2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Surrogate generation algorithm for pseudoperiodic time series

Xiaodong Luo, Tomomichi Nakamura and Michael Small
EIE Department, Hong Kong Polytechnic University, Hung Hom, Hong Kong.
Email: enxdluo@eie.polyu.edu.hk, entomo@eie.polyu.edu.hk, ensmall@eie.polyu.edu.hk

Abstract—In this paper we propose an effective surrogate generation algorithm for pseudoperiodic time series, which can properly cope with a large scope of stochastic perturbations in the noisy data sets. As an example of application, we will demonstrate the ability of this algorithm to distinguish chaotic time series from pseudoperiodic ones of the Rössler system. In addition, we will briefly introduce another surrogate generation algorithm for nonlinearity detection whose central idea is similar with that in this algorithm.

1. INTRODUCTION

The main idea of surrogate tests [1] is that, by presuming a time series possesses a property of interest, we devise an algorithm to produce surrogates of the original time series which aims to preserve the potential property but destroy all others. The generated surrogates, together with the original time series, provide necessary stuff for a Monte Carlo hypothesis test on the property we are interested in. Hence the advantage of surrogate tests is that, even though we only have a limited amount of data, we can still assess the confidence interval of our calculations. Such assessments are important because in many situations the presence of noise will impair the statistical reliability of our calculation results.

Here we consider an algorithm to generate surrogates for pseudoperiodic time series. By pseudoperiodic time series we mean the representative of a periodic orbit perturbed by dynamical or observational noise, whose states within one cycle are largely independent of those within previous cycles given a cycle length. Initially, Theiler [2] proposed the cycle shuffling algorithm to generate pseudoperiodic surrogates. But the difficulty in applying this algorithm is that, we have to know the precise periodicity of the data set. But due to the quantization error when measuring the data, along with other noise sources, it is inevitable that we will introduce certain discontinuity into the surrogates by randomly shuffling individual cycles of the time series. Sometimes, such discontinuity might lead to spurious results if not dealt with cautiously [4].

Later, Small *et al.* [4] proposed the pseudoperiodic surrogate (PPS) algorithm. First they conduct the time delay embedding reconstruction on the time series to produce a vector field, among which the underlying system of the original time series is embedded. Then they utilize the local linear modelling techniques to generate surrogates, which will approximate the original time series. As the authors reported, this method works well for pseudoperiodic time series even with a very high dynamical noise level.

In this paper we propose another slightly simpler algorithm to generate surrogates for pseudoperiodic time series. First we elucidate our null hypothesis which specifies the property of interest, then in the next section we devise the corresponding algorithm to generate the surrogates. The null hypothesis is that, the stationary time series is pseudoperiodic with noise components which are (approximately) identically distributed and uncorrelated for sufficiently large temporal translations. By noise components we mean any perturbations to the underlying deterministic system of the time series, hence they can be dynamical noise, observational noise or their combinations. However, we shall note that, if an unstable periodic orbit (UPO) is perturbed by a large amount of dynamical noise, it is possible that the observed orbit will not stay close to the original one, and the pseudoperiodicity of the time series might be broken. In addition, the constraints of the noise components are stronger than that of Theiler's algorithm, which only implicitly requires the noise components will not change after cycles shuffling. Nevertheless, the situations described in our null hypothesis still covers a large scope, including the case of UPOs perturbed by small enough dynamical noise together with observational noise, which will be shown later.

We will discuss the surrogate generation algorithm in the next section. As an example of application of this algorithm, we will demonstrate its ability to distinguish chaotic time series from pseudoperiodic ones of the Rössler system. In addition, we will briefly introduce another surrogate generation algorithm for nonlinearity detection proposed in our another recent works, which is somewhat similar with the one to be introduced below.

