acta physica slovaca vol. 45 No. 5, 567 – 574 October 1995

CORRELATION DIMENSION UNDERESTIMATION

Anna Krakovská[1]
Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 842 19 Bratislava, Slovakia

Received 16 May 1994, in final form 5 December 1994, accepted 7 December 1994

The Grassberger-Procaccia algorithm for the calculation of correlation dimension provides a value which underestimates the real correlation dimension, even for precise and noiseless data. This fact is connected with the data set size used for calculation. In this paper, relation between the amount of data points and precision of the dimension estimate is derived. Numerical experiments with Lorenz system are used to illustrate the results.

1. Introduction

This paper deals with correlation dimension of attractor and its estimation by the Grassberger - Procaccia algorithm (GPA) [1], [2]. The GPA is a widely used way of the identification of low-dimensional determinism. The result of this algorithm may provide an indication of chaos and information about the lower bound on the number of a suitable model's degrees of freedom. We focus our attention on some problems with the application of the GPA.

Let us consider dynamical systems described in terms of maps or differential equations. The dynamics of such systems is usually investigated in state space where the evolution of the system from an initial state corresponds to a trajectory. If the trajectories approach some subset of the state space, then this set is called an attractor. First of all, we are interested in nonlinear dynamical systems with chaotic behaviour. A typical feature of chaotic systems is a sensitivity to initial conditions what causes that any small uncertainty in the initial state estimation grows exponentially fast. Seemingly, this divergence of initially nearby trajectories cannot occur in the case of dissipative systems for which the volume of initial conditions contracts to some subset of the state space. In fact, the system with sensitive dependency on initial conditions can evolve to an attractor, but the dynamics must exhibit stretching and folding. It means that in some directions the trajectories of the system are stretched, whereas in other directions they have to be contracted. It gives rise to chaotic attractors - sets of highly complicated fractal structure [3].

[1]E-mail address: UMERKRAK@SAVBA.SAVBA.SK

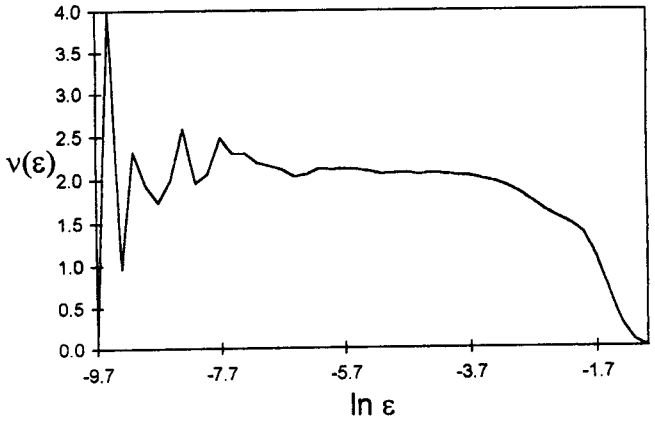

Fig. 1. Correlation exponent curve for Lorenz attractor. $M = 3$, $N = 40000$. Typical regions of correlation exponent curve can be distinguished: the plateau lies between the range of statistical fluctuations for vanishing ε and the range where ε exceeds the data set diameter.

Low-dimensional nonlinear systems which exhibit complex and apparently unpredictable behaviour resemble stochastic systems. In many cases of interest, distinguishing between determinism and stochasticity represents a difficult problem. In order to provide such a distinction a lot of new techniques have been developed. One of them is explained further.

2. Grassberger-Procaccia algorithm

The geometrical character of the attractor may provide an important information about the system. To characterize the structure of the attractor the spectrum of generalized dimensions D_q is widely used [1], [4]:

$$D_q = (q-1)^{-1} \lim_{\varepsilon \to 0} \frac{\ln \sum_{i=1}^{\mathcal{N}(\varepsilon)} p_i^q}{\ln \varepsilon}, \tag{1}$$

where $\mathcal{N}(\varepsilon)$ is the total number of hypercubes of dimension M and side length ε which cover the attractor, and p_i is the probability of finding a point in the hypercube i.

The three commonly investigated dimensions are D_0 (capacity dimension), D_1 (information dimension) and D_2 (correlation dimension). To be exact, we note that D_1 cannot be calculated by substituting $q = 1$ to (1). D_1 means $\lim_{q \to 1} D_q$.

