Eur. Phys. J. B **13**, 399–403 (2000)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Revisiting the role of correlation coefficient to distinguish chaos from noise

J. Bhattacharya[1,2,a] and P.P. Kanjilal[2]

[1] Max-Planck-Institut für Physik Komplexer Systeme, Nöthnitzer Strasse 38, Dresden 01187, Germany
[2] Department of Electronics & ECE, Indian Institute of Technology, Kharagpur 721302, India

Received 8 March 1999

Abstract. The correlation coefficient *vs.* prediction time profile has been widely used to distinguish chaos from noise. The correlation coefficient remains initially high, gradually decreasing as prediction time increases for chaos and remains low for all prediction time for noise. We here show that for some chaotic series with dominant embedded cyclical component(s), when modelled through a newly developed scheme of periodic decomposition, will yield high correlation coefficient even for long prediction time intervals, thus leading to a wrong assessment of inherent chaoticity. But if this profile of correlation coefficient *vs.* prediction horizon is compared with the profile obtained from the surrogate series, correct interpretations about the underlying dynamics are very much likely.

PACS. 05.45.Ac Low-dimensional chaos – 05.45.Tp Time series analysis

1 Introduction

Low-dimensional chaos but seemingly random behaviour is an ubiquitous phenomenon in nature. For sparse time series (those containing data points in the order of ∼1000–2000), one is simply interested to detect the presence of low-dimensional chaos. In the past, the identification of chaotic dynamics in time series relied heavily on the estimation of the dimension of the underlying attractor [1–3]. But this approach [3] has certain limitations [4]. Other approaches [5,6] to detect the underlying chaos exploit the intrinsic deterministic properties of the chaotic system. Since chaotic systems obey certain definite rules, limited predictions are possible although long term predictions are not possible due to the extreme sensitivity of it's initial conditions. Sugihara and May [7] proposed a modified version of the prediction algorithm originally proposed by Lorenz [8], which are based upon a library of past patterns in a data series. The predicted series is statistically compared to the original series and corresponding Pearson's linear correlation coefficient (ρ) *vs.* prediction time (T_p) profile is used to differentiate between chaos and additive noise. In chaotic processes, ρ is shown to fall gradually starting from a high value whereas in random processes it remains steady at much lower value. Numerous research reports have been published using this feature of ρ *vs.* T_p profile for analysing the underlying dynamics of processes in diverse areas, *e.g.*, childhood epidemics [7], lynx pelt data [9], vertical ground movements of an active caldera [10], and spatial patterns [11].

[a] e-mail: joy@mpipks-dresden.mpg.de

In this paper, we initially raise a question concerning the role of ρ as a reliable measure to distinguish chaos from noise. We show that by seeing only the ρ *vs.* T_p profile may infer incorrectly for cyclical chaotic process (a process with some degree of oscillation). Since nature is unbounded with such irregular cyclical phenomena, it will be noteworthy to pay attention to these processes. Here we use a relatively new method for the detection of multiple *amplitude modulated* cyclical components from a composite time series. The unique feature of this decomposition is that the method is completely data-adaptive, so the periodic waveform may be nonsinusoidal in nature. A new scheme of periodic prediction is also used. The variation of ρ for the original data series is statistically compared with the ρ for the surrogate data sets.

2 Modelling and prediction

The modelling and prediction procedure used involves the following steps: (i) Detection of the presence of a dominant amplitude modulated component. (ii) Extraction of the associated component. (iii) Repetition of steps (i) and (ii) until no further component is detectable. (iv) Long term prediction of the given series through the periodic prediction of the individual components.

The details of the procedure is as follows.

Consider an observable scalar time series $\{x(k), k = 1, 2, \ldots, T\}$, where T is the length of time series.

