Data Dimensionality Estimation Methods: A survey

Francesco Camastra

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

Abstract

In this paper, data dimensionality estimation methods are reviewed. The estimation of the dimensionality of a data set is a classical problem of pattern recognition. There are some good reviews [1] in literature but they do not include more recent developments based on fractal techniques and neural autoassociators. The aim of this paper is to provide an up-to-date survey of the dimensionality estimation methods of a data set, paying special attention to the fractal-based methods.

Key words: Intrinsic Dimensionality; Topological Dimension; Fukunaga-Olsen's algorithm; Fractal Dimension; MultiDimensional Scaling

1 Introduction

Pattern recognition problems deal with data represented as vectors of dimension d. The data is then embedded in $\mathbb{R}^d$, but this does not necessarily imply that its actual dimension is d. The dimensionality 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 [2], a data set $\Omega \subset \mathbb{R}^d$ is said to have *Intrinsic Dimensionality* (*ID*) equal to M if its elements lie entirely within an M-dimensional subspace of $\mathbb{R}^d$ (where $M < d$).
ID estimation 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 even 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 reduced computation time. Then it can be hard to make reliable classifiers when the dimensionality of input data is high (*curse of dimensionality* [3]). According to the statistical learning theory approach [4], the capacity and the generalization capability of the

classifiers depend on ID. Therefore the use of vectors with smaller dimension often leads to improved classification performance. Besides, when using an autoassociative neural network [5] to perform a nonlinear feature extraction (e.g. nonlinear principal component analysis), the ID can suggest a reasonable value for the number of hidden neurons. Finally, ID estimation methods are used to fix the model order in a time series, that is crucial in order to make reliable time series predictions. This paper reviews the methods to estimate ID paying special attention to the fractal-based techniques, which are generally neglected in the surveys on ID estimation.

Following the classification proposed in [1], there are two approaches for estimating ID. In the first one (*local*) ID is estimated using the information contained in sample neighborhoods, avoiding the projection of the data onto a lower-dimensional space. In the second approach (*global*), the data set is unfolded in the d-dimensional space. Unlike local approaches that use only the information contained in the neighborhood of each data sample, global approaches make use of the whole data set.

The paper is organized as follows: in Section 2 local approaches are reviewed; Section 3 present global approaches to estimate ID; Section 4 is devoted to describe specific global approaches i.e. fractal-based techniques; in Section 5 some applications are described; in Section 6 a few conclusions are drawn.

2 Local methods

Local (or *topological*) methods try to estimate the topological dimension of the data manifold. The definition of topological dimension was given by Brouwer [6] in 1913. Topological dimension is the basis dimension of the local linear approximation of the hypersurface on which the data resides, i.e. the tangent space. For example, if the data set lies on an m-dimensional submanifold, then it has an m-dimensional tangent space at every point in the set. For instance, a sphere has a two-dimensional tangent space at every point and may be viewed as a two-dimensional manifold. Since the ID of the sphere is three, the topological dimension represents a lower bound of ID. If the data does not lie on a manifold, the definition of topological dimension does not directly apply. Sometimes the topological dimension is also referred to simply as the *local dimension*. This is the reason why the methods that estimate the topological dimension are called local. The basic algorithm to estimate the topological dimension was proposed by Fukunaga and Olsen [7]. Alternative approaches to the Fukunaga-Olsen's algorithm have been proposed to estimate locally ID. Among them the *Near Neighbor Algorithm* [8] and the methods based on *Topological Representing Networks* (TRN) [9] are the most popular.

2.1 Fukunaga-Olsen's algorithm

Fukunaga-Olsen's algorithm is based on the observation that for vectors embedded in a linear subspace, the dimension is equal to the number of non-zero eigenvalues of the covariance matrix. Besides, Fukunaga and Olsen assume that the intrinsic dimensionality of a data set can be computed by dividing the data set in small regions (*Voronoi tesselation* of data space). Voronoi tesselation can be performed by means of a clustering algorithm, e.g. LBG [10]. In each region (*Voronoi set*) the surface in which the vectors lie is approximately linear and the eigenvalues of the local covariance matrix are computed. Eigenvalues are normalized by dividing them by the largest eigenvalue. The intrinsic dimensionality is defined as the number of normalized eigenvalues that are larger than a threshold T. Although Fukunaga and Olsen proposed for T, on the basis of heuristic motivations, values such as 0.05 and 0.01, it is not possible to fix a threshold value T good for every problem.

