IDIAP RESEARCH REPORT

Estimating the Intrinsic Dimension of Data with a Fractal-Based Method

Francesco Camastra [a]

Alessandro Vinciarelli [b]

IDIAP–RR 02-02

January 2002

to appear in

IEEE Transactions on Pattern Analysis and Machine Intelligence

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11

fax +41 – 27 – 721 77 12

e-mail secretariat@idiap.ch

internet http://www.idiap.ch

[a] INFM-DISI, University of Genova, Via Dodecaneso 35, 16146 Genova, Italy (e-mail: camastra@disi.unige.it).

[b] Institut Dalle Molle d'Intelligence Artificielle Perceptive, Rue du Simplon 4, 1920 Martigny, Switzerland (e-mail: alessandro.vinciarelli@idiap.ch).

IDIAP Research Report 02-02

Estimating the Intrinsic Dimension of Data with a Fractal-Based Method

Francesco Camastra Alessandro Vinciarelli

January 2002

to appear in
IEEE Transactions on Pattern Analysis and Machine Intelligence

Abstract. In this paper, the problem of estimating the Intrinsic Dimension of a data set is investigated. A fractal-based approach using the Grassberger-Procaccia algorithm is proposed. Since the Grassberger-Procaccia algorithm performs badly on sets of high dimensionality, an empirical procedure, that improves the original algorithm, has been developed. The procedure has been tested on data sets of known dimensionality and on time series of Santa Fe competition.

1 Introduction

Pattern Recognitions problems involve data represented as vectors of dimension d. The data is then embedded in the space $\mathbb{R}^d$, but this does not mean necessarily that its Intrinsic Dimension (ID) is d. The ID of a data set is the minimum number of free variables needed to represent the data without information loss. In more general terms, following Fukunaga [9], a data set $\Omega \subset \mathbb{R}^d$ is said to have an ID equal to M if its elements lie entirely within an M-dimensional subspace of $\mathbb{R}^d$ (where $M < d$).
Estimation of the ID is important for many reasons. The use of more dimensions than strictly necessary leads to several problems. The first one is the space needed to store the data. As the amount of available information increases, the compression for storage purposes becomes ever more important. The speed of algorithms using the data depends on the dimension of the vectors, so a reduction of the dimension can result in saving computational time. Moreover, in the statistical learning theory approach [32], the capacity and the generalization capability of the classifiers depend on ID and the use of vectors with smaller dimension often leads to better performance in classification. Finally, when using an autoassociative neural network [18] to perform a nonlinear feature extraction (e.g. nonlinear PCA), the ID can suggest a reasonable value for the number of hidden neurons.
This paper presents an approach, based on fractal techniques, to ID estimation. Fractal techniques have been successfully applied to estimate the attractor dimension of underlying dynamic systems generating time series [17]. The literature presents results in the study of chaotic systems (e.g. Hénon map, Rössler oscillator)[22], in the analysis of ecological time series (e.g. Canadian lynx population) [15], in biomedical signal analysis [31], in radar clutter identification [12] and in the prediction of financial time series [24]. Notwithstanding this, in pattern recognition, fractal methods are mainly used to measure the fractal dimension of an image [13]. As far as we know, the application of fractal approaches to the problem of ID estimation has never been proposed before. The proposed ID estimation method is tested on both artificial and real data showing good results.
The structure of this paper is as follows: Section 2 presents several techniques for estimating ID; in Section 3 fractal methods are reviewed; the procedure to estimate ID is described in Section 4; in Section 5 some experimental results are reported; in Section 6 some conclusions are drawn.