Consider the $m \times n$ matrix,

$$\mathbf{A}_n = \begin{bmatrix} x(1) & x(2) & \dots & x(n) \\ x(n+1) & x(n+2) & \dots & x(2n) \\ \vdots & \vdots & \ddots & \vdots \\ x(n(m-1)+1) & x(n(m-1)+2) & \dots & x(nm) \end{bmatrix} \quad (1)$$

It may be noted that the configured matrix $\mathbf{A}_n$ is different from the conventional trajectory matrix formed by the time-lagged delay vectors. This type of matrix configuration is advantageous in the context of latent cycle detection from the original data series. If the series is repetitive with cyclicity N, a rank redundancy will develop for the matrix $\mathbf{A}_n$ with $n = N$. We employ singular value decomposition (SVD) [12] for this purpose. For any matrix $\mathbf{A}_n$, it yields $\mathbf{A}_n = \mathbf{USV}^{\mathrm{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{S} = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_{\mathrm{p}} : \mathbf{0})$, the diagonal elements of $\mathbf{S}$ are called the singular values (s.v.) of $\mathbf{A}_n$. The singular values of the matrix $\mathbf{A}_n$ are precisely the lengths of the semi-axes of the hyper-ellipsoid E defined by $E = \mathbf{A}_n\mathbf{v} : ||\mathbf{v}||_2 = 1$. A physical analogy to this procedure is that if we think of the row vectors of the matrix $\mathbf{A}_n$ as the mass distribution of a body in the n-dimensional space, finding the right singular vectors corresponds to finding the axes of inertia of the body. The singular values are then the root-mean-square projections of its mass distributions onto each axis of inertia. If $\{x(k)\}$ is a periodic process with periodicity n, all the row vectors of $\mathbf{A}_n$ are equal or linearly correlated, so only one axis of inertia of an n-dimensional sphere is needed to describe the evolution of the process. This phenomenon is reflected in the rank of the matrix or more specifically in the singular values. The rank of $\mathbf{A}_n$ is equal to the number of nonzero s.v. of $\mathbf{A}_n$. The exact rank redundancy is developed for strict periodic process, but one will hardly get any periodic process in ideal sense. So when the row length is just equal to the underlying dominant periodicity or its higher integer multiples, one is expected to get matrices with high condition number though they may be of full rank. Since we are in search of a dominant repetitive component, we measure σ_1/σ_2 which acts as a measure of closeness to rank-oneness. For different configurations of $\mathbf{A}_n$ (with varying row length n), we find the ratio of the σ_1/σ_2. If the series $\{x(k)\}$ is having a latent rhythmic component of cycle length N, one is supposed to find a series of peaks at $n = iN$ (i, any positive integer) in the profile of this s.v. ratio *vs.* varying n. The range of i depends on the stability of the associated component. This spectrum of σ_1/σ_2 *vs.* the row length or orbit length (n) may be called the *p-spectrum* [13] due to its ability for the detection of any latent periodic component in an irregular cyclical series.

After detection of the dominant cycle N, configure $\{x(k)\}$ as matrix $\mathbf{A}_n$ with row length $n = N$ and find its sv-decomposition; $\mathbf{u}_1\sigma_1\mathbf{v}_1^{\mathrm{T}}$ is the *best cyclical component of period-N* obtainable from the series in least-squares sense. $\mathbf{u}_1\sigma_1$ (which we model as the series $\{g(k)\}$) represents the scaling factors or amplitudes associated with the successive periods or orbits and $\mathbf{v}_1^{\mathrm{T}}$ is the most representable pattern among all the orbits normalized to unit length. The time series formed by the successive rows of $\mathbf{u}_1\sigma_1\mathbf{v}_1\mathrm{T}$ will be at least nearly periodic with fixed period length N or it can be called as an amplitude modulated periodic time series. The unique feature of the proposed decomposition is that the periodic component may be non-sinusoidal in nature. Actually it depends on the nature of $\mathbf{v}_1$.

The matrix $(\mathbf{A}_N - \mathbf{u}_1\sigma_1\mathbf{v}_1^{\mathrm{T}})$ is converted to a residual time series $\{x_r(k)\}$. The p-spectrum of $\{x_r(k)\}$ will show presence of additional cyclical component (if any), which can be extracted the same ways as above, and the process is repeated. The extraction stops when the p-spectrum does not show any repetitive peaks.

