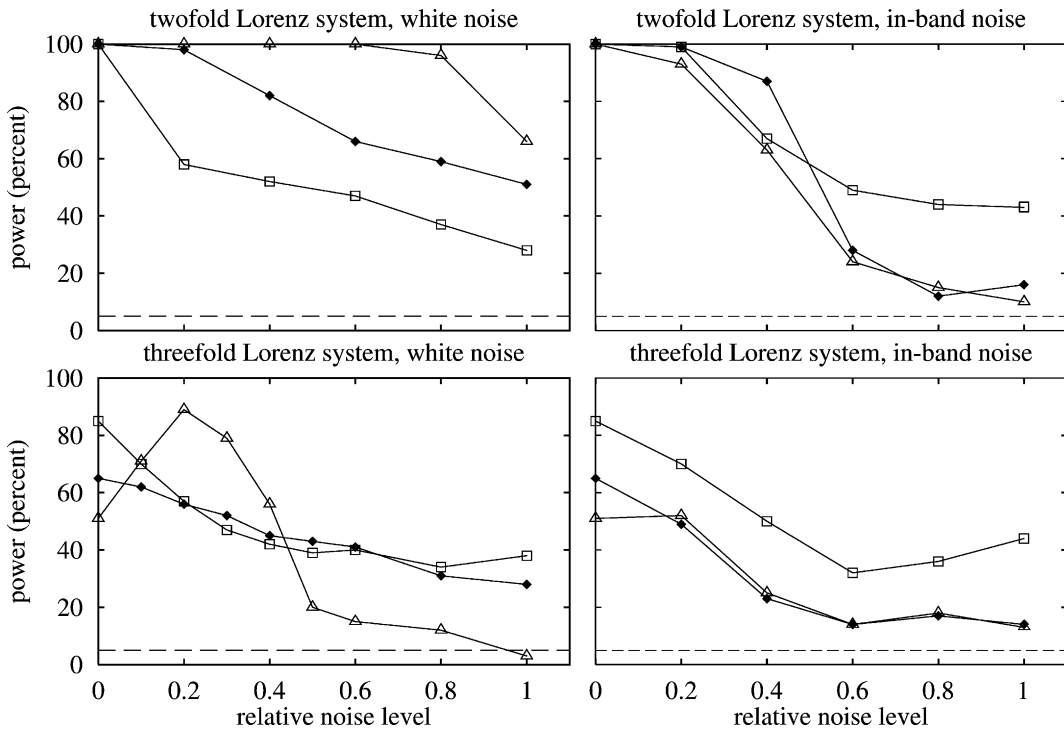

Fig. 5. Discrimination power (i.e. fraction of correct rejections of the null hypothesis) under surrogate data testing versus relative noise level, estimated from 100 time series (for each case and noise level) of 8192 points length sampled from the sum of the x-coordinates of the twofold Lorenz system (upper panels) and the threefold Lorenz system (lower panels), contaminated by white noise (left panels) and in-band noise (right panels). Results are shown for the nonlinearity test statistic δ as defined in Eq. (8) (filled diamonds), for the r.m.s. prediction error of local constant prediction (triangles) and for the time irreversibility statistic defined in Ref. [24] (squares). The dashed line at 5% denotes the size of the test, i.e. the level of power corresponding to a erroneous rejection of the null hypothesis.

From the figure we see that there is not a unique winner among the three test statistics, but rather, depending on the dimension of the underlying dynamical system and the amount and spectral properties of the noise, different test statistics provide the highest power of discrimination. First of all we note that for the noise-free twofold Lorenz system all three test statistics achieve perfect discrimination, i.e. a power of 100%. We would like to mention that we were unable to reproduce this success with test statistics based on the bispectrum [6]. For the threefold Lorenz system it is no longer possible to achieve such perfect discrimination, rather we find highest power (85%) for the irreversibility statistic (henceforth abbreviated by "IRR"), and lowest (51%) for the local constant predictor ("LCP"). The δ statistic ("δ") achieves 65%.