2.2 The Near Neighbor Algorithm

The first attempt to use near neighbor techniques in order to estimate ID is due to Trunk [11]. Trunk's method works as follows. An initial value of an integer parameter k is chosen and the k nearest neighbors to each pattern in the given data set are identified. The subspace spanning the vectors from the i^{th} pattern to its k nearest neighbors is constructed for all patterns. The angle between the $(k+1)^{th}$ near neighbor of pattern i and the subspace constructed for pattern i is then computed for all i. If the average of these angles is below a threshold, ID is k. Otherwise, k is incremented by 1 and the process is repeated. The weakness of Trunk's method is that it is not clear how to fix a suitable value for the threshold.

An improvement (*Near Neighbor Algorithm*) of Trunk's method was proposed by Pettis et al. [8]. Assuming that the data are locally uniformly distributed, they derive the following expression for ID:

$$ID = \frac{\langle r_k \rangle}{(\langle r_{k+1} \rangle - \langle r_k \rangle)k} \tag{1}$$

where $\langle r_k \rangle$ is the mean of the distances from each pattern to its k nearest neighbors.

The algorithm presents some problems. It is necessary to fix a suitable value for k and it is performed on a heuristic basis. Besides, Pettis et al. derived ID, using the equation (1), for the special case of three uniformly distributed one-dimensional vectors. They found $ID = 0.9$. Therefore it seems that Pettis' estimator is biased even for this simple case. Besides, Pettis described an iterative algorithm, based on an arbitrary number of neighbors, for the ID estimation. Then Verveer and Duin [12] found that Pettis' iterated algorithm yielded a uncorrect value for ID. Therefore Verveer and Duin proposed an iterative algorithm (*near neighbor estimator*) that provides a non iterative solution for ID estimation.

If $\langle r_k \rangle$ is observed for $k = k_m$ to $k = k_M$ a least square regression line can be fit to $\langle r_k \rangle$ as a function of $(\langle r_{k+1} \rangle - \langle r_k \rangle)k$. Verveer and Duin obtained the following estimation for ID:

$$ID = \left[\sum_{k=k_m}^{k_M - 1} \frac{(\langle r_{k+1} \rangle - \langle r_k \rangle)\langle r_k \rangle}{k} \right] \left[\sum_{k=k_m}^{k_M - 1} (\langle r_{k+1} \rangle - \langle r_k \rangle)^2 \right]^{-1} \qquad (2)$$

Since the estimate yielded by Verveer-Duin's algorithm is generally not an integer it has to be rounded to the nearest integer. Since the vectors are usually locally uniformly distributed, Verveer and Duin advise that the values $k = k_m$ and $k = k_M$ should be small as possible. When the data is very noisy Verveer and Duin suggest to ignore the first nearest neighbor i.e it should be $k_m > 1$.

As a general comment it is necessary to remark that both Pettis' and Verver-Duin's algorithms are sensitive to outliers. The presence of outliers tends to significantly affect the ID estimate. Another problem is the influence of the *edge effect*. Data close to the cluster edge are not uniformly distributed. Therefore if the percentage of this data, on the whole data set, is not negligible, ID estimate is distorted. This happens when the dimensionality of the data set is high and the data density is low.

2.3 TRN-based methods

Topology Representing Network (TRN) is a unsupervised neural network proposed by Martinetz and Schulten [9]. They proved that TRN are optimal topology preserving maps i.e TRN preserves in the map the topology originally present in the data.

Bruske and Sommer [13] proposed to improve Fukunaga-Olsen's algorithm using TRN in order to perform the Voronoi tesselation of the data space. In detail, the algorithm proposed by Bruske and Sommer is the following. An optimal topology preserving map G, by means of a TRN, is computed. Then, for each neuron $i \in G$, a PCA is performed on the set Q_i consisting of the differences between the neuron i and all of its m_i closest neurons in G. Bruske-Sommer's algorithm shares with Fukunaga-Olsen's one the same limitations: since none of the eigenvalues of the covariance matrix will be null due to noise, it is necessary to use heuristic thresholds in order to decide whether an eigenvalue is significant or not.