In the context of estimation as well as prediction, we assume $\mathbf{v}_1^{\mathrm{T}}$ to remain stationary throughout the evolution process. So the prediction is to be performed for the $\{g(k)\}$ series. Since the length of the $\{g(k)\}$ series is short (usually ~ 20 or depends on the number of oscillations), a parsimonious model has to be produced. The prediction procedure is basically two folded. The variables are rotated by modified QR with column pivoting factorization [14] and best set of variables are selected through the minimization of the Mallow's C_{p} criterion [15]. So conventional p-step ahead prediction transforms in to p-period ahead prediction as $g(k+p|k)\mathbf{v}_1^{\mathrm{T}}$ where $g(k+p|k)$ is the p-step ahead prediction of the series $\{g(k)\}$ produced at time k; this is the inherent advantage of the periodic modelling over the existing prediction schemes.

The performance of the prediction is assessed through computing the Pearson's product moment correlation coefficient (ρ) between the original series and the predicted series. For pairs of quantities (x_i, y_i), ρ can be calculated as follows [16]:

$$\rho = \frac{\sum_{i=1}^{T_{\mathrm{p}}}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{T_{\mathrm{p}}}(x_i - \overline{x})^2 \sum_{i=1}^{T_{\mathrm{p}}}(y_i - \overline{y})^2}} \quad (2)$$

where $\overline{x}$ and $\overline{y}$ are the mean of the series $\{x_i\}$ and $\{y_i\}$, and T_{p} is the length of the series; here it refers to the prediction horizon. The values of ρ are typically 0 and 1 for uncorrelated and identical predictions respectively.

3 Results

Consider the 3 variable (X, Y, Z) Rössler attractor [17]:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -X - Z \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= X + 0.2Y \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= -0.4 + XZ - 5.7Z. \end{aligned} \quad (3)$$

The set of the equations is numerically integrated with the 4-th order Runge-Kutta method and first 3000 points is discarded as initial transients. We generate 4200 data

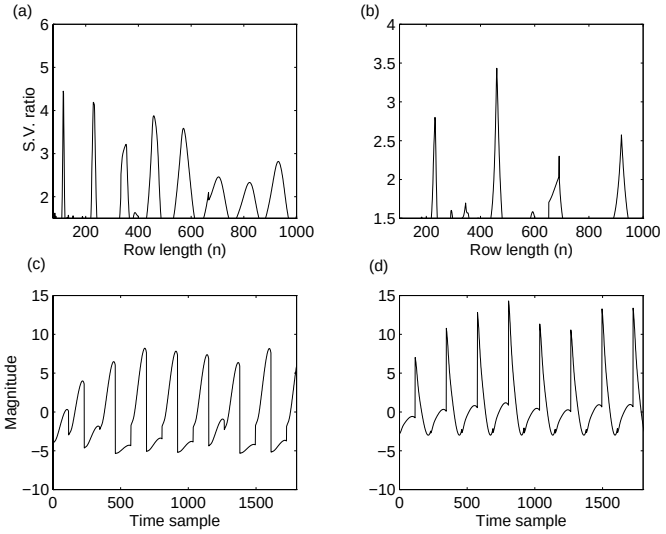

Fig. 1. (a) The p-spectrum of the X variable, showing peaks appearing at $n = 115, 230, 345, 460, \ldots$ clearly stand for the presence of one dominant component of periodicity 115. (b) The p-spectrum of the residual series obtained through the subtraction of the first cyclical component from the original series. The peaks in the p-spectrum at $n = 230, 460, 690, \ldots$ suggest presence of additional cyclical component of period 230. (c–d) The individual cyclical components extracted from the X variable of the Rössler map with periodicity 115 and 230 respectively. For the first component, antipersistence behaviour (large amplitude followed by small amplitudes and *vice versa*) in scaling factor sequences is obtained.