2 ID estimation techniques

Following the classification proposed in [16], there are two approaches for estimating ID. The first (*global*) estimates the dimensionality of the data set as a whole. The second (*local*) estimates ID using the information contained in sample neighborhoods and avoiding the projection of the data onto a lower-dimensional space.
The literature presents only a few results, mostly based on *projection techniques* and topological methods (*topological dimension of data* [3]). Projection techniques, which are generally used as global methods, search for the best subspace to project data by minimizing the reconstruction error. These methods can be divided into two families: linear and nonlinear. Linear methods such as *Principal Component Analysis* (PCA) are inadequate estimators since they tend to overestimate the ID[1]. For example, a set formed by points lying on a curve for PCA has dimension 2 rather than 1. On the other hand, though nonlinear PCA [18] performs better than linear PCA in some contexts [7], it presents problems when estimating ID [19]. Among projection techniques it is worth mentioning the *Whitney reduction network* recently proposed by Broomhead and Kirby [2][18]. This method is based on Whitney's concept of *good projection*, namely a projection obtained by means of an injective mapping. As pointed out in [18], finding good projections can be difficult and can sometimes involve empirical considerations. Finally, the fractal approach presented in this paper can be classified as a global method.
Topological methods, that are local, try to estimate the *topological dimension* of the data manifold. Some authors [8] use Topology Representing Networks (TRN) [21], that optimally represent the data topology, to estimate ID. They use the number n of cross-correlations learnt by each neuron as an

indicator of the local dimension of Ω.
The number n of cross-correlations is used as an indicator of the local dimension of Ω, assuming that n is quite close to the number k of spheres which touch a given sphere, in the *Sphere Packing Problem*[4]. This approach presents some drawbacks: the above assumption has not yen been proved, the number k is known exactly only for dimension values between 1 and 8, and k tends to grow exponentially with the space dimension. This last peculiarity strongly limits the use of the conjecture in practical applications where data can have high dimensionality. Other authors [10] perform a Voronoi tesselation of data space and in each Voronoi set a local PCA is performed. Finally, Bruske and Sommer [3] after the Voronoi tesselation of the data space, compute, by means of TRN, an optimal topology preserving map G. Then for each node $i \in G$, a PCA is performed on the set Q_i consisting of the differences between the node i and all of its m_i neighbours in G. The above mentioned [3][10] methods present some limits: since, due to noise, none of the eigenvalues of the covariance matrix will be null it is necessary to use heuristic thresholds in order to decide whether an eigenvalue is significant or not.

3 Fractal Methods

A unique definition of the dimension has not been given yet. A popular definition, among many proposed, is the so-called *Box-Counting* dimension [20], belonging to a dimension definition family defined *fractal* [5]. The Box-Counting dimension is a simplified version of Haussdorff dimension [5][22]. The Box-Counting dimension D_B (or *Kolmogorov Capacity*) of a set Ω is defined as follows: If $\nu(r)$ is the number of the boxes of size r needed to cover Ω, then D_B is

$$D_B = \lim_{r \to 0} \frac{\ln(\nu(r))}{\ln(\frac{1}{r})}. \tag{1}$$

Unfortunately, the Box-Counting dimension can be computed only for low-dimensional sets, because the algorithmic complexity grows exponentially with set dimension. Therefore, in our opinion, a good substitute for the Box-Counting dimension can be the *Correlation Dimension* [11]. Correlation dimension is successfully used, because of its computational simplicity, to estimate the dimension of attractors of dynamical systems. Correlation dimension is defined as follows: let $\Omega = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be a set of points in $\mathbb{R}^n$ of cardinality N. If the *Correlation Integral* $C_m(r)$ is defined as:

$$C_m(r) = \lim_{N \to \infty} \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} I(\|\mathbf{x}_j - \mathbf{x}_i\| \le r) \tag{2}$$

where I is an *indicator function*[1], then the *Correlation Dimension* D of Ω is:

$$D = \lim_{r \to 0} \frac{\ln(C_m(r))}{\ln(r)}. \tag{3}$$

It has been proved [11] that the Correlation Dimension is a lower bound of the Box-Counting dimension, but because of noise, the difference between the two is negligible in real applications. Some methods [28][27] have been studied to obtain an optimal estimate for the correlation dimension, but all of these techniques work only when the correlation integral assumes a given form [2], otherwise the estimators can perform poorly [29]. Moreover, these methods generally require some heuristics to set the radius r [30]. Therefore, in our work we used the original procedure (*GP algorithm*) proposed by Grassberger and Procaccia that consists of plotting $\ln(C_m(r))$ versus $\ln(r)$ and of measuring the slope of the linear part of the curve (Figure 1).
It has been proved [6][26] that in order to get an accurate estimate of the dimension D, the set cardinality N has to satisfy the following inequality:

$$D < 2 \log_{10} N. \tag{4}$$

[1] $I(\lambda)$ is 1 iff condition λ holds, 0 otherwise.
[2] For example, Takens' method is optimal only iff $C_m(r) = ar^D[1 + br^2 + o(r^2)]$ where a, b are constants.