2. SURROGATE ALGORITHM FOR PSEUDOPERIODIC TIME SERIES

Let $\{x_i\}_{i=1}^N$ be a data set with N observations. We assume $\{x_i\}_{i=1}^N$ is stationary and can be (approximately) decomposed into the deterministic components and the noise components, which are independent of each other. Therefore we can write a data point x_i as $x_i = p_i + n_i$, where p_i and n_i denote the periodic component and the noise

component respectively. Without losing generality, we set $E(p_i) = E(n_i) = 0$ where E is the expectation operator. To generate the surrogates, let

$$y_i^\tau = \alpha x_i + \beta x_{i+\tau} = (\alpha p_i + \beta p_{i+\tau}) + (\alpha n_i + \beta n_{i+\tau}) \quad (1)$$

with $i = 1, 2, ..., N - \tau$, where coefficients α and β satisfy $\alpha^2 + \beta^2 = 1$ and parameter τ is the temporal translation between subsets $\{x_i\}_{i=1}^{N-\tau}$ and $\{x_{i+\tau}\}_{i=1}^{N-\tau}$. If τ is sufficiently large, under our hypothesis, n_i and $n_{i+\tau}$ are uncorrelated, then $var(\alpha n_i + \beta n_{i+\tau}) = var(n_i)$. If we also require the translation τ to satisfy that the autocovariance function $cov(p_i, p_{i+\tau}) = 0$, then $var(\alpha p_i + \beta p_{i+\tau}) = var(p_i)$. Hence the surrogates $\{y_i^\tau\}_{i=1}^{N-\tau}$ and the original time series $\{x_i\}_{i=1}^{N}$ will have the same noise level $(var(n_i)/var(x_i))^{1/2}$.

In addition, if the deterministic components $\{p_i\}_{i=1}^{N}$ are from a periodic orbit, then the summation $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$ of $\{\alpha p_i\}_{i=1}^{N-\tau}$ and $\{\beta p_{i+\tau}\}_{i=1}^{N-\tau}$ will also form a periodic orbit except when $\alpha p_i + \beta p_{i+\tau} \approx 0$ holds nearly for all indices i, as to be explained below. Note $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$ and $\{p_i\}_{i=1}^{N}$ might be quite different periodic orbits, e.g., $\{p_i\}_{i=1}^{N}$ is a period 4 orbit while $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$ could turn to be a period 2 one. Nevertheless, the correlation dimensions of both the periodic orbits will theoretically be the same since $\{p_i\}_{i=1}^{N}$ and $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$ have the same degree-of-freedom. Hence we can adopt the correlation dimension as the suitable discriminating statistic in this situation.

We now consider several computational issues:

1. We require the translation τ to satisfy the condition $cov(p_i, p_{i+\tau}) = 0$. The reason is that we want the deterministic components $\{p_i\}_{i=1}^{N-\tau}$ to be orthogonal to $\{p_{i+\tau}\}_{i=1}^{N-\tau}$, otherwise the projection of $\{\alpha p_i\}_{i=1}^{N-\tau}$ onto $\{\beta p_{i+\tau}\}_{i=1}^{N-\tau}$ might counteract $\{\beta p_{i+\tau}\}_{i=1}^{N-\tau}$ under some situations, for example, if $p_i \approx -p_{i+\tau}$ and $\alpha = \beta$, the deterministic components $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$ will almost vanish while the noise components $\{\alpha n_i + \beta n_{i+\tau}\}_{i=1}^{N-\tau}$ remain. Hence the correlation dimensions calculated are actually those of the noise components instead of the deterministic components, which will certainly lead to the false rejection of the null hypothesis.

2. We attempt to preserve the noise level during the generation process of the surrogates, which requires τ to be sufficiently large to guarantee the decorrelation between the noise components. However, we expect $\{x_i\}_{i=1}^{N-\tau}$ and $\{x_{i+\tau}\}_{i=1}^{N-\tau}$ shall have at least some overlaps to make use of the information of the whole data set $\{x_i\}_{i=1}^{N}$, which means τ shall not exceed $N/2$. Moreover, it is recommended that the length of a data set shall not be too short in order to appropriately calculate its correlation dimension, which also implies τ shall not be too large.