Finally Frisone et al. [14] use Topology Representing Networks to get directly an ID estimate. If the data manifold Ω is approximated by a TRN, the number n of cross-correlations learnt by each neuron of the TRN is an indicator of the local dimension of the data set Ω. Frisone et al. conjectured that the number n is close to the number k of spheres that touch a given sphere, in the *Sphere Packing Problem* (*SPP*) [15]. For space dimensions from 1 to 8, k is: 2, 6, 12, 24, 40, 72, 126, 240. Besides, SPP has only been solved for a 24-dimensional space and k is 196560. Hence it is adequate to measure k to infer ID. Frisone's approach presents some drawbacks: the conjecture has not been proved yet, the number k is known exactly only for few dimension values and 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.

3 Global Methods

Global methods try to estimate the ID of a data set, unfolding the whole data set in the d-dimensional space. Unlike local methods that use only the information contained in the neighborhood of each data sample, global methods make use of the whole data set.

Global methods can be grouped in three big families: *Projection techniques*, *Multidimensional Scaling Methods* and *Fractal-Based Methods*.

3.1 Projection techniques

Projection techniques search for the best subspace to project the data by minimizing the projection error. These methods can be divided into two families: linear and nonlinear.

Principal Component Analysis (*PCA*) [5,16] is a widely used linear method. PCA projects the data along the directions of maximal variance. The method consists of computing eigenvalues and eigenvectors of the covariance matrix of data. Each of the eigenvectors is called a *principal component*. ID is given by the number of the non-null eigenvalues. The method presents some problems. PCA is an inadequate estimator, since it tends to overestimate the ID [17]. As shown in Figure 1, a data set formed by points lying on a circumference for PCA has dimension 2 rather than 1.

In order to cope with these problems, nonlinear algorithms have been proposed to get nonlinear PCA. There are two different possible approaches to get a nonlinear PCA: an autoassociative approach (*Nonlinear PCA*) [5,18] and the one based on the use of Mercer kernels (*Kernel PCA*) [19].

Nonlinear PCA is performed by means of a five-layers neural network. The neural net has a typical bottleneck structure, shown in Figure 2. The first (*in-*

put) and the last (*output*) layer have the same number of neurons, while the remaining hidden layers have less neuron than the first and the last ones. The second, the third and the fourth layer are called respectively *mapping*, *bottleneck* and *demapping* layer. Mapping and demapping layers have usually the same number of neurons. The number of the neurons of the bottleneck layer provides an ID estimate. The targets used to train Nonlinear PCA are simply the input vector themselves. The network is trained with the backpropagation algorithm, minimizing the square error. As optimization algorithm, the *conjugate-gradient algorithm* [20] is generally used. Though nonlinear PCA performs better than linear PCA in some contexts [21], it presents drawbacks when estimating ID. As underlined by Malthouse [22], the projections onto curves and surfaces are suboptimal. Besides, NLPCA cannot model curves or surfaces that intersect themselves. Kernel PCA consists of making a nonlinear projection of the data set, by means of an appropriate *positive definite function* (*Mercer kernel*) [23] in a new space (*Feature Space*). Then the eigenvalues of the covariance matrix in the Feature Space are computed and ID is given by the number of the non-null eigenvalues. The method presents some problems. The performance of the method is heavily influenced by the kernel choice [24]. Moreover, due to the data noise, last eigenvalues, even if very small, are not null. Therefore it is necessary to ignore the eigenvalues whose magnitude is lower than a threshold value that can be only fixed in a heuristic way. Among projection techniques it is worth mentioning the *Whitney reduction network* recently proposed by Broomhead and Kirby [5,25]. This method is based on Whitney's concept of *good projection* [26], namely a projection obtained by means of an injective mapping. An injective mapping between two sets U and V is a mapping that associate a unique element of V to each element of U. As pointed out in [5], finding projections, by means of injective mappings, can be difficult and can sometimes involve empirical considerations.

3.2 Multidimensional Scaling Methods

Multidimensional Scaling (MDS) [27,28] methods are projection techniques that tend to preserve, as much as possible, the distances among data. Therefore data that are close in the original data set should be projected in such a way that their projections, in the new space (*output space*), are still close. Among multidimensional scaling algorithms, the best known example is *MDSCAL*, by Kruskal [29] and Shepard [30]. The criterion for the goodness of the projection used by MDSCAL is the *stress*. This depends only on the distances between data. When the rank order of the distances in the output space is the same as the rank order of the distances in the original data space, stress is zero.