Inequality 4 shows that the number N of data points needed to accurately estimate the dimension of a D-dimensional set is at least $10^{\frac{D}{2}}$. Even for low dimensional sets this leads to huge values of N.
The effect of N on the measure of the dimension can be seen in table 1. This table reports the value of the measures of D obtained using the GP algorithm over sets of points randomly distributed in 10-dimensional hypercubes (supposed to have $D = 10$). The dimension is measured for different values of N and the error with respect to the supposed true dimension is reduced by increasing the number of data points from 10^3 to $10^{\frac{D}{2}} = 10^5$.
In order to cope with this problem and improve the reliability of the measure for low values of N, we developed an empirical procedure as described in the following section.

4 Intrinsic Dimension Estimation Procedure

Consider the set Ω of cardinality N. Our empirical procedure (EP) consists of the following steps:

1. A set Ω', whose ID d is known, with the same cardinality N as Ω is created. For instance, Ω' could be constituted by N points randomly generated in a d-dimensional hypercube.

2. The correlation dimension D of Ω' is measured with the GP algorithm.

3. The previous steps are repeated for T different values of d. The set $C = \{(d_i, D_i) : i = 1, 2, \ldots, T\}$ is obtained.

4. A best-fitting to the points in C is performed. A plot (*reference curve*) Γ of D versus d is generated (see figure 2). The reference curve allows to estimate the value of D when d is known.

5. The Correlation Dimension D of Ω is computed and, using Γ, the Intrinsic Dimension of Ω can be estimated.

The previously described method is based on the following assumptions:

1. Γ depends on N.

2. Since the GP algorithm gives close estimates on sets of the same dimensionality and cardinality, the dependence of Γ on the Ω' sets used for its setup is negligible.

In comparison with topological methods, our approach presents the following advantages: it allows one to estimate the ID of high-dimensional data, unlike TRN-based method. Moreover, the approach proposed is based on the estimation of a fractal dimension and therefore allows one to obtain non-integer values. The latter is quite important, since due to the presence of noise, real data can sometimes lie within a *fractal-like* submanifold, whose dimension is usually non-integer.

5 Experimental Results

The EP was tested by first creating reference curves for different values of the cardinality N, then by using each one of them to estimate the dimension of data sets with the same cardinality and known dimension.
During our experimentation the sets Ω' used for the reference curve setup are formed by points randomly generated in a d-hypercube. A plot was generated for the following cardinality values: 1000, 2000, 5000, 10000, 30000 and 100000. In correspondence with each value, a pair (d, D) was calculated for

$$d \in \{2, 3, 5, 10, 15, 18, 20, 25, 28, 30, 33, 35, 38, 40, 43, 45, 48, 50\} \tag{5}$$

The plot function is estimated by a Multi-Layer-Perceptron (MLP) [1]. Its structure (input, hidden and output layer composed of 1,2 and 1 neurons respectively) was set up by means of the *Bayesian Information Criterion* [25]. The resulting reference curves can be seen in figure 2.

In order to test the method, several sets (with cardinalities corresponding to the values indicated above) were created composed of random points generated in hypercubes in spaces with dimension 8 and 23. These sets were assumed to have ID 8 and 23 respectively and were not used in generating the reference curves Γ.
Following the procedure described in section 4, the GP algorithm is first applied for each set, then the plot Γ corresponding to the same cardinality as the set being measured is used to compute the ID. The results are reported in table 2.
The table shows the dimension estimation obtained with the GP algorithm and with the empirical procedure proposed here. A remarkable improvement is obtained indeed when the cardinality is low.
Afterwards in order to validate the EP procedure, the data set A [14] and D [23] of the Santa Fe time series competition were considered. Data Set A is a real data time series generated by a Lorenz-like Chaotic System, implemented by NH_3-FIR lasers. The data set D is a synthetic time series generated by a particle motion, simulated on a computer, with nine freedom degrees. The goal of the experimentation was to estimate, with the GP procedure, the attractor dimension of time series A and D. In order to estimate the attractor dimension, we used the method of delays [17][22].
Given a time series $x(t)$, with $(t = 1, 2, \dots, N)$, a set of points $\{X(t) : X(t) = [x(t), x(t-1), \dots, x(t-d+1)]\}$ is obtained. If d is adequately large, between the manifold M, generated by the points $X(t)$ and the attractor U of the dynamic system that generated the time series, there is a diffeomorphism. Therefore it is adequate to measure the dimension of M to infer the dimension of U.
We applied the method of delays to the data set A, considering its first 1000, 2000, 5000, 10000 points. The results, obtained with the GP and EP algorithm, are reported in table 3. Since the value of the fractal dimension of the attractor of Lorenz's system is approximately 2.06, the result can be considered satisfactory. Then we applied the method of delays to the data set D, considering the first 1000, 2000, 5000, 10000, 30000 points. The results, with GP and EP, are shown in table 3. Since the system that generated the data D has 9 freedom degrees, the result can be considered particularly satisfactory.