3. From Eqn. (1) we note that the coefficient ratio α/β shall not be too large or too small, otherwise $\{y_i^\tau\}_{i=1}^{N-\tau}$ will be very close to $\{x_i\}_{i=1}^{N-\tau}$ or $\{x_{i+\tau}\}_{i=1}^{N-\tau}$, which will lead to approximately the same correlation dimensions of $\{x_i\}_{i=1}^{N}$ and $\{y_i^\tau\}_{i=1}^{N-\tau}$ regardless of the dynamical behavior of $\{x_i\}_{i=1}^{N}$, and thus impair the discriminating power of the correlation dimension. We suggest to let α be uniformly drawn from the interval $[-0.8, -0.6] \cup [0.6, 0.8]$ and $\beta = \sqrt{1 - \alpha^2}$, which provides the ratio α/β moderate values.

4. $\{x_i\}_{i=1}^{N}$ and $\{y_i^\tau\}_{i=1}^{N-\tau}$ have the same noise level, but the distribution of their noise components, $\{n_i\}_{i=1}^{N}$ and $\{\alpha n_i + \beta n_{i+\tau}\}_{i=1}^{N-\tau}$, might be different. We choose the Gaussian kernel algorithm (GKA) [5] to calculate the correlation dimensions as it is reported that the GKA can reasonably estimate the correlation dimensions of noisy data sets with different noise distributions.

3. APPLICATION

In this section we apply this surrogate generation algorithm to the Rössler system to demonstrate its ability of discrimination between chaotic and pseudoperiodic time series. The central idea is that, if $\{p_i\}_{i=1}^{N}$ is periodic, its linear combination, $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$, shall also be periodic. However, if the deterministic part $\{p_i\}_{i=1}^{N}$ is chaotic, then the summation $\{\alpha p_i + \beta p_{i+\tau}\}_{i=1}^{N-\tau}$ may have a new dynamical structure with a different correlation dimension from that of $\{p_i\}_{i=1}^{N}$ due to the sensitivity of chaotic systems to initial conditions, hence by adopting the correlation dimension as the discriminating statistic we might detect this difference.

The equations of the Rössler system are given by

$$\begin{cases} \dot{x} = -y - z, \\ \dot{y} = x + ay, \\ \dot{z} = b + z(x - c). \end{cases} \quad (2)$$

with the initial conditions $x(0) = y(0) = z(0) = 0.1$. We fix parameters $b = 2$, $c = 4$ and choose the integration time step $= 0.1$ time units. We integrate the system $10,000$ times and take the x components as the time series to be studied (we discard the first $1,000$ data points to avoid the possible transient states for safety). When parameter $a = 0.39095$, the Rössler system exhibits limit cycle behavior of period 6. To obtain the pseudoperiodic orbit, we first add 0.15% Gaussian white noise (w.r.t the standard deviation) to the x component at each integration step, and then introduce observational noise into the obtained data set. Although Gaussian white observational noise is the most common choice in this situation, in order to demonstrate the ability of our surrogate algorithm to deal with colored noise, we instead adopt the noise generated from the $AR(1)$ process $\xi_{i+1} = 0.8\xi_i + \epsilon_i$ with the variable ϵ

