Neural Processing Letters **14:** 27–34, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Intrinsic Dimension Estimation of Data: An Approach Based on Grassberger–Procaccia's Algorithm

FRANCESCO CAMASTRA[1,3] and ALESSANDRO VINCIARELLI[2]
[1] *Elsag spa, Via G. Puccini 2, 16154 Genova, Italy; e-mail: francesco.camastra@elsag.it*
[2] *IDIAP – Institut Dalle Molle d'Intelligence Artificielle Perceptive, Rue du Simplon 4, CP592 – 1920 Martigny, Switzerland; e-mail: alessandro.vinciarelli@idiap.ch*
[3] *Computer Science Department, University of Genoa, Via Dodecaneso 35, 16146 Genova, Italy; e-mail: camastra@disi.unige.it*

Abstract. In this paper the problem of estimating the intrinsic dimension of a data set is investigated. An approach based on the Grassberger–Procaccia's algorithm has been studied. Since this algorithm does not yield accurate measures in high-dimensional data sets, an empirical procedure has been developed. Grassberger–Procaccia's algorithm was tested on two different benchmarks and was compared to a TRN-based method.

Key words: correlation dimension, dimensionality estimation conjecture, Grassberger–Procaccia's algorithm, intrinsic dimension estimation, Multi-Layer Perceptron, Topology Representing Network

Abbreviations: BIC – Bayesian Information Criterion; DEC – Dimensionality Estimation Conjecture; GP – Grassberger–Procaccia's; ID – Intrinsic Dimension; NLPCA – Nonlinear Principal Component Analysis; MLP - Multi-Layer Perceptron; PCA – Principal Component Analysis; PP – Projection Pursuit; TRN – Topology Representing Network

1. Introduction

The intrinsic dimension (*ID*) of a data set is the minimum number of free variables needed to generate the data. More generally, a data set in d dimension is said to have an *intrinsic dimensionality* equal to M ($M < d$) if the data lie entirely within a M-dimensional subspace S [9]. ID estimation, giving a lower bound on the number of variables needed to describe a data set, is very important in pattern recognition and statistics. In particular, the knowledge of the intrinsic dimensionality is helpful for data visualization, and for classifier and regressor design in order to limit as much as possible the number of variables used to represent data sets. Furthermore, ID can suggest a value to dimension the middle layer of the autoassociative neural networks [1] often used to realize powerful feature extractors.

In spite of the importance of study of the intrinsic dimensionality of data very few results, mostly based on *projection techniques* and on the estimate of *topological dimension of data* [2], are reported in the literature. The projection techniques search the best subspace to project data, by minimizing the projection error. These methods can be divided into two families: linear and non-linear.

Linear methods such as *Principal Component Analysis* (*PCA*) and *Projection Pursuit* (*PP*) [7] are inadequate estimators of intrinsic dimension since they tend to use more dimensions than necessary [1]. For example, points lying on a curve for PCA and PP have dimension 2 instead of 1.

On the other hand, nonlinear methods such as *Nonlinear PCA (NLPCA)* [13] present some drawbacks [14]. The projections onto curves and surfaces are suboptimal; they cannot model curves or surfaces that intersect themselves (e.g. circles). Therefore NLPCA can sometimes lead to incorrect results and does not represent a reliable estimator of the intrinsic dimensionality.

Topological dimension estimators try to estimate the *topological dimension* of the data manifold. Frisone et al. [8] proposed a method based on the *Dimensionality Estimation Conjecture* (*DEC*). If the data manifold Ω is approximated by a *Topological Representing Network* (*TRN*) [16], the number n of cross-correlations learnt by each neuron is an indicator of the local dimension of Ω. The number n is quite close to the number k (*kissing number*[1]) of spheres which touch a given sphere, in the *Sphere Packing Problem* [4]. This approach presents some drawbacks. The DEC conjecture has not been proved yet; the number k is known exactly only for few dimension values. Besides, k tends to grow exponentially with the space dimension. The last peculiarity strongly limits the use of the conjecture in practical applications where data can have high dimensionality.

In this paper, an approach to the intrinsic dimensionality estimation, alternative to the above described methods, is presented. Such approach is based on the Grassberger–Procaccia's algorithm. This technique has been successfully applied to different fields of physics, such as the study of chaotic system attractors [17] and the dynamics reconstruction in natural time series [12]. The structure of the paper is as follows: in Section 2 Grassberger–Procaccia's algorithm is illustrated; the procedure for the intrinsic dimension estimation is described in Section 3; in Section 4 some experimental results are reported; in Section 5 some conclusions are drawn.

2. Grassberger–Procaccia's Algorithm

Though many definitions of the dimension of a set, based on mathematical proofs or on empirical conjectures have been proposed, a univocal definition has not been given yet.