We study the correlation dimension which can be calculated using the GPA. This algorithm is easy to implement but we will show that the obtained result have to be interpreted carefully. It is evident, that the correlation dimension of simple attractor take on integer values. For example, fixed point has dimension zero, limit cycle has dimension one, 2-torus has dimension two. Fractal attractors, however, are of noninteger correlation dimension. It means that estimating a finite noninteger correlation dimension indicates the presence of complex deterministic dynamics. Therefore, this

quantity may be used as a tool for distinguishing between stochasticity and determinism. Moreover, the value of the attractor dimension provides an information about the minimum number of degrees of freedom of the system which generates this attractor.

Next we recall the basic ideas which the GPA is based on [1]. From (1) one can see that correlation dimension is given by

$$D_2 = \lim_{\varepsilon \to 0} \frac{\ln \sum_{i=1}^{\mathcal{N}(\varepsilon)} p_i^2}{\ln \varepsilon} = \lim_{\varepsilon \to 0} \frac{\ln C_2(\varepsilon)}{\ln \varepsilon} , \tag{2}$$

where correlation integral $C_2 = \sum_{i=1}^{\mathcal{N}(\varepsilon)} p_i^2$ is the probability that a hypercube of volume ε^M contains two points of the attractor. It is approximately equal to the probability that the distance $|X_i, X_j|$ between two points X_i, X_j of the attractor is less than ε. Therefore, correlation integral C_2 can be approximated as

$$C_2 \approx \lim_{N \to \infty} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \theta(\varepsilon - |X_i, X_j|) ,$$

$$\theta(x) = \begin{cases} 0 & , \text{if } x < 0 \\ 1 & , \text{if } x > 0 . \end{cases}$$

N is the number of data used for calculation. Then the GPA simply counts the pairs of points with mutual distance less than ε.

From (2) one can see that, in order to find the correlation dimension, we have to plot $\ln C_2(\varepsilon)$ as a function of $\ln \varepsilon$ and follow the slope of the obtained curve. This slope $\nu(\varepsilon) = \frac{d(\ln C_2(\varepsilon))}{d(\ln \varepsilon)}$ is called correlation exponent, and the limit of $\nu(\varepsilon)$ for vanishing ε represents the correlation dimension. But the estimate of the $\lim_{\varepsilon \to 0} \nu(\varepsilon)$ is often not easy.

In order to analyse the behaviour of correlation exponent $\nu(\varepsilon)$ it is usefull to plot $\nu(\varepsilon)$ versus $\ln \varepsilon$. As it is shown in Fig. 1, for finite sample size this plot displays more regions of distinct types of behaviour. For small ε, the graph depends on a statistically unsufficient number of points, so for this part large differences in $\nu(\varepsilon)$ are typical. It means, that in the case of finite number of data, it is impossible to determine the limit behaviour of correlation exponent. If a flat part (plateau) follows, one assumes that the limit value has been reached here already. Then the value of the plateau is taken as the searched correlation dimension. Finally, the slope approaches zero for ε close to the diameter of the sample set. It is obvious that due to the poor statistics at small ε and the edge effect at large ε, only limited part of the graph is usable for the dimension estimation.

Recently, many authors have discussed the number of data necessary for correlation dimension calculation [2], [5], [6], [7]. Surprisingly, they have not obtained the same results. For instance, for reliable dimension estimation (5% error) of an attractor L.A.Smith requires 42^M data [5], and J.Theiler requires 5^M data [2], where M is the minimum dimension of space containing the whole attractor. The first requirement

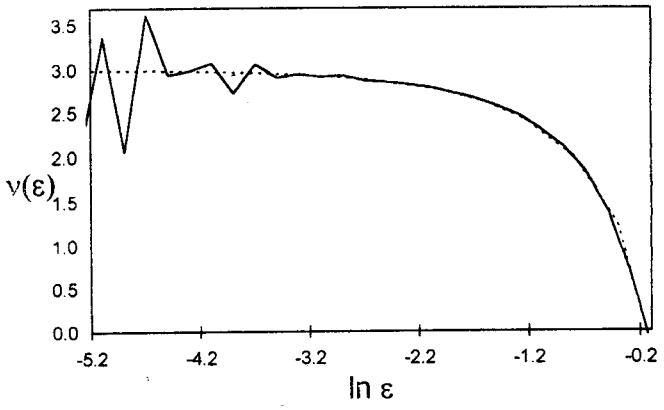

Fig. 2. Correlation exponent curve for uniformly distributed random data. $M = 3$, N is infinite (pointed), $N = 10000$ (full).