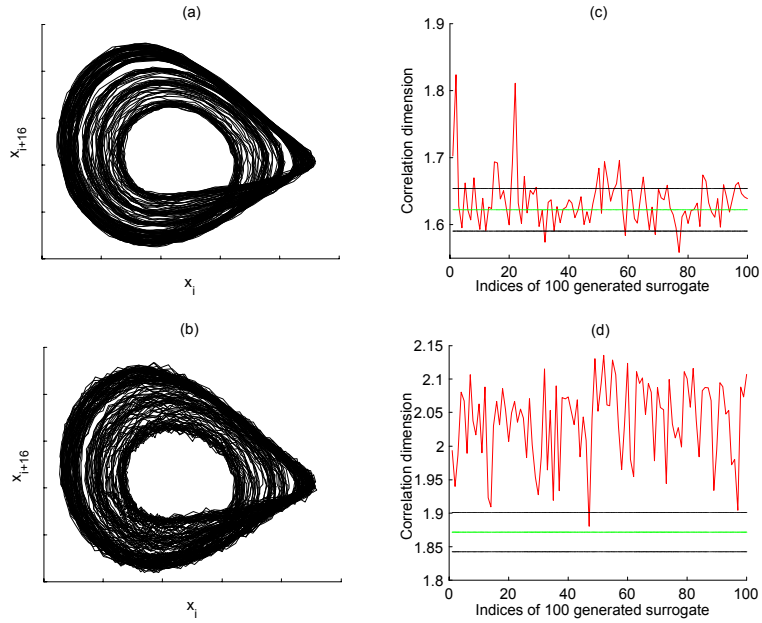

Figure 1: (a) State space x_{i+n} vs. x_i of the pseudoperiodic orbit from the Rössler system with $n = 16$; (b) State space x_{i+n} vs. x_i of the chaotic orbit from the Rössler system with $n = 16$; (c) Surrogate test for the pseudoperiodic orbit. The abscissa is the indices of 100 surrogates and the ordinate is the corresponding correlation dimensions. The middle line is the mean correlation dimension of the original time series calculated 100 times using the GKA with the embedding dimension $m = 10$, the upper and lower lines denote the correlation dimensions twice the standard deviation away from the mean value and the curve indicates the correlation dimensions of 100 surrogates. (d) Surrogate test for the chaotic orbit. The meaning of the lines and the curve is the same as that in Panel (c).

following the normal Gaussian distribution $N(0, 1)$ and set the standard deviation of the $AR(1)$ noise to be 3% of that of the obtained data set. However, one shall note that other colored noise modelled by the $ARMA(p, q)$ processes in principle shall also lead to the same results. When parameter $a = 0.392$, the system exhibits chaotic behavior. To obtain the chaotic orbit, we first integrate Eqn. (2) without introducing any dynamical perturbations, and then add the obtained data set with 3% $AR(1)$ observational noise. The phase spaces x_{i+n} vs. x_i of both the pseudoperiodic and chaotic orbits with $n = 16$ are plotted in panel (a) and (b) of Fig. 1 respectively. Note that, although it is easy to visually distinguish between the pseudoperiodic and chaotic orbits, without any *a priori* knowledge, it will be more difficult to appropriately infer the underlying dynamical behaviors from only visual inspections, hence it is nontrivial to devise an algorithm to analyze quantitatively.

To choose the suitable temporal translations for surrogate generation, we select a interval of $[100, 150]$. Due to the quantization error in generating the original time series, there might be no integer translations within this interval which exactly satisfy the condition $cov(x_i, x_{i+\tau}) = 0$, hence we have to instead search the local minimum integer translations which make the measure $|cov(x_i, x_{i+\tau})|$ most close to zero. We adopt the integers thus obtained as the temporal translations with the same probability to generate 100 surrogates, and then utilize the GKA implemented in [6] to calculate their correlation dimensions. Before calculating the correlation dimensions, we need to reconstruct the data sets via time delay embeddings. For this purpose, we select the time delay according to the algorithm proposed in [7] and let the embedding dimension vary from 2 to 20, among which the calculation results of $m = 10$ are indicated, see Fig. 1. According to Taken's embedding theorem, other embedding dimensions will lead to the same conclusions (to be shown below) if they are large enough. Note that to speed up the calculation, only 2000 data points are used as the reference points for the GKA. There are some statistical fluctuations even for the same data set when calculating its correlation dimension, therefore we calculate 100 times to estimate its mean correlation dimension and standard deviation. As shown in panel (c) and (d) of Fig. 1, there are three lines parallel to the abscissas in both panels. The middle lines denote the estimation of the mean correlation dimensions of the pseudoperiodic and chaotic data sets, while the upper and lower lines indicate the positions twice the standard deviation away from the mean values. For the surrogates of both the pseudoperiodic and chaotic data sets, however, we calculate their correlation dimensions only once to save time. The results are illustrated as the curves in panel (c) and (d) respectively.