Kruskal's stress S_K is:

$$S_K = \left[\frac{\sum_{i<j} \left[rank(d(x_i, x_j)) - rank(D(x_i, x_j))\right]^2}{\sum_{i<j} rank(d(x_i, x_j))^2} \right]^{\frac{1}{2}} \tag{3}$$

where $d(x_i, x_j)$ is the distance between the data x_i and x_j and the $D(x_i, x_j)$ is the distance of the projections of the same data in the output space. When the stress is zero a perfect projection exists. Stress is minimized by iteratively moving the data in the output space from their initially randomly chosen positions according to a gradient-descent algorithm. The intrinsic dimensionality is determined in the following way. The minimum stress for projections of different dimensionalities is computed. Then a plot of the minimum stress versus dimensionality of the output space is performed. ID is the dimensionality value for which there is a knee or a flattening of the curve. Kruskal and Shepard's algorithm presents a main drawback. The knee or the flattening of the curve could not exists. MDS approaches close to Kruskal and Shepard's one are the *Bennett's algorithm* [31] and the *Sammon's mapping* [32]

3.2.1 Bennett's algorithm

Bennett's algorithm is the based on the assumption that data are uniformly distributed inside a sphere of radius r in an L-dimensional space. Let X_1 and X_2 be random variables representing points in the sphere and R_L be the normalized Euclidean distance (the *interpoint distance*) between them. If

$$R_L = \frac{|X_1 - X_2|}{2r} \tag{4}$$

then the variance of R_L is a decreasing function of L, which may be expressed as:

$$L \;\; var(R_L) \approx constant \tag{5}$$

where $var(R_L)$ is the variance of R_L. Therefore increasing the variance of the interpoint distances has the effect of decreasing the dimensionality of the representation, i.e. it "*flattens*" the data set.

Bennett's algorithm involves two stages. The first stage moves the patterns, in the original input space, in order to increase the variance of the interpoint distances. The second stage adjusts the position of the patterns in order to make the rank orders of interpoint distances in local regions are the same. These steps are repeated until the variance of the interpoint distances levels off. Finally the covariance matrix of the whole data set, yielded by the previous stages, is computed. The ID is determined by the number of significant eigenvalues of the covariance matrix.

Bennett's algorithm presents some drawbacks. First of all, as in Fukunaga-Olsen's algorithm, in order to decide if an eigenvalue is significant, it is necessary to fix heuristically a threshold value. Besides, as underlined previously in the PCA description, this method tends to overestimate the dimensionality of a data set.

Chen and Andrews [33] proposed to improve Bennett's algorithm by introducing a cost function to make Bennett's rank-order criterion more sensitive to local data regions. The basic idea is still to mantain rank order of local distances in the two spaces.

3.2.2 Sammon's mapping

Sammon proposed to minimize a stress measure similar to Kruskal's one. The stress S_S proposed by Sammon has the following expression:

$$S_S = \left[\sum_{i<j} \frac{(d(x_i, x_j) - D(x_i, x_j))^2}{d(x_i, x_j)} \right] \left[\sum_{i<j} d(x_i, x_j) \right]^{-1} \tag{6}$$

Where $d(x_i, x_j)$ is the distance between patterns x_i and x_j in the original data space and $D(x_i, x_j)$ is the distance in the two- or three- dimensional output space. The stress is minimized by the gradient-descent algorithm.

Kruskal [34] demonstrated how a data projection very similar to Sammon's mapping could be generated from MDSCAL. An improvement to Kruskal's and Sammon's methods has been proposed by Chang and Lee [35]. Unlike Sammon and Kruskal who move all points simultaneously in the output space to minimize the stress, Chang and Lee have suggested to minimize the stress by moving the points two at a time. In this way, it tries to preserve local structure while minimizing the stress. The method requires heavy computational resources even when the cardinality of the data set is moderate. Besides, the results are influenced by the order in which the points are coupled.

Several other approaches to ID estimation have been proposed. It is worth mentioning Shepard and Carroll's *index of continuity* [36], Kruskal's *indices of condensation* [37] and Kruskal and Carroll's parametric mapping [38]. Surveys of the classical Multidimensional Scaling methods can be found in [27,28,39]. Recently local versions of MDS methods, i.e. ISOMAP algorithm [40] and Local Linear Embedding [41], have been proposed. We do not describe these methods for the sake of the brevity.