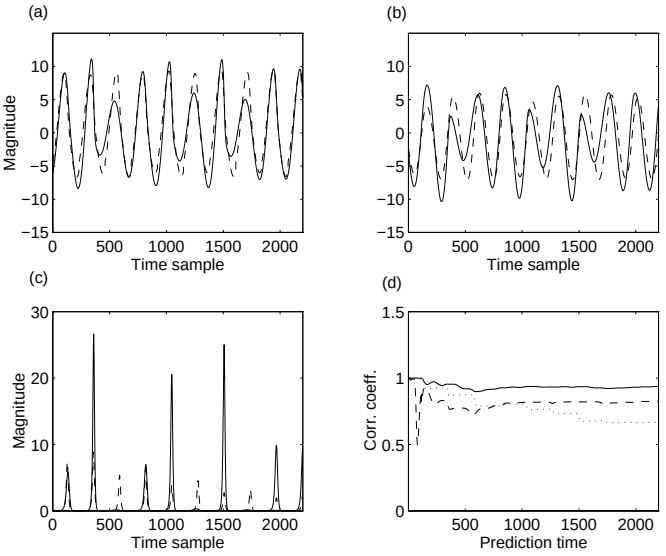

Fig. 2. (a–c) True predictions (original series in solid line), predicted series in dashed line) of X, Y, and Z variables. Note that for Z variable, comparatively poor amplitude matching is obtained due to its point process like behaviour (*i.e.* the energy being nonuniformly distributed throughout the flow). (d) Prediction performance in terms of ρ *vs.* $T_{\rm p}$ for the 3 states (X variable in solid, Y in dash, and Z in dotted line respectively).

points of each of the state variables. First 2000 data points are used for the estimation and the rest 2200 data points are used for assessing the prediction in out-of-sample sense (*i.e.*, when predicting the later part of the data, we have no access to it). This attractor has sharp peaks in the power-spectrum due to very small diffusion constant [19]. In the p-spectrum (Fig. 1a), we find the repetitive comparatively large peaks appearing at a regular interval at $n = 115$, 230, 345, 460, 575, ... indicating the presence of most dominant component of cycle length 115.

The first residual series is formed after subtracting this cyclical component from the original series. The p-spectrum of the first residual is shown in Figure 1b. Here also repetitive peaks are observed at $n = 230, 460, 690, \ldots$ which is due to an additional cyclical component of periodicity 230. In the subsequent residual series, no further periodicity is detectable through p-spectrum. So we model the X variable as the sum of two cyclical components of periodicity 115, and 230 respectively. The cyclical components for the two periodicities of 115 and 230 are shown in Figures 1c, d. It is very explicit that each cyclical component acts as an amplitude modulated wave where the modulated wave is fairly *nonsinusoidal*. And it has to be mentioned that although the periodicity of the second component is twice that of the first, it is not a harmonic one in the Fourier sense which can be easily justified by inspecting their patterns.

Similarly, we detect one cyclical component of periodicity 232 from the Y series, and one cyclical component of periodicity 230 from the Z series. If we compare the phase space of the true trajectories and the reconstructed through the estimated state variables (figures not shown), it suggests that the proposed method successfully captures the underlying dynamics of the Rössler equation macroscopically through the proposed periodic decomposition of each of the state variables.

Figures 2a–c show the true prediction (over a prediction horizon of 1 to 2200 data points) for the X, Y and Z variables respectively, based on the detected periodic components. For the X variable, initially good fit is obtained in the first 2 oscillations which is followed by an overestimation in amplitude prediction in the 3rd oscillation. Finally for the last 2 oscillations, we get again good match in terms of phase and amplitude. For Y and Z variables, we find that the predicted series are able to capture the *directions of excursions* of the states quite closely, although the amplitudes do not match so closely. So for each variable, the detection of the amplitude modulated component(s) is well justified. Figure 2d demonstrates the prediction performance in terms of correlation coefficient. For X variable, we find asymptotically flat ρ *vs.* $T_{\rm p}$ profile saturating at very high value (> 0.9). For Y variable, ρ drops to 0.5 around $T_{\rm p} = 80$ which is followed by a steep increase and finally saturates at high value (> 0.8). For Z variable, a step wise decrease curve is obtained, where the length of a step is somewhat related to the periodicity.

The narrow bandwidth of the power spectra of Rössler also indicates the existence of few number of predominant

frequency of oscillation[1]; thus uncertainty lies in the measurement of the time for one cycle of the periodic component, but it does not grow in time [20]. We think that for a cyclical series with or without chaoticity with almost fixed periodicity, a periodic model (like the one adopted) is always likely to produce high ρ irrespective of $T_{\rm p}$ unless the associated pattern is highly unstable and not always unidirectional. If one looks carefully Figure 2, in almost all the cases close phase (the term "phase" here is only related with the directions of excursions) match is obtained though mismatch in amplitudes are clear. Here the used statistic is correlation coefficient which is completely linear. So good prediction in period will yield very high ρ though the amplitudes may mismatch. That happens in our case also. The model captures the inherent dominant periodicity correctly which is reflected in the phase matching in the predicted domain. So highlighting only this profile, one is tempted to infer about the non-chaoticity of the process.