We use the ranking criterion [3] to determine whether

the null hypothesis shall be rejected or not. The idea of this criterion is that, suppose the discriminating statistic of the original data set is Q_0, and those of N_S surrogates are $\{Q_1, Q_2, ..., Q_{N_S}\}$. Rank the statistics $\{Q_0, Q_1, ..., Q_{N_S}\}$ and denote the rank of Q_0 by r_0, if the data set is consistent with the hypothesis (i.e., no evidence to reject), r_0 can be any integer value between 1 and $N_S + 1$. However, if the hypothesis is false, Q_0 might be distracted from the realization $\{Q_1, Q_2, ..., Q_{N_S}\}$ of the surrogate distribution, i.e, Q_0 will be the smallest or largest amongst $\{Q_0, Q_1, ..., Q_{N_S}\}$, hence we can reject the null hypothesis if $r_0 = 1$ or N_S+1, the probability of a false rejection is $1/(N_S + 1)$ for one-sided tests and $2/(N_S + 1)$ for two-sided tests. In our calculations with 100 surrogates, for the pseudoperiodic data set, we cannot distinguish the correlation dimension of the original data set from those of the surrogates, as shown in panel (c) of Fig. 1, which means we cannot reject the null hypothesis. However, for the chaotic data set, the correlation dimension of the original data set is obviously different from those of the surrogates shown in panel (d), hence we can reject the null hypothesis with a confidence level up to 98% even for a two-sided test. By this way we can appropriately distinguish chaotic orbits from pseudoperiodic orbits.

4. FURTHER DISCUSSION

Theiler *et al.* [1] proposed three hierarchical algorithms for nonlinearity detection, i.e., the Fourier transform (FT) algorithm, the amplitude adjusted Fourier transform algorithm (AAFT) and iterative AAFT algorithm (IAAFT). However, the relatively strong constraints in their hypothesis for the IAAFT algorithm narrow its application scope in some situations. For example, for a chaotic neuron network with several neurons, its output is the average of all of the outputs of each neuron. Via Theiler's three algorithms, we cannot find the evidence of nonlinearity, i.e., we have to conclude that the output of the chaotic neuron network most likely comes from a linear stochastic system. This is obviously not true.

In one of our recent works [8], we are investigating another surrogate algorithm to detect the nonlinearity in the time series from the chaotic neuron network. The central idea is similar with that to distinguish pseudoperiodic orbits from chaotic ones as described previously, i.e., if the underlying system of a time series is linear, there shall be no fractal structure. By choosing suitable temporal translations, the surrogates generated by summing up several subsets of the original time series might not change the linear property of the original system. However, if the underlying system is chaotic, summing up several subsets will very likely destroy its original fractal structure. Since the correlation dimension is not an appropriate statistic for stochastic systems, we adopt the inter-point distribution as the discriminating statistic instead to detect the possible nonlinearity.

5. CONCLUSION

In this paper a simple but effective algorithm is proposed to generate surrogates for pseudoperiodic time series. The main idea of this algorithm is that a linear combination of any two segments of the same periodic orbit will generate another periodic orbit, by properly choosing the temporal translation, the surrogates thus generated will have the same noise level as that of the original time series. As an application of this algorithm, we can use it to distinguish chaotic time series from pseudoperiodic ones since a linear combination of two segments of a chaotic orbit is very possible to destroy its dynamical structure. The idea is not limited to pseudoperiodic time series, we can also apply to detect nonlinearity in stochastic-like time series. This is under investigation now.

References

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

[2] J. Theiler, "On the evidence for low-dimensional chaos in an epileptic electroencephalogram ", *Phys. Lett. A*, Volume 196, pp.335-341, 1995.

[3] 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 Inst. Commun.*, Volume 11, pp.99-113, 1997.

[4] M. Small, D. Yu, and R.G. Harrison, "A surrogate test for pseudo-periodic time series data", *Phys. Rev. Lett.*, Volume 87, 188101, 2001.

[5] C. Diks, "Estimating invariants of noisy attractors", *Phys. Rev. E*, Volume 53, pp.4263-4266, 1996.

[6] D.J. Yu, M. Small, R.G. Harrison, and C. Diks, "Efficient implementation of the Gaussian kernel algorithm in estimating invariants and noise level from noisy time series data", *Phys. Rev. E*, volume 61, pp.3750-3756, 2000.

[7] X. Luo and M. Small, "Using geometric measures of redundancy and irrelevance tradeoff coefficient to choose suitable delay times for continuous systems", submitted, 2003.

[8] T. Nakamura, X. Luo and M. Small, "Topological surrogate data technique for chaotic time series", in preparation, 2004.