6 Conclusions

This work presented a new Intrinsic Dimension estimation technique based on a fractal method, namely the Grassberger-Procaccia algorithm. GP algorithm is effective when the dimension is low, but presents severe limits for high dimensionalities. Therefore an empirical procedure, that improves GP algorithm, has been developed.
A comparison performed over sets, randomly generated in hypercubes of known dimensionality shows that the new proposed procedure allows a significant improvement of the GP algorithm for low dimensional sets. Further experiments, performed on data series of the Santa Fe competition, confirm this opinion. The proposed procedure is based on the assumption that the dependence of reference curves on the data used for their generation is negligible. Therefore further work will consist in applying the method over different kinds of data in order to confirm the validity of this assumption.

References

[1] C. Bishop, *Neural Networks for Pattern Recognition.* Cambridge, UK: Cambridge University Press, 1995.

[2] D.S. Broomhead and M.Kirby, "A New Approach to Dimensionality Reduction: Theory and Algorithms", *SIAM Journal of Applied Mathematics*, vol. 60, no. 6, pp. 2114-2142, 2000.

[3] J. Bruske and G. Sommer, "Intrinsic Dimensionality Estimation with Optimally Topology Preserving Maps", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 572- 575, May 1998.

[4] J.H. Conway and N.J.A. Sloane, "Sphere Packings, Lattices and Groups", Grundlehren der mathematischen Wissenschaften 290. Springer-Verlag, New York, 1988.

[5] J.P. Eckmann and D. Ruelle, "Ergodic Theory of Chaos and Strange Attractors", *Review of Modern Physics*, vol. 57, pp. 617-659, 1985.

[6] J.P. Eckmann and D. Ruelle, "Fundamental Limitations for Estimating Dimensions and Lyapounov Exponents in Dynamical Systems", *Physica*, vol. D56, pp. 185-187, 1992.

[7] D.Fotheringhame and R.J.Baddeley, "Nonlinear principal component analysis of neuronal spike train data", *Biological Cybernetics*, vol. 77, pp. 282-288, 1997.

[8] F. Frisone, F. Firenze, P. Morasso and L. Ricciardiello, "Application of Topological-Representing Networks to the Estimation of the Intrinsic Dimensionality of Data", *Proceedings of International Conference on Artificial Neural Networks*, Paris, France, 1995.

[9] K. Fukunaga, "Intrinsic Dimensionality Extraction", *Classification, Pattern Recognition and Reduction of Dimensionality, Vol. 2 of Handbook of Statistics*, P.R. Krishnaiah and L.N. Kanal, eds., pp. 347-360. Amsterdam: North Holland, 1982.

[10] K. Fukunaga and D.R. Olsen, "An Algorithm for Finding Intrinsic Dimensionality of Data", *IEEE Transactions on Computer*, vol. 20, no. 2, pp. 165-171, 1976.

[11] P. Grassberger and I. Procaccia, "Measuring the Strangeness of Strange Attractors", *Physica* vol. D9, pp. 189-208, 1983.

[12] S.Haykin and X. Bo Li, "Detection of Signals in Chaos", *Proceedings of IEEE*, vol. 83, no. 1, pp. 95-122, 1995.

[13] Q. Huang, J.R. Lorch, and R.C. Dubes, "Can the Fractal Dimension of Images be Measured?", *Pattern Recognition*, vol. 27, no. 3, pp. 339-349, 1994.