Grassberger–Procaccia's algorithm (*GP algorithm*) allows to estimate the so-called *correlation dimension* [10]. The correlation dimension belongs to a dimension family called *fractal* [15]. Among fractal dimension definitions [5], the correlation dimension is the most popular because of its computational simplicity. Correlation dimension is defined in the following way. Let Ω be a set of points

[1] For space dimensions from 1 to 8, k is: 2, 6, 12, 24, 40, 72, 126, 240. Besides, the value k for a 24-dimension space, is also known and is 196560.

X ($X \in \mathbb{R}^n$) of cardinality N. If the *correlation integral* $C_m(r)$ is defined as:

$$C_m(r) = \lim_{N \to \infty} \frac{2}{N(N-1)} \sum_{1 \leqslant i<j \leqslant N} I(\|X_j - X_i\| \leqslant r) \tag{1}$$

where I is an *indicator function*[2], then the *correlation dimension* D of Ω is equal to:

$$D = \lim_{r \to 0} \frac{\ln(C_m(r))}{\ln(r)} \tag{2}$$

There are several methods [21, 20] in the literature to obtain an optimal estimation of the correlation dimension, but all techniques are optimal only when the correlation integral assumes a given form[3], in the other cases the estimators can perform poorly [22]. Besides, these methods generally require the use of some heuristics to fix the *radius* r [23]. Therefore, in our work, we used the original procedure proposed by Grassberger and Procaccia that consists in plotting $\ln(C_m(r))$ versus $\ln(r)$ and measuring the slope of the linear part of the curve (Figure 2).

It has been proved [6] that in order to get an accurate estimate of the dimension D, the set cardinality N has to satisfy the following inequality:

$$D < 2 \log_{10} N \tag{3}$$

The inequality (3) implies that the number of points N, necessary to get accurate estimate, grows exponentially when the dimension to estimate increases. According to a calculus of L. A. Smith [19], for the estimation of a fractal dimension D to be accurate within 5%, N should be 42^D. Since, for high dimension, N becomes huge, it is very hard to have a precise estimate of correlation dimension. In order to show the dependence of the estimated value on the number of set points, we performed the following experiments. We generated 10-dimension sets containing different number of points. Then we estimated the correlation dimension of each set with GP algorithm. The results are reported in Table 1 and show the strong dependence of the estimated value on the number of points of the set. Therefore, in order to cope with this bottleneck, an empirical procedure has been developed.

3. Intrinsic Dimension Estimation Procedure

Consider the set Ω, whose intrinsic dimension has to be estimated. The procedure consists of the following steps:

[2] $I(\lambda)$ is 1 iff condition λ holds, 0 otherwise.

[3] For example, Takens' method is optimal only iff $C_m(r) = ar^D[1 + br^2 + o(r^2)]$ where a, b are constants.

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

Points number	Estimated dimension
1000	7.83
2000	7.94
5000	8.30
10000	8.56
30000	9.11
100000	9.73

1. Another set Ω',with the same cardinality N is generated. Ω' is formed by points uniformly distributed in a d-dimensional hypercube. We assumed that the intrinsic dimensionality of Ω' is d.
2. The correlation dimension (D) of Ω' is measured by the GP algorithm.
3. Previous steps are repeated for T different values of d, obtaining a collection of measures $C = \{(d_i, D_i) : i = 1 \ldots T\}$.
4. A best-fitting of the points of C by means of a nonlinear function is performed. A plot Γ of (D) versus (d) is generated (see Figure 1).
5. Finally the correlation dimension of Ω is computed and, by using the best-fit, the intrinsic dimension of Ω is estimated.

The method previously described is based on the following assumptions:

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

In comparison with the DEC-based method, our approach presents the following advantages. First of all, the approach allows to get ID estimation of high-dimensional space, unlike the DEC-based method that estimates ID dimension up to 8. Moreover the approach proposed, since it is based on the estimation of a fractal dimension, allows to obtain non-integer values. This is preferable since, because of the noise, real data can sometimes lie within a *fractal-like* submanifold, whose dimension is usually non-integer.

4. Experimental Results

GP algorithm was applied to estimate the intrinsic dimension on two different benchmarks.