Finally it is worth mentioning the *Curvilinear Component Analysis* (*CCA*) proposed by Demartines and Herault [42]. The principle of CCA is a *self-organizing* neural network performing two tasks: vector quantization of the data set, whose dimensionality is n, and a nonlinear projection of these quantizing vectors onto a space of dimensionality p ($p < n$). The first task is performed by means of *SOM* [43]. The second task is performed by means of a technique very similar to MDS methods previously described. Since a MDS that preserve all distances is not possible, a cost function E measures the goodness of the projection. The cost function E is the following:

$$E = \frac{1}{2} \sum_{i} \sum_{j \neq i} (d(x_i, x_j) - D(x_i, x_j))^2 F(D(x_i, x_j), \lambda) \tag{7}$$

where $d(x_j, x_j)$ are Euclidean distances between the points x_i and x_j of data space and $D(x_i, x_j)$ are Euclidean distances between the projections of the points in the output space; λ is a set of parameters to set up and $F(\cdot)$ is a function (e.g. a decreasing exponential or a sigmoid) to be chosen in an opportune way.
CCA seems to have very close performances to Shepard's MDS based on index of continuity [42].

4 Fractal-Based Methods

Fractal-based techniques are global methods that have been successfully applied to estimate the attractor dimension of the underlying dynamic system generating time series [44]. Unless other global methods, they can provide as ID estimation a non-integer value. Since fractals are generally [1] characterized by a non-integer dimensionality, for instance the dimension of Cantor's set and Koch's curve [45] is respectively $\frac{\ln 2}{\ln 3}$ and $\frac{\ln 4}{\ln 3}$, these methods are called *fractal*. In nonlinear dynamics many definitions of *fractal* dimensions [46] have been proposed. The *Box-Counting* and the *Correlation* dimension are the most popular.

4.1 Box-Counting Dimension

The first definition of dimension (*Hausdorff dimension*) [46,47] is due to Hausdorff [48]. The *Hausdorff dimension* D_H of a set Ω is defined by introducing the quantity

$$\Gamma_H^d(r) = \inf_{s_i} \sum_i (r_i)^d \tag{8}$$

[1] Fractals have not always non-integer dimensionality. For instance, the dimension of *Peano's curve* is *2*.

where the set Ω is covered by cells s_i with variable diameter r_i, and all diameters satisfy $r_i < r$.
That is, we look for that collection of covering sets s_i with diameters less than or equal to r which minimizes the sum in (8) and we denote that minimized sum $\Gamma_H^d(r)$. The *d-dimensional Hausdorff measure* is then defined as

$$\Gamma_H^d = \lim_{r \to 0} \Gamma_H^d(r) \tag{9}$$

The d-dimensional Hausdorff measure generalizes the usual notion of the total length, area and volume of simple sets. Haussdorf proved that Γ_H^d, for every set Ω, is $+\infty$ if d is less than some critical value D_H and is 0 if d is greater than D_H. The critical value D_H is called the *Hausdorff dimension* of the set. Since the Hausdorff dimension is not easy to evaluate, in practical application it is replaced by an upper bound that differs only in some constructed examples: the *Box-Counting dimension* (or *Kolmogorov capacity*) [47].
The Box-Counting dimension D_B 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

$$D_B = \lim_{r \to 0} \frac{\ln(\nu(r))}{\ln(\frac{1}{r})} \tag{10}$$

It can show that if in the definition of Hausdorff dimension the cells have the same diameter r, Hausdorff dimension reduces to Box-Counting dimension. Although efficient algorithms [49],[50], [51] have been proposed, the Box-Counting dimension can be computed only for low-dimensional sets because the algorithmic complexity grows exponentially with the set dimensionality.

4.2 Correlation Dimension

A good substitute for the Box-Counting dimension can be the *Correlation dimension* [52]. Due to its computational simplicity, the Correlation dimension is successfully used to estimate the dimension of attractors of dynamical systems.
The Correlation dimension is defined as follows:
let $\Omega = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ be a set of points in $\mathbb{R}^n$ of cardinality N. If the *correlation integral* $C_m(r)$ is defined as:

$$C_m(r) = \lim_{N \to \infty} \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(\|\mathbf{x}_j - \mathbf{x}_i\| \leq r) \tag{11}$$

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

$$D = \lim_{r \to 0} \frac{\ln(C_m(r))}{\ln(r)} \tag{12}$$

Correlation and Box-Counting dimension are strictly related. It can be shown that both dimensions are special cases of the *generalized Renyi dimension.*
If the *generalized correlation integral* C_p is:

$$C_p(r) = \frac{1}{N(N-1)^{p-1}} \sum_{i=1}^{N} \left(\sum_{j \neq i}^{N} I(\|\mathbf{x}_j - \mathbf{x}_i\| \leq r) \right)^{p-1} \tag{13}$$