The same study is now continued on the surrogate data sets, where the null hypothesis is that the data is nothing but the monotonous transformation of a Gaussian linear process. The iterative scheme [18] is implemented so that the surrogate possesses the same power spectrum and distributions of the original data. We find that the surrogates of the X variable show only one periodic component ($\sim$ 230) and ρ sets at a much lower value than the prediction for original data, which is as expected. The two distributions (Fig. 3) are significantly different; Mann-Whitney rank-sum statistic (M) [21] gives 21.67[2]. Similar results have been found for Y and Z variables. So through the surrogate analysis, the determinism is corroborated for the Rössler series.

Next we analyse a simple linear stochastic model [22]:

$$W_{k+1} = 0.925(W_k + 1.07(W_k - W_{k-1}) + 0.25\eta) \qquad (4)$$

where η is white noise between -0.5 and 0.5. The model produces almost periodic time series with a narrow band power spectrum. Estimating over 500 data points, two periodic components are found with periodicity 23 and 22. Here very good estimation (energy of the error $\sim$ 3% of the original energy) is obtained, whereas in the prediction domain ρ decays very rapidly and sets to a much lower value, typically $\rho = 0.28$ for $T_{\rm p} = 430$. The surrogate of W also shows two periodicities and the ρ of the surrogates are compared with the original; the corresponding statistic (M) is 1.76 thus leading to the acceptance of the null hypothesis of linear Gaussian process.

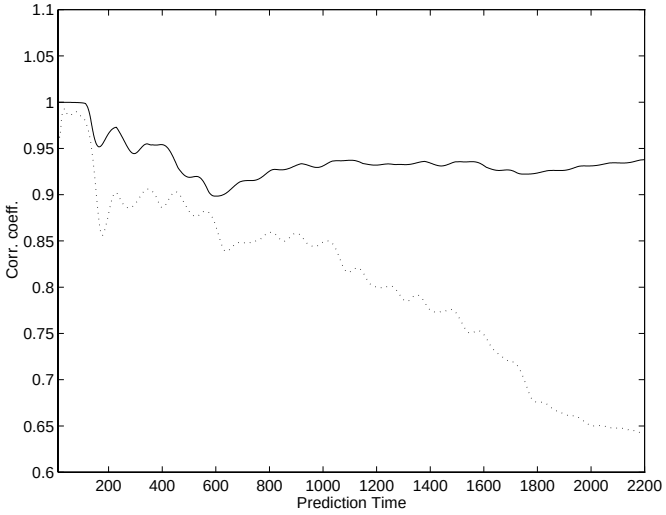

Fig. 3. Prediction performance (dotted line) of surrogates with the original data (solid line). The dashed curve is obtained by averaging the ρ *vs.* $T_{\rm p}$ curves for 10 set of surrogates.

It is also reported [23] that the initial decay (may be microscopically small) of ρ is somewhat related to the average rate of loss of information which gives an estimate of the Kolmogorov entropy (K_2), the sum of positive Lyapunov exponents. On the basis of certain statistical properties of the predicted series (equality of mean and variance), K_2 may be obtained by computing the initial slope of $\log(1-\rho)$ *vs.* $T_{\rm p}$. For the X variable of the Rössler series, using first 10 points (in this range, the value of ρ is greater than 0.9999), we get $K_2 = 0.135$ which is very close to the theoretical value =0.13 [24]. Similar analysis on the Y and Z variables yield 0.1491 and 0.1100 respectively. But it has to be noted that this estimation is very much dependent on the number of points used since accommodating longer prediction time span results in a deviation from required linearity.

For the much debated measles series of New York [7], we obtain two annual components. Here a high ρ (= 0.936) is obtained even for wide $T_{\rm p}$ (= 60 months). The surrogate of the measles series shows only one seasonal component, and ρ for $T_{\rm p} = 60$ is typically 0.823; the M-W statistic is 4.6, thus the measles series is accepted to be deterministic. Here the estimated entropy is 0.92 though the direct computations of the SEIR equations indicates values ranging between 0.4–0.5 bpy [25]. The overestimation may be due to the presence of the strong seasonal cycle [23].