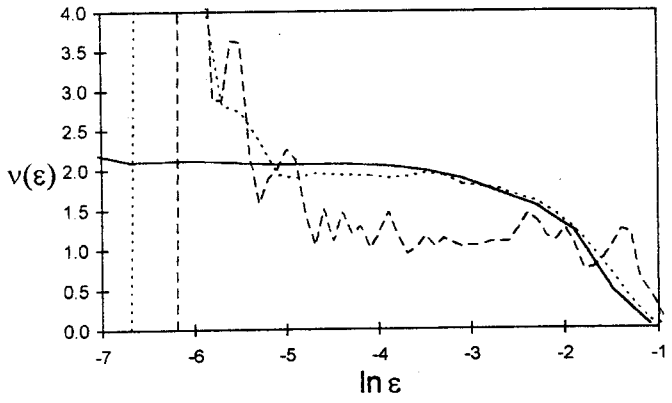

Fig. 3. Cut of correlation exponent curve for Lorenz attractor. Minimum embedding dimension $M = 3$, number of data $N = 125$ (dashed), $N = 1158$ (pointed) and $N = 74088$ (full).

is regarded to be too pessimistic the second one is too optimistic. Differences in the results have arisen because the argumentations of the authors contain one vague part regarding the acceptable range of plateau.

In this paper, we show what is the number of data needed for succesful application of the GPA and what is the possible underestimation of the resultant correlation dimension estimate. The validity of our results is clearly supported by numerical experiments.

3. Results

Let us derive how many data are necessary for the computation of correlation dimension by the GPA. In contrast to other analyses, our estimate of number of data necessary for succesful dimension calculation does not depend on expected width of the plateau. We only utilize the knowledge of underestimation for uniform M-dimensional hypercube. On that account, we consider calculation of correlation dimension for random data, similarly, as it has been done by L.A.Smith [5]. Since random data are not

produced by a finite-dimensional system, we expect that in M-dimensional space the GPA will result in the value M. It is true for infinite amount of points, but finite number of data leads to the underestimated value of correlation dimenson, as the following analysis shows.

Let us have a uniformly covered M-dimensional cube with edge length equal one. Then $C_2(\varepsilon)$ corresponds to the probability that the distance between two randomly chosen points is less than ε. It is easy to show that for one-dimensional cube (interval) this probability is

$$P(|x-y|<\varepsilon)=\varepsilon(2-\varepsilon)\ .$$

Next, the maximum norm is used, so that the distance $|X,Y|$ between two points $X=(x_1,...,x_M)$, $Y=(y_1,...,y_M)$ is given by

$$|X,Y|=max\{|x_i-y_i|,\ i=1,...,M\}\ .$$

Then in the case of uniformly covered M-dimensional cube the searched probability P is

$$P(|X,Y|<\varepsilon)=P(|x_1-y_1|<\varepsilon\wedge|x_2-y_2|<\varepsilon\wedge...\wedge|x_M-y_M|<\varepsilon)=(\varepsilon(2-\varepsilon))^M.$$

Thus

$$C_2(\varepsilon)=(\varepsilon(2-\varepsilon))^M$$

and

$$\nu(\varepsilon)=\frac{d\,(\ln C_2(\varepsilon))}{d\,(\ln\varepsilon)}=M\left(1-\frac{\varepsilon}{2-\varepsilon}\right)\ . \tag{3}$$

From Eq. (3) it is obvious that for nonzero ε the computed ν is less then the expected M, therefore it underestimates the true value. This underestimation is presented in Fig. 2. But the relation (3) is derived for the case of infinite amount of uniformly distributed points.

