Comparison of Novel Methods for Correlation Dimension Estimation

[1]H. Budáčová, [2,1]S. Štolc

[1]Institute of Measurement Science, Department of Theoretical Methods,
Slovak Academy of Sciences, Bratislava, Slovakia

[2]Austrian Institute of Technology, GmbH, Seibersdorf, Austria

Email: hana.budacova@savba.sk

Abstract. *We introduced and implemented two numerical methods, which estimate the correlation dimension from a finite set of data. The first is focused on identification of the scaling region in the correlation sum computed from the data. We find candidates for the linear region of various lengths and then combine the obtained results to compute the correlation dimension estimate. The second method uses Gaussian mixture method to predict behavior of the correlation integral and location of the scaling region. The aim of this paper is to compare these methods on various datasets with known correlation dimension.*

Keywords: Chaotic Attractors, Correlation Dimension, Scaling Region, Gaussian Mixture Model

1. Introduction

Correlation dimension is a measure of the dimensionality of a set of points and can be understood as a parameter that characterizes the complexity of strange attractors. It is a frequently used tool for detecting chaotic behavior in dynamical systems. According to [1] the correlation dimension (further denoted as D_2) of the attractor reconstructed from one variable using time delays and embeddings in higher dimensions is equal to the correlation dimension of the original attractor. These results engendered an extensive study of computational estimates of D_2 from measured data. In practice, this property can be used to help us understand the dynamics of many biological, chemical, climatological or financial dynamical systems.

In the past decades, the computational power experienced a rapid development. That leaves space for further improvements of methods that compute D_2 more accurately.

Definition of Correlation Dimension

Let us denote $x_1, x_2, \cdots, x_N$ the set of k-dimensional data points lying on a chaotic attractor of our interest. We follow the Grassberger-Procaccia algorithm [2] and define the correlation sum $C(r)$ as

$$C(r) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta\left(r - \left\|x_i - x_j\right\|\right) \tag{1}$$

where $\|\cdot\|$ computes the Euclidean distance, N is the number of the data points, $\Theta(\cdot)$ is called Heaviside function defined by

$$\Theta(s) = \begin{cases} 1 & \text{for } s > 0, \\ 0 & \text{for } s \leq 0. \end{cases} \tag{2}$$

The right-hand side of the Eq. (1) computes number of pairs of given data points whose distance is less than some given radius $r > 0$, normalized by the total number of pairs. Equivalently, it is the cumulative distribution of probability that two random points from given dataset are closer than r.

Furthermore, it can be interpreted as an estimate of the cumulative distribution of probability of the pair distances of the data on the original attractor.

Taking increasingly larger data sets, the probability is expected to behave as the power law r^{D_2} for small r. Subsequently, the logarithm of the correlation sum $C(r)$ is expected to be a linear function of the logarithm of r for small r. The correlation dimension is defined as the slope of this linear function, i.e.,

$$D_2 = \lim_{r \to 0} \lim_{N \to \infty} \frac{\partial \ln(C(r))}{\partial \ln(r)} \tag{3}$$

Problems with Finite Datasets

In practice, we encounter only finite datasets. Hence, the Eq. (3) cannot be applied exactly in practical computations. Our choice of the size of the dataset is restricted by various factors from measurement requirements and limitations (practical reasons) to time complexity of computations (computational reasons).

Due to the self-similarity of strange attractor, we assume that $\ln(C(r))$ behaves as a linear function of $\ln(r)$ for some range of r. The key problem is to detect this so-called scaling region. The size of the dataset limits the smallest reasonable choice of r. For very small r, due to the lack of data, there are none or only a few pairs within such a small distance. That makes $C(r)$ a poor estimate of the cumulative distribution function of pair distances for such small values. For r that approaches the actual size of the attractor, the edge effects begin to play an important role and $\ln(C(r))$ is no longer linear, but becomes saturated.

In the past years, the most common method to find the scaling region was the so-called visual inspection. The plot of $\ln(C(r))$ was shown to an expert, who decided where the linear region was and the slope in this region was denoted as D_2. Many have tried to improve and automatize this method. Surprisingly, many ideas failed, giving poorer results than this oracular method. In [3], the method where the middle quarter of the vertical axis is used to determine the scaling region (the time series is normalized to [0,1]) is mentioned. In [4] the approximation of $\ln(C(r))$ by a sum of linear and non-linear functions of $\ln(r)$ is used, followed by the limit of r to 0. The nonlinear term is chosen so that its limit is 0. This approach gives very good results especially for some systems with slow convergence. A new method using K-means is introduced in [5]. The latter paper contains also citations of several notable methods examined recently.

Aside from the aforementioned problems we also need to be careful with temporal and geometrical correlations. In a time series some points are close not due to geometrical attributes of the attractor, but due to the closeness in time. This was proven to cause underestimation of the correlation dimension and the cure called the Theiler window was introduced in [6]. Another factor that can affect the accuracy of computed results is the embedding dimension of the data. The higher the dimension, the more data we need to capture the attractor. The estimate of the size of the dataset and the underestimation of correlation dimension in case of deficient data can be found for example in [7]. In this paper we overcame these last two problems by choosing N to be large enough.

2. Subject and Methods

We propose two methods, each with its own way of dealing with the problem of identification of the scaling region. The main aim of our work was to find a method, which is accurate, fast and can be used to evaluate D_2 automatically for large datasets.

Method 1: Linear Region Finding (LRF)
In the first method we begin with computation of the correlation sum for logarithmically equidistant values of r. The data are normalized, so that the size of attractor is less than 1. We concentrate on finding the region on which $\ln(C(r))$ is linear and this region has certain defined length. For this purpose we use linear regression and least square method to compute and compare the error terms. Then we compare the obtained values of slopes and the position of these regions for various lengths and combine these results to obtain the value of D_2.