[14] U.Hübner, C.O. Weiss, N.B. Abraham and D.Tang, "Lorenz-like Chaos in NH_3-FIR Lasers", *Time Series Prediction. Forecasting the Future and Understanding the Past*, A.Weigend and N.A. Gershenfeld, eds., pp. 73-104. Reading, MA: Addison Wesley, 1994.

[15] V. Isham, "Statistical Aspects of Chaos: a Review", *Networks and Chaos-Statistical and Probabilistic Aspects*, O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall, eds., pp. 124-200. London: Chapman & Hall, 1993.

[16] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice Hall, 1988.

[17] D.Kaplan and L.Glass, *Understanding Nonlinear Dynamics.* New York, NY: Springer-Verlag, 1995.

[18] M.Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns.* New York, NY: John Wiley and Sons, 2001.

[19] E.C. Malthouse, "Limitations of nonlinear PCA as performed with generic neural networks", *IEEE Trans. on Neural Networks*, vol. 9, no.1, pp. 165-73, 1998.

[20] B. Mandelbrot, *Fractals: Form, Chance and Dimension.* San Francisco, CA: Freeman, 1977.

[21] T. Martinetz and K. Schulten, "Topology Representing Networks", *Neural Networks*, vol. 3, pp. 507-522, 1994.

[22] E. Ott, *Chaos in Dynamical Systems.* Cambridge, UK: Cambridge University Press, 1993.

[23] F.J.Pineda and J.C.Sommerer, "Estimating Generalized Dimensions and Choosing Time Delays: A Fast Algorithm", *Time Series Prediction. Forecasting the Future and Understanding the Past*, A.Weigend and N.A. Gershenfeld, eds., pp. 367-385. Reading, MA: Addison Wesley, 1994.

[24] J.A. Scheinkman and B. Le Baron, "Nonlinear Dynamics and Stock Returns", *Journal of Business*, vol. 62, pp. 311-337, 1989.

[25] G. Schwartz, "Estimating the Dimension of a Model", *Annals of Statistics*, vol. 6, pp. 497-511, 1978.

[26] L.A. Smith, "Intrinsic Limits on Dimension Calculations", *Physics Letters*, vol. A133, pp. 283-288, 1988.

[27] R.L. Smith, "Optimal Estimation of Fractal Dimension", *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity vol. XII*, M. Casdagli, S. Eubank, eds., pp. 115-135. New York: Addison-Wesley, 1992.

[28] F. Takens, "On the Numerical Determination of the Dimension of an Attractor", in B. Braaksma, H. Broer and F. Takens (eds), "Dynamical Systems and Bifurcations, Proceedings Groningen 1984", Lecture Notes in Mathematics No. 1125, pp. 99-106, Springer-Verlag, Berlin, 1985

[29] J. Theiler, "Lacunarity in a Best Estimator of Fractal Dimension", *Physics Letters*, vol. A133, pp. 195-200, 1988.

[30] J. Theiler, "Statistical Precision of Dimension Estimators", *Physical Review*, vol. A41, pp. 3038-3051, 1990.

[31] W.S. Tirsch, M. Keidel, S. Perz, H. Scherb and G. Sommer, "Inverse Covariation of Spectral Density and Correlation Dimension in Cyclic EEG Dimension of the Human Brain", *Biological Cybernetics*, vol. 82, pp. 1-14, 2000.

[32] V.Vapnik, *Statistical Learning Theory.* John Wiley and Sons, NY: 1998.