Let us assume a finite number of uniformly distributed points in hypercubes of increasing dimension M. The GPA is based on calculating the mutual distance of points. In the case of finite number of data, the most probable distance ε_{max} is of great importance. Because of poor statistics, for $\varepsilon<\varepsilon_{max}$ the correlation exponent $\nu(\varepsilon)$ presents sudden changes so this part of graph does not give usable information. If the number of uniformly distributed data points in the M-dimensional cube is N then $\varepsilon_{max}=N^{-\frac{1}{M}}$. Therefore,

$$\nu(\varepsilon_{max})=M\left(1-\frac{1}{2N^{\frac{1}{M}}-1}\right) \tag{4}$$

yields the maximum value, which is acceptable for N random data embedded in a M-dimensional cube. This statement is supported by numerical experiment (Fig. 2).

The relation (4) helps us to answer two practical questions.

1. How many data are necessary for the GPA if investigating uniformly distributed random points?
2. What is the measure of correlation dimension underestimation if the number of points used for the GPA is N and the dimension of space is M ?

To answer the first question, let us assume that we want to calculate the dimension with precision k. Then $\nu(\varepsilon_{max}) = Mk$, what results in the requirement $N = (\frac{k-2}{2k-2})^M$. If $k = 95\%$ than $N = (10.5)^M$. It means that obtaining an 5% underestimated result by the GPA demands using $(10.5)^M$ data. Of course, this is valid for uniformly covered M-dimensional hypercube.

The concrete example of the second problem may be the following one. For $N = 100000$ uniformly distributed points in cube of dimension $M = 5$ the estimated correlation exponent will be $1 - \frac{1}{2.100000^{\frac{1}{5}}-1} = 94.7\%$ of the expected value five, i.e. 4.73.

We note that the above analysis is valid for uniformly distributed points which represents the case of maximum value of underestimation. Now we can focus on deterministic systems. Their attractors are more "compressed", than a set of uniformly distributed points. Therefore in most cases using the same amount of data should result in higher precision in dimension estimation. Assume that the investigated attractor is embedded in M-dimensional space, where M is as small as possible. It means that $M = int(D_2)+1$. Here $int(D_2)$ denotes the integer part of the attractor's dimension. In the further text this value of M will be reffered to as the minimum embedding dimension. The value of correlation exponent for the most probable distance ε_{max} on the attractor we denote ν_p. In practice, ν_p is obtained as the value of the plateau. Then an analogy of (4) holds for deterministic systems:

$$\nu_p = D_2 \left(1 - \frac{1}{2N^{\frac{1}{D_2}} - 1}\right) ,$$

where ν_p is the value of the plateau, D_2 is the searched dimension, N is the number of data points, M is the minimum embedding dimension. The measure of underestimation is given by the expression in brackets. To approximate this measure we replace D_2 by M. It leads to the estimate

$$\nu_p > D_2 \left(1 - \frac{1}{2N^{\frac{1}{M}} - 1}\right) .$$

Then

$$D_2 < \nu_p \left(1 + \frac{1}{2(N^{\frac{1}{M}} - 1)}\right) . \tag{5}$$

where ν_p is the value of the plateau, D_2 is the searched dimension, N is the number of data points, M is the minimum embedding dimension.

Because of the above argumentation we suggest to determine the correlation dimension in case of deterministic systems as follows. First a clear plateau in correlation exponent function is required. If the plateau exists, its value yields the lower bound of the interval the investigated dimension lies in. The upper bound is given by (5).

To summarize, the next relation holds for the correlation dimension D_2:

$$D_2 \in \left(\nu_p, \quad \nu_p \left(1 + \frac{1}{2(N^{\frac{1}{M}} - 1)}\right)\right) \tag{6}$$

where ν_p is the value of the plateau, N is the number of data points, and M is the minimum embedding dimension.

In order to illustrate our results, some numerical experiments are presented in the following section.

4. Examples

1. First example is the uniformly covered hypercube. We will show that the analytically derived results are confirmed by numerical computations. Fig. 2 presents correlation exponent function for uniformly distributed data. The pointed line denotes the hypotetical curve of $\nu(\varepsilon)$ for infinite amount of points as it is given by Eq. (3). The example of finite amount of data points is drawn fully. In this case, the state space dimension coincides with the minimum embedding dimension so that $M = 3$. Number of data $N = 10000$. Eq. (4) says that $\varepsilon_{max} \approx 0.046$ ($\ln 0.046 \approx -3.08$) represents the most probable distance therefore $\nu(0.046) \approx 2.929$ is the maximum acceptable value of correlation exponent. The amount of pairs of points with mutual distance ε less than $\varepsilon_{max} \approx 0.046$ is decreasing and it leads to irregular changes in $\nu(\varepsilon)$. Fig. 2 clearly illustrates this effect. For $\ln\varepsilon > -3.08$ both curves coinside, for $\ln\varepsilon$ below -3.08 the full curve presents sudden changes so it is not more usable.