4 Conclusions

The results presented here initially raises one concern about the role of ρ in differentiating low-dimensional

[1] We have tried with approximating the Rössler attractor with few sinusoidal components (= 5) estimated through Fourier spectra. In the estimated domain, good fit is obtained due to the forcing by least-squares, but the prediction is completely meaningless in terms of both phase and magnitude. So this also proves that although the power spectra of the Rössler attractor contains very sharp peaks, the underlying dynamics lies beyond the scope of the Fourier analysis. Again instead of using the proposed scheme of prediction, we have tried with mean value prediction for the $\{\mathbf{u}_1\sigma_1\}$ series and compute the ρ. But here it falls gradually thus exhibiting the superiority of the adopted scheme.

[2] For a two-tailed test at $\alpha = 0.05$, $t_{0.05(2),\infty} = 2.33$; if $|M| > 2.33$, we can say that the probability that the two sets of ρ *vs.* $T_{\rm p}$ distributions are random samples of the same distributions is < 0.01.

chaos from noise. Earlier numerous research reports have confirmed the underlying chaoticity of various systems through the decaying nature of ρ with increasing prediction time. With different examples, we have shown that great cautions are needed to interpret the profile of ρ. It has been shown here that for cyclical process, ρ remains flat or maintains high value irrespective of $T_{\rm p}$ when the system is periodically modelled. Through different examples along with the surrogate analysis, we have shown that the underlying dynamics can also be truly verified with such profile.

Helpful suggestions from Holger Kantz, David Wales are thankfully acknowledged. We are also grateful to Thomas Halsey for his valuable comments for the improvements of the paper. One of the authors (J. B.) also thanks Deutscher Akademischer Austauschdienst for financial assistance.

References

1. A. Babloyantz, A. Destexhe, Proc. Nat. Acad. Sci. USA **83**, 3513 (1986).
2. A.A. Tsonis, J.B. Elsner, Nature **333**, 545 (1988).
3. P. Grassberger, I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
4. J. Theiler, J. Opt. Soc. Amer. A. **7**, 1055 (1990).
5. J.D. Farmer, J.J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).
6. M. Casdagli, Physica D **35**, 335 (1989).
7. G. Sugihara, R.M. May, Nature **344**, 734 (1990).
8. E. Lorenz, J. Atmos. Sci. **26**, 636 (1969).
9. A. Mees, in *Nonlinear Modelling and Forecasting*, edited by M. Casdagli, S. Eubank (Addison-Wesley, 1992), pp. 3–24.
10. M. Cortini, C.C. Barton, Geology **21**, 239 (1993).
11. D.M. Rubin, Chaos **2**, 525 (1992).
12. D.S. Broomhead, G.P. King, Physica D **20**, 217 (1986).
13. P.P. Kanjilal, J. Bhattacharya, G. Saha, Phys. Rev. E **59**, 4013 (1999).
14. P.P. Kanjilal, G. Saha, T.J. Koickal, IEEE Trans. Syst. Man. Cybernet. B **29**, 1 (1999).
15. C.L. Mallows, Technometrics **15**, 661 (1973).
16. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, 1992).
17. N.H. Packard, J.P. Crutchfield, J.D. Farmer, R.S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).
18. T. Schreiber, A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).
19. M.G. Rosenblum, A.S. Pikovsky, J. Kürths, IEEE Trans. Circuit. Syst-I **44**, 874 (1997).
20. E.F. Stone, Phys. Lett. A **163**, 367 (1992).
21. J.H. Zar, *Biostatistical Analysis*, 2nd edn. (Prentice Hall, NJ, 1984).
22. C.J. Stam, J.P.M. Pijn, W.S. Pritchard, Physica D **112**, 361 (1998).
23. D.J. Wales, Nature **350**, 485 (1991).
24. A. Wolf, J.B. Swift, H. Swinney, J.A. Vastano, Physica D **16**, 285 (1985).
25. W.M. Schaffer, IMA J. Math. Appl. Med. Biol. **2**, 22 (1985).