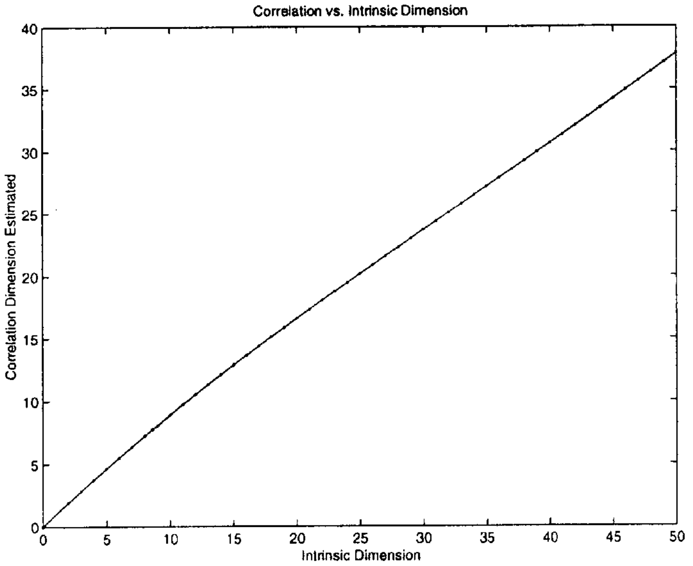

Figure 1. Correlation versus Intrinsic Dimension.

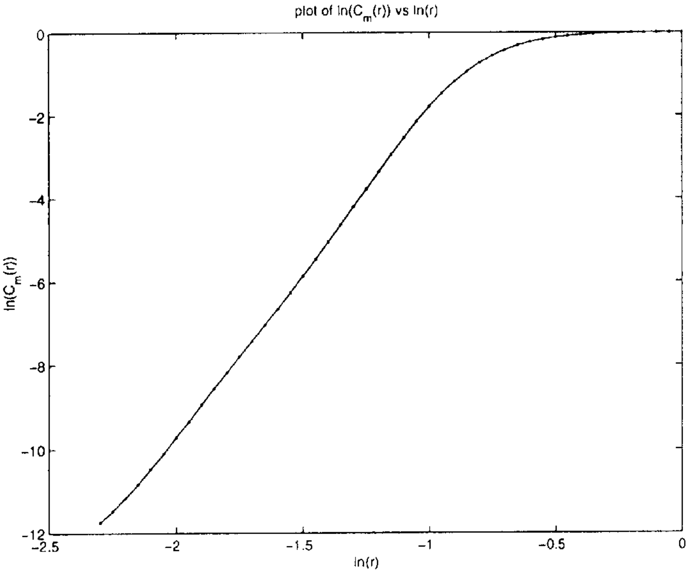

Figure 2. Plot of $\ln(C_m(r))$ versus $\ln(r)$ in the benchmark 2.

4.1. THE FIRST BENCHMARK

The first benchmark is a set formed by 30,000 points uniformly distributed in a 8-dimensional hypercube.

First of all, the plot Correlation versus Intrinsic Dimension was generated. For the setup of Γ, the 8-dimensional Ω' set has not been used. The nonlinear function used for the best-fitting was a linear combination of two logistic functions, generated by means of a MLP [1] with 1 input, 2 hidden and 1 output neurons. MLP structure was set up by means of the *bayesian information criterion* (*BIC*) [18]. The plot Correlation versus Intrinsic Dimension is shown in Figure 1.

Then GP algorithm was applied to the data set. The Correlation Dimension was set up to *7.16* and using the MLP that best-fits curve Γ in Figure 1, the intrinsic dimension was fixed to *7.90*. In comparison with the fractal dimension technique, on the same data set, the DEC-based method was tested. Many TRN maps were tried, selecting the one that minimizes the *quantization error*. The maximum cross-correlation neuron number in the map is *66*. Since the number obtained is closer to 72 than 40, that are respectively the kissing number k for the dimension 6 and 5, the intrinsic dimension, estimated by TRN, is ~ 6.

4.2. THE SECOND BENCHMARK

The second benchmark is a data set constituted by 30,000 feature vectors extracted by isolated cursive handwritten characters[4]. Each feature vector has 34 components [3]. ID of the second benchmark is unknown. GP algorithm was applied and Figure 2 shows the log-log plot obtained. The Correlation Dimension was set up to *7.76* and using Γ, the intrinsic dimension was fixed to *8.62*. TRN was also tried on the data set. The maximum cross-correlation neuron number in the case was *90*. Since the number is between 72 and 126, the intrinsics dimension, estimated by TRN, is *between 6 and 7*. As in the first benchmark, ID estimated by TRN-based method is lower than the one computed by GP.

5. Conclusions

We presented an empirical approach based on Grassberger–Procaccia's algorithm in order to estimate the intrinsic dimensionality of data. The approach was applied to two different benchmarks and the results obtained were compared to those given by a TRN-based method. In the benchmark, with known ID, GP algorithm performed better than TRN-based method. Besides, differently to TRN method our approach allows to obtain intrinsic dimension estimate also when the dimension is high. In our opinion, the method can be applied to set up the structure of a Nonlinear PCA network. In fact the middle layer of the autoassociative

[4] Characters were selected by words belonging to CEDAR database [11].

five-layer-bottleneck network, used to realize Nonlinear PCA, should have a neuron number quite close to the intrinsic dimensionality of data. Therefore, the intrinsic dimension estimate could be used to get the structure of a NLPCA network avoiding heavy experimental trials.