As random data are not generated by a finite-dimensional system, ν should theoretically equal 3. But using a finite number of data leads to underestimation and the curve of correlation exponent looks like indicating a deterministic dynamics. Now it is clear that a number of computed values of "low dimensions" in the literature has been a reflection of a small number of data rather than of a low-dimensional dynamics.

2. The second example concerns the Lorenz system,

$$\begin{aligned} \dot{x} &= 10(y - x) \\ \dot{y} &= -y - xz + 28x \\ \dot{z} &= xy - (8/3)z \quad . \end{aligned}$$

Correlation dimension of Lorenz attractor lies between 2.06 and 2.08. 95% of this value is between 1.957 and 1.976. In order to reach this value, $5^3 = 125$ data is recommended to use in paper [2]. The corresponding correlation exponent curve (dashed) in Fig. 3 shows that this amount of data is statistically unsufficient. On the other hand, $42^3 = 74088$ data, required in [5] leads to much better result than 95% of the true value of correlation dimension (full line). We have recommended using $10.5^3 = 1158$ data points and the calculation confirmes that this amount of data really leads to the plateau about 1.96.

5. Conclusion

The meaningful dimension estimation is conditioned by the existence of a clear plateau. But the determination of the correlation exponent plateau is often a difficult problem. Besides the influence of the amount of data, there are other aspects as the lacunarity of

the attractor or the additive noise in data which can destroy the linear scaling range. The effect of low-amplitude noise is often not significant but if the signal to noise ratio is below some critical value, minimizing the effect of noise becomes important. This question has been succesfully addressed by many authors and several nonlinear filtering methods have been employed to reduce the noise in complex nonlinear time series [9], [10], [11]. With reference to the above arguments, specifying the amount of data sufficient for creation of a clear plateau and, consequently, acceptable dimension estimation is not possible unless the quality of data is guarranteed.

It is clear that Grassberger-Procaccia method may be succesful only if having sufficiently large data set. Our main result says that this amount is about $(10.5)^M$ for random points, uniformly distributed over the M-dimensional cube. According to our experience $(10.5)^M$ data points of an attractor are also satisfactory for obtaining an evident plateau in the case of deterministic systems's attractor which is embedded in M-dimensional space. The points of the attractor are nonuniformly distributed so the underestimation will be lower than in the extreme case represented by the uniform hypercube. It results in the estimate (6). This relation allows very precise estimate of correlation dimension, what brings us closer to the exact values of correlation dimension for known chaotic attractors.

Acknowledgements This work was supported, in part, by Slovak Grant Agency for Science (grant No 999006/93). The author wish to thank P.Krakovský for his help with the computational part of this paper.

References

[1] P. Grassberger, I. Procaccia: *Physica* **9D** (1983), 189;
[2] J. Theiler: *J. Opt. Soc. Am. A* **7** (1990), 1055;
[3] B. Mandelbrot: *Fractals - Form, Chance and Dimension.* Freeman, San Francisco 1977;
[4] J. D. Farmer, E. Ott, J. A. Yorke: *Physica* **7D** (1983), 153;
[5] L. A. Smith: *Phys. Lett. A* **133** (1987), 283;
[6] M. A. H. Nerenberg, C. Essex: *Phys. Rev. A* **42** (1990), 7065;
[7] J.-P. Eckmann, D. Ruelle: *Physica D* **56** (1992), 185;
[8] C. Grebogi, E. Ott, J. A. Yorke *Phys. Rev. A* **38** (1988), 3688;
[9] E. J. Kostelich, J. A. Yorke: *Phys. Rev. A* **38** (1988), 1649;
[10] T. Sauer: *Physica D* **58** (1992), 193;
[11] T. Schreiber: *Phys. Rev. E* **48** (1993), R13;