When adding white noise, this picture changes considerably: for the twofold Lorenz IRR soon looses much of its power, whereas δ performs substantially better, and LCP is almost unaffected even by a relative noise level of up to 0.8. Considering that local constant prediction provides a very efficient noise reduction algorithm [24], this result should not be too surprising. For the threefold Lorenz the power of LCP even increases with increasing noise level, a result which seems to indicate the presence of "noise-induced detection of nonlinearity", an effect which we are currently unable to explain. The result can easily be reproduced by employing a set of 100 different time series. For higher noise levels, however, the power of LCP soon drops down towards the size of the test (5%), and IRR and δ provide much better power, without significant differences of power between these two.

When adding in-band noise, again a completely different result is obtained: LCP cannot cope well with this kind of noise and frequently yields the lowest power of the three statistics. δ yields only slightly better result than LCP, although for the twofold Lorenz system the decrease of power sets in at higher noise levels. For this discrimination task IRR seems to be the best choice, especially at higher dimension and noise level.

These somewhat inconsistent results illustrate the difficulty of proper choice of test statistics in surrogate data testing and also the need for further work

on designing optimised statistics. However, it should be stressed that our main motivation for designing the δ statistic was to devise a quantitative and immediately meaningful measure for nonlinearity and to reduce the role of the surrogate data approach. It seems that the price for an improved interpretability of the test statistic (with respect to explicit modelling) is a certain loss of power, since we have found situations for which either LCP or IRR provide higher power of discrimination. On the other hand, IRR is a measure of time irreversibility, and this property (or more precisely, *absence* of a property) is not equivalent to nonlinearity, so a rejection of the null hypothesis provides less information on the potential nonlinear characteristics of a time series than the δ statistic which is based directly on a specific class of nonlinear models.

10. Discussion and conclusion

In this paper we have designed a new measure for nonlinearity in time series from continuous dynamical systems; to this end we have combined discrete parametric modelling with a subsampling approach which is analogous to time-delay embedding with delay times larger than unity. It is a central result of this paper that the well-known problems of modelling continuous dynamical systems by discrete parametric models can be addressed successfully by a subsampling approach. As a by-product the subsampling approach also provides an estimate of the optimum delay time with respect to the detection of nonlinearity; such information may prove to be useful for designing refined models.

Our motivation came from the wish to replace conventional surrogate data testing by a more quantitative test for nonlinearity, in order to obtain a test which provides more specific information on the time series and reduces the detrimental effects of various well-known problems and weak points of surrogate data testing. Conventional surrogate data testing essentially yields just a "yes/no" information, which in a world where weak nonlinearities are probably ubiquitous [27] may not provide very deep insights. By comparing the predictive performance of linear and ExpAR models through their values of AIC we have

obtained a more quantitative measure of the nonlinear characteristics of a time series.

Since these models cannot yet cope with nonlinear static transforms, it is still recommended to combine this test statistic with surrogate data testing. We intend to remove this shortcoming in future work. Surrogate data testing is also helpful in assessing the significance of a deviation from zero of our measure.

We have demonstrated that certain technical details of the surrogate time series generation algorithm may have severe effects on testing for nonlinearity; but due to the subsampling approach our test is rather robust against these effects (in the lower panels of Fig. 3 the distinction between the original time series and the surrogate time series is still clearly possible). Nevertheless these effects, which seem to be closely related to the issue of preferring "constrained-randomisation" over "typical-realisation" algorithms for surrogate time series generation [9], deserve further investigation.

By comparing the discrimination power of our test statistic with two other well-tried test statistics under conventional surrogate data testing, when applied to high-dimensional time series contaminated by strong noise components, we have shown that it strongly depends on the dimension of the underlying dynamical system and the amount and spectral properties of the noise, which statistic provides the highest power. It seems that the price for an improved interpretability of our test statistic is a certain loss of power, but at least this statistic can essentially keep up with the other statistics which were tested. We presume that the performance of our test can be improved further by employing a more general class of nonlinear models. The main drawback of such models will be the need to perform explicitly nonlinear fitting of parameters, which will considerably increase the computational time expenses.

Finally it has to be admitted that the reliable detection and quantitative description of nonlinearity in time series of limited length and limited precision remains a difficult task, especially if the dynamics is characterised by higher complexity. Failure to detect nonlinearity will always have two possible reasons: either the dynamics is actually linear, or there is non-

linearity which cannot reliably be detected on the basis of the available data. But in the latter case the time series would have to be regarded as "operationally linear", and there would be no justification to apply nonlinear tools of analysis. By the example of the experimental time series from Taylor–Couette flow we have seen an example of significant nonlinearity despite of a rather high dimension; in contrast to this result we have gained the impression that for many time series from human electroencephalogram which we have tested it is not possible to find significant nonlinearity.