The generalized Renyi dimension D_p is defined in the following way:

$$D_p = \lim_{r \to 0} \frac{1}{p-1} \frac{\ln(C_p(r))}{\ln(r)} \tag{14}$$

It can be shown [52] that for $p = 0$ and $p = 2$ D_p reduces respectively to the Box-Counting and the Correlation dimension. Besides, it can be proved that Correlation Dimension is a lower bound of the Box-Counting Dimension. Nevertheless, due to noise, the difference between the two dimensions is negligible in applications with real data.

4.3 Methods of Estimation of Fractal Dimension

The most popular method to estimate Box-Counting and Correlation dimension is the log-log plot. This method consists in plotting $\ln(C_m(r))$ versus $\ln(r)$. The Correlation dimension is the slope of the linear part of the curve (Figure 3). The method to estimate Box-Counting is analogous, but $\ln(\nu(r))$ replaces $\ln(C_m(r))$.
The methods to estimate Correlation and Box-Counting dimension present some drawbacks. Though Correlation and Box-Counting dimension are asymptotic results and hold only for $r \to 0$; r cannot be too small since too few observations cannot allow to get reliable dimension estimates. In fact the noise has most influence at small distance. Therefore there is a trade-off between taking r small enough to avoid non-linear effects and taking r sufficiently large to reduce statistical errors due to lack of data. The use of least-square method makes the dimension estimate not adequately robust towards the outliers. Moreover, log-log plot method does not allow to compute the error in dimension estimation.

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

Some methods [53],[54],[55] have been studied to obtain an optimal estimate for the correlation dimension. Takens [55] has proposed a method, based on *Fisher's method of Maximum Likelihood* [56,57], that allows to estimate the correlation dimension with a standard error.

4.3.1 Takens' method

Let Q be the following set $Q = \{q_k \mid q_k < r\}$ where r_k is the the Euclidean distance between a generic couple of points of Ω and r (*cut-off radius*) is a real positive number.

Using the Maximum Likelihood principle it can prove that the expectation value of the Correlation Dimension $\langle D_c \rangle$ is:

$$\langle D_c \rangle = - \left(\frac{1}{|Q|} \sum_{k=1}^{|Q|} q_k \right)^{-1} \tag{15}$$

where $|Q|$ stands for the cardinality of Q.

Takens' method presents some drawbacks. It requires some heuristics to set the radius [58]. Besides, the method is optimal only if the correlation integral $C_m(r)$ assumes the form $C_m(r) = ar^D[1 + br^2 + o(r^2)]$ where a and b are constants. Otherwise Takens' estimator can perform poorly [59].

4.4 Limitations of Fractal Methods

In addition to the drawbacks previously exposed, estimation methods based on fractal techniques have a fundamental limitation.

It has been proved [60,61] that in order to get an accurate estimate of the dimension D, the set cardinality N has to satisfy the following inequality:

$$D < 2 \log_{10} N \tag{16}$$

Inequality (16) 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.

In order to cope with this problem and to improve the reliability of the measure for low values of N, *the method of surrogate data* [62] has been proposed. The method of surrogate data is an application of a well-know statistic technique called *bootstrap* [63]. Given a data set Ω, the method of surrogate data consists of creating a new synthetic data set Ω', with greater cardinality, that has the same statistical properties of Ω, namely the same mean, variance and Fourier Spectrum. Although the cardinality of Ω' can be chosen arbitarily, the method of surrogate data cannot be used when the dimensionality of the data set is high. As pointed out previously, a data set whose dimension is *18* requires at least, on the base of (16), a data set with 10^9 points. Therefore the method of surrogate data becomes computationally burdensome.

Finally heuristic methods [64,65] have been proposed in order to estimate how fractal techniques underestimate the dimensionality of a data set when its cardinality is unadequate. These heuristic methods permit inferring the actual dimensionality of the data set. Since the methods are not theoretically well-grounded they have to be used with prudence.