This method can be considered as an improved visual inspection method, with the slight difference that not the expert, but the computer chooses the linear region and computes D_2.

Method 2: Gaussian Mixture Model (GMM)
In the second method we use a two-pass approach to estimation of the cumulative distribution function of the pair distances and assessment of the scaling region.

In the first pass through the data, we begin with computing mean values and standard deviations of distances to the $1\text{st}, 2\text{nd}, 3\text{rd}, \ldots, (N-1)\text{-th}$ nearest neighbors in the dataset. We assume the distribution of i-th nearest neighbor is log-normal. Then, by using the Gaussian Mixture Model method, we compute a model-based estimate for the cumulative distribution $C(r)$.

In the second pass, we exploit estimate of $C(r)$ to generate a non-uniform binning of the interval between minimum and maximum distance found in the data, so that a reasonable number of pairs is accumulated in each bin. This approach efficiently addresses the problem of insufficient statistics for a good estimation $C(r)$ in small radii.

Finally the scaling region is assumed between the right boundary of the first bin and the radius where $\partial \ln(C(r)) / \partial \ln(r)$ reaches its maximum.

Data Specification
We tested both our algorithms on two different data groups:

1. Cantor set, Sierpinsky triangle, and Fractal pyramid,
2. Logistic map ($r = 4$), Normally distributed one-, two-, and three-dimensional data.

The first group was generated using random generator and iterative algorithm that placed each point deeper and deeper into the fractal structure. We decided for these fractal sets because their correlation dimensions are well known. The data for logistic map were obtained using iterative algorithm $x_{n+1} = r \cdot x_n \cdot (1 - x_n)$ with $r = 4$. The correlation dimension of this dataset is exactly 1 (proof can be found in, e.g., [4]). The last three datasets were generated by means of an independent identically distributed Gaussian pseudo-random generator which, in theory, delivers data with correlation dimension equal to the number of dimensions.

Each data set consisted of 100 000 points. In the experiments with smaller size of data we used a subsample of the original data.

3. Results and Discussion

In Table 1., the results provided by our methods are shown. GMM gives in general better results than LRF. The computations from randomly distributed data are falling behind with increasing dimension, which is in consensus with the ideas in Introduction. This happens because in higher dimensions more data points are needed to capture the behavior of the correlation integral with the same accuracy. It follows that the accuracy must drop if the number of data is kept constant.

The poorest accuracy is obtained in the computation of D_2 of the logistic map. In [4] one can find the proof that the convergence of the correlation sum is very slow. It is only natural that, in such a case, much more data is needed to reach certain accuracy and that is the reason why our results are underestimated. As already mentioned, in [4] authors address this issue by adding a non-linear term, which led very precise results limiting r to 0. However, the results for 1D and 2D random data are more accurate with our methods, with even less data points that they had in their experiment (they obtained 1.072 and 2.133 for N greater than 10^6).

Table 1. Comparison of computed values of correlation dimension by the Linear Region Finding (LRF) method and Gaussian Mixture Model (GMM) method for $N = 10\,000$ and $N = 100\,000$ with the exact values. The grayed cells indicate the winning method given the dataset and its size.

Dataset	$N = 10\,000$		$N = 100\,000$		Theoretical value
	LRF	GMM	LRF	GMM	
Cantor set	0.6201	0.6298	0.6217	0.6297	0.6309
Sierpinsky triangle	1.5818	1.5814	1.5683	1.5811	1.5849
Fractal pyramid	2.3041	2.3179	2.3099	2.3110	2.3219
Logistic map (r=4)	0.9041	0.9284	0.9041	0.9254	1.0000
1D normal data	0.9964	1.0002	0.9999	1.0023	1.0000
2D normal data	1.9798	1.9815	1.9923	1.9931	2.0000
3D normal data	2.9420	2.9489	2.9780	2.9766	3.0000

4. Conclusions

We have shown that the results computed by the GMM and LRF methods are in good consensus with the exact values of the correlation dimension of various tested datasets. Moreover, the GMM method gives consistently better results than the LRF method.

It is quite clear, that these methods work well for computer generated noise-free data. However, most real-world data are usually contaminated with some level of noise. In many practical situations, one may need to reconstruct the system from a small number observed variables. That means our methods have to be tested under such conditions before applied to such tasks. The level of accuracy that we obtained in our existing experiments hints on promising results in this area too.

Acknowledgements

This work has been supported by the Slovak Grant Agency for Science (project No. 2/0043/13) and the Slovak Research and Development Agency (project No. APVV-0096-10).

References

[1] Sauer T.D., Yorke J.A. Are the dimensions of a set and its image equal under typical smooth functions? *Ergodic theory and dynamical systems,* 17(4): 941-956, 1997.

[2] Grassberger P., Procaccia I. Measuring the strangeness of strange attractors. *Physical Review Letters*, 50(5):346 - 349, 1983.

[3] Mekkler A. Calculation of EEG correlation dimension: Large massifs of experimental data. *Computer methods and programs in biomedicine,* 92(1):154-160, 2008.

[4] Sprott J.C., Rowlands G. Improved correlation dimension calculation. *International journal of bifurcation and chaos*, 11(7):1865 - 1880, 2001.

[5] Ji C.C., Zhu H., Jiang W. A novel method to identify the scaling region for chaotic time series correlation dimension calculation. *Chinese Sci Bull*, 56:925 - 932, 2011.

[6] Theiler J. Spurious dimension from correlation algorithms applied to limited time-series data. *Physical Review A*, 34(3):2427 - 2432, 1986.

[7] Krakovská A. Correlation dimension underestimation. *Acta physica slovaca*, 45(5):567 - 574, 1995.