Acknowledgements

References

[1] M.C. Casdagli, S. Eubank, J.D. Farmer, J. Gibson, State space reconstruction in the presence of noise, Physica D 51 (1991) 52–98.

[2] T. Schreiber, A. Schmitz, Surrogate time series, Physica D 142 (2000) 346–382.

[3] J. Timmer, The power of surrogate data testing with respect to non-stationarity, Phys. Rev. E 58 (1998) 5153–5156.

[4] P.E. Rapp, A.M. Albano, I.D. Zimmerman, M.A. Jiménez-Montaño, Phase-randomized surrogates can produce spurious identifications of non-random structure, Phys. Lett. A 192 (1994) 27–33.

[5] D. Kugiumtzis, Test your surrogate data before you test for nonlinearity, Phys. Rev. E 60 (1999) 2808–2816.

[6] M.J. Hinich, Testing for gaussianity and linearity of a stationary time series, J. Time Ser. Anal. 3 (1982) 169–176.

[7] A. Galka, Topics in Nonlinear Time Series Analysis — With Implications for EEG Analysis, Advanced Series in Nonlinear Dynamics, Vol. 14, World Scientific, Singapore, 2000.

[8] T. Schreiber, A. Schmitz, Improved surrogate data for nonlinearity tests, Phys. Rev. Lett. 77 (1996) 635–638.

[9] J. Theiler, D. Prichard, Constrained-realization Monte Carlo method for hypothesis testing, Physica D 94 (1996) 221–240.

[10] J. Theiler, S. Eubank, Don't bleach chaotic data, CHAOS 3 (1993) 771–782.

[11] T. Schreiber, A. Schmitz, Discrimination power of measures for nonlinearity in a time series, Phys. Rev. E 55 (1997) 5443–5447.

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

[13] M. Barahona, C.-S. Poon, Detection of nonlinear dynamics in short noisy time series, Nature 381 (1996) 215–217.

[14] T. Ozaki, H. Oda, Nonlinear time series model identification by Akaike's information criterion, in: B. Dubuisson (Ed.), Information and Systems, Pergamon Press, Oxford, 1978, pp. 83–91.

[15] Zh. Shi, Y. Tamura, T. Ozaki, Nonlinear time series modelling with the radial basis function-based state-dependent autoregressive model, Int. J. Syst. Sci. 30 (1999) 717–728.

[16] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (1974) 716–723.

[17] Y. Sakamoto, M. Ishiguro, G. Kitagawa, Akaike Information Criterion Statistics, Reidel, Dordrecht, 1986.

[18] A. Buse, The likelihood ratio, Wald, and Lagrange multiplier test: an expository note, The American Statistician 36 (1982) 153–157.

[19] R. Schlittgen, B.H.J. Streitberg, Zeitreihenanalyse, 8th Edition, Oldenburg, Munich, 1999.

[20] L.A. Aguirre, S.A. Billings, Retrieving dynamical invariants from chaotic data using NARMAX models, Int. J. Bifurc. Chaos 5 (1995) 449–474.

[21] J. Theiler, P.E. Rapp, Re-examination of the evidence for low-dimensional nonlinear structure in the human electroencephalogram, Electroen. Clin. Neurophys. 98 (1996) 213–222.

[22] C. Egbers, G. Pfister, Physics of Rotating Fluids, Lecture Notes in Physics, Vol. 549, Springer, Berlin, 2000.

[23] J. von Stamm, U. Gerdts, T. Buzug, G. Pfister, Symmetry breaking and period doubling on a torus in the VLF regime in Taylor–Couette flow, Phys. Rev. E 54 (1996) 4938–4957.

[24] R. Hegger, H. Kantz, T. Schreiber, Practical implementation of nonlinear time series methods: the TISEAN package, CHAOS 9 (1999) 413–435.

[25] C. Diks, J.C. van Houwelingen, F. Takens, J. DeGoede, Reversibility as a criterion for discriminating time series, Phys. Lett. A 201 (1995) 221–228.

[26] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D 9 (1983) 189–208.

[27] J. Timmer, What can be inferred from surrogate data testing? Phys. Rev. Lett. 85 (2000) 2647.