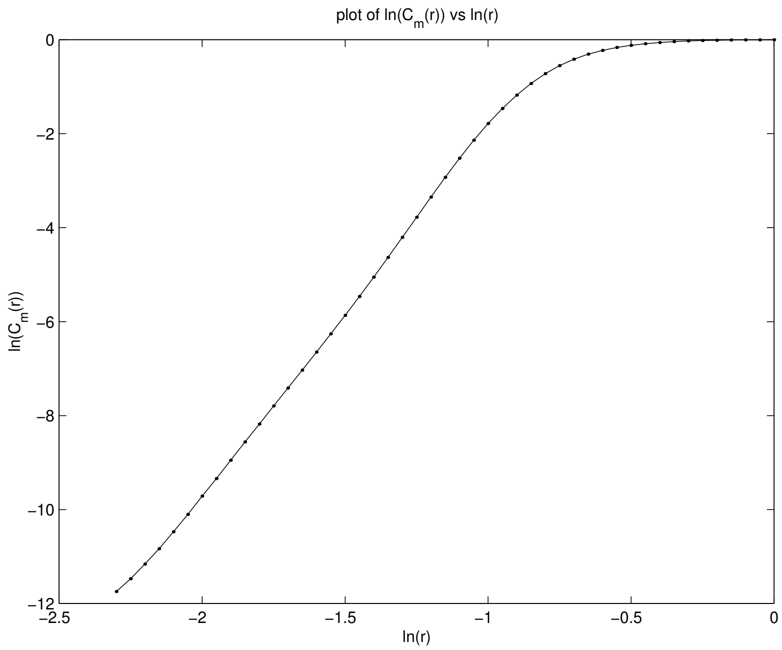

Figure 1: Plot of $\ln(C_m(r))$ vs $\ln(r)$.

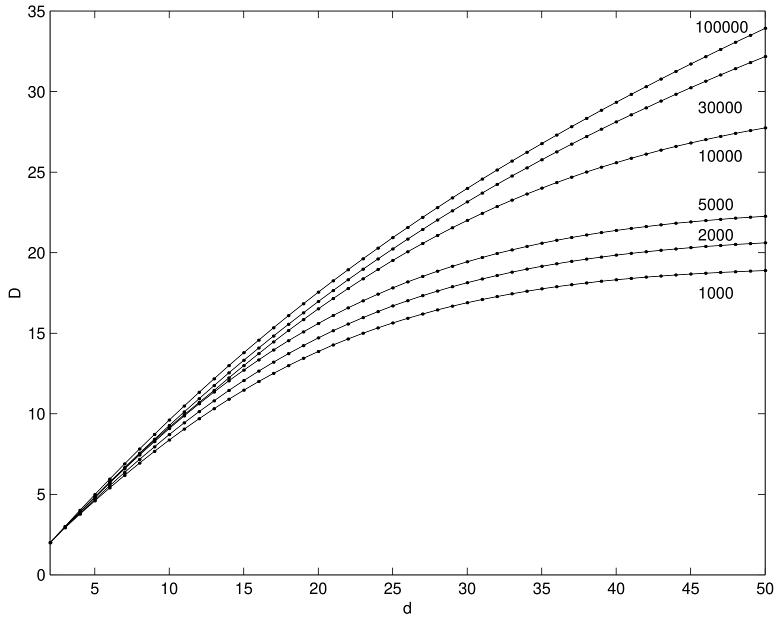

Figure 2: Reference curves for different values of the cardinality N.

points number	estimated dimension
1000	7.83
2000	7.94
5000	8.30
10000	8.56
30000	9.11
100000	9.73

Table 1: Dependence of the estimated correlation dimension on the number of data points used (the actual dimension of data is 10).

points	GP (d=8)	EP (d=8)	GP (d=23)	EP (d=23)
1000	6.83	7.86	14.99	22.95
2000	6.94	7.75	15.76	22.48
5000	7.42	7.98	17.09	23.21
10000	7.51	8.20	18.04	22.43
30000	7.65	8.13	19.10	23.20
100000	7.83	8.03	19.78	23.24

Table 2: ID estimation, by the GP algorithm and EP, of 8 and 23-dimensional data sets. The first column reports the number of points used for the estimation. The second and third columns show the value of d estimated with GP and EP respectively for the 8-dimensional set. The last two columns show the estimates of d obtained with GP and EP for the 23-dimensional set.

points	GP (D)	EP (D)	GP (A)	EP (A)
1000	7.54	8.84	2.00	2.01
2000	7.87	8.90	2.01	2.02
5000	8.13	8.83	2.03	2.03
10000	8.25	9.09	2.03	2.03
30000	8.48	9.07		

Table 3: Estimation of the attractor dimension of the series D and A of Santa Fe competition. The first column reports the number of points used for the estimation. The dimension values estimated for D series with GP and EP respectively are reported in second and third column. Last two columns show the the d estimated with GP and EP using the data of the A series. A series has only 10000 points.