Acknowledgements

We wish to acknowledge S. Bengio (IDIAP), J. Luettin (IDIAP) and D. Ugolini (PNB) for commenting on the draft, and M. Spinetti (PNB) for technical support.

The present work has been supported by a grant within the Italian Minister Contract: "Dispositivi con reti neurali e sensori per riconoscimento voce e operazioni varie," given to Polo Nazionale Bioelettronica–Parco Scientifico e Tecnologico dell'Elba.

References

1. Bishop, C.: *Neural Networks for Pattern Recognition*, Cambridge University Press, 1995.
2. Bruske, J. and Sommer, G.: Intrinsic dimensionality estimation with optimally topology preserving maps, *IEEE Trans. on Patt. Anal. and Mach. Intell.* (PAMI), **20**(5) (1998) 572–575.
3. Camastra, F. and Vinciarelli, A.: Isolated cursive character recognition based on neural nets, *Kuenstliche Intelligenz*, special issue on handwriting, R. Rojas (ed.), **2** (1999) 17-19.
4. Conway, J. H. and Sloane, N. J. A.: Sphere packings, lattices and groups, *Grundlehren der mathematischen Wissenschaften 290*. Springer-Verlag, New York, 1988.
5. Eckmann, J. P. and Ruelle, D.: Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.* **57** (1985) 617–659.
6. Eckmann, J. P. and Ruelle, D.: Fundamental limitations for estimating dimensions and Lyapounov exponents in dynamical systems, *Physica*, **D56** (1992) 185–187.
7. Friedman, J. K. and Tukey, J. W.: A projection pursuit algorithm for exploratory data analysis, *IEEE Trans on Computer*, **23** (1974) 881–889.
8. Frisone, F., Firenze, F., Morasso, P. and Ricciardiello, L.: Application of topological-representing networks to the estimation of the intrinsic dimensionality of data. In: *Proceedings of ICANN'95*, October 9-13, Paris, France, 1995.
9. Fukunaga, K.: Intrinsic dimensionality extraction. In: P. R. Krishnaiah and L. N. Kanal (eds.), Classification, Pattern Recognition and Reduction of Dimensionality, *Handbook of Statistics*, Vol. 2, Amsterdam, North Holland, 1982, pp. 347–360.
10. Grassberger, P. and Procaccia, I.: Measuring the strangeness of strange attractors, *Physica*, **D9** (1983) 189–208.
11. Hull, J. J.: A database for handwritten text recognition research, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **16**(5) (1994) 550–554.
12. Isham, V.: Statistical aspects of chaos: a review. In: O. E. Barndorff-Nielsen, J. L. Jensen and W. S. Kendall (eds.), *Networks and Chaos-Statistical and Probabilistic Aspects*, Chapman Hall, London, 1993, pp. 124–200.
13. Karhunen, J. and Joutsensalo, J.: Representations and separation of signals using nonlinear PCA type learning, *Neural Networks*, **7**(1) (1994) 113–127.

14. Malthouse, E. C.: Limitations of Nonlinear PCA as performed with Generic Neural Networks. In: *Proceedings of NIPS'97 Workshop on Advances in Autoencoders/Autoassociators Based Computations*, December 5, 1997.
15. Mandelbrot, B.: *Fractals: Form, Chance and Dimension*, Freeman, San Francisco, 1977.
16. Martinetz, T. and Schulten, K.: Topology Representing Networks, *Neural Networks*, **3** (1994) 507–522.
17. Ott, E.: *Chaos in Dynamical Systems*, Cambridge University Press, 1993.
18. Schwartz, G.: Estimating the dimension of a model, *Ann. Stat.*, **6** (1978) 497–511.
19. Smith, L. A.: Intrinsic Limits on Dimension Calculations, *Phys. Lett.*, **A133** (1988) 283–288.
20. Smith, R. L.: Optimal Estimation of Fractal Dimension, In: M. Casdagli and S. Eubank (eds.), *Nonlinear Modeling and Forecasting, SFI Studies in the Science of Complexity*, Vol. XII, Addison-Wesley, 1992, pp. 115–135.
21. Takens, F.: 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, Springer-Verlag, Berlin, 1985, pp. 99–106.
22. Theiler, J.: Lacunarity in a best estimator of fractal dimension, *Phys. Lett.*, **A133** (1988) 195–200.
23. Theiler, J.: Statistical precision of dimension estimators, *Phys. Rev.*, **A41** (1990) 3038–3051.