5 Applications

As mentioned before, estimation methods of the dimensionality of data sets are useful in pattern recognition to develop powerful feature extractors. For instance, when using an autoassociative neural network to perform a nonlinear feature extraction (e.g. nonlinear principal component analysis), the ID can suggest a reasonable value for the number of hidden neurons. Besides, ID has been used as feature for the characterization of human faces [66] and, in general, some authors [67,68] have measured the fractal dimension of an image with the aim to establish if the dimensionality was a distinctive feature of the image itself.

The estimate of the dimensionality of a data set is crucial in the analysis of signals and time series. For instance, ID estimation is fundamental in the study of chaotic systems (e.g. Hénon map, Rössler oscillator) [47], in the analysis of ecological time series (e.g. Canadian lynx population) [69], in biomedical signal analysis [70,71], in radar clutter identification [72], in speech analysis [73], and in the prediction of financial time series [74].

Finally ID estimation methods are used to fix the model order in time series. This is fundamental to make reliable time series predictions. In order to understand the importance of the knowledge of ID, we consider a time series $x(t)$, with $(t = 1, 2, \ldots, N)$. It can be described by the equation:

$$x(t) = f(x(t-1), x(t-2), \ldots, x(t-(d-2)), x(t-(d-1))) + \epsilon_t \quad (17)$$

The term ϵ_t represents an indeterminable part originated either from unmodelled dynamics of the process or from real noise. The function $f(\cdot)$ is the *skeleton* of the time series [75,76]. If $f(\cdot)$ is linear, we have an *autoregressive model of order* (*AR(d-1)*), otherwise a *nonlinear autoregressive model of order* (*NAR(d-1)*).

The key problem in the autoregressive models is to fix the model order *(d-1)*. Fractal-based techniques can be used to fix the order in a time series. In particular, these techniques can be used for the *model reconstruction* (or *re-*

construction of dynamics) of the time series. This is performed by the method of delays [77].
The time series in the equation (17) is represented as a series of a set of points $\{X(t) : X(t) = [x(t), x(t-1), \dots, x(t-d+1)]\}$ in a d-dimensional space. 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 [3].
Takens-Mañé embedding theorem [78,79] states that, in order to obtain a faithful reconstruction of the system dynamics, it must be:

$$2S + 1 \leq d \tag{18}$$

where S is the dimension of the system attractor and d is called the *Embedding dimension* of the system.
Therefore, it is adequate to measure S to infer the Embedding dimension d and the order of the time series $d-1$. The estimation of S can be performed by means of fractal techniques (e.g Box-Counting and Correlation dimension estimation) previously described in Section 4.
There are many applications of fractal techniques to fix the model order to natural time series: in the economic field [74], in engineering [80], in the analysis of electroencephalogram data [81], in metereology [82,83], in the analysis of astronomical data [84] and many others. A good review about these application can be found in [69].

6 Conclusions

In this paper, the data set dimensionality estimation methods have been reviewed. The estimation of the dimensionality of a data set is a classical problem of pattern recognition. Recently the use of fractal-based techniques and neuroassociators seems to get new force to the research on reliable estimation methods of data set dimensionality. The aim of this paper has been to provide a survey of the estimation methods, focusing on the methods based on fractal techniques and neural autoassociators. In spite of fractal techniques have been successfully applied to estimate the dimensionality of a data set, they seem to fail dramatically when, at the same time, the cardinality of the data set is low and the dimensionality is high. To get reliable estimators, when the dimensionality is high and the set cardinality is low, still remains an open problem.

Acknowledgements First the author wishes to thank the anonymous reviewer for his valuable comments. The author thanks M. Kanevski and A.

[3] M is *diffeomorphic* to U iff there is a differentiable map $m : M \mapsto U$ whose inverse m^{-1} exists and is also differentiable.

Vinciarelli for comments and useful discussions. The work is dedicated to my parents, Attilio and Antonia Nicoletta.

References

[1] A. K. Jain, R. C. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.

[2] K. Fukunaga, Intrinsic dimensionality extraction, in: Classification, Pattern Recognition and Reduction of Dimensionality, Vol. 2 of Handbook of Statistics, North Holland, 1982, pp. 347–362.

[3] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961.

[4] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.

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

[6] A. Heyting, H. Freudenthal, Collected Works of L.E.J Brouwer, North Holland Elsevier, 1975.

[7] K. Fukunaga, D. R. Olsen, An algorithm for finding intrinsic dimensionality of data, IEEE Transactions on Computers 20 (2) (1976) 165–171.

[8] K. Pettis, T. Bailey, T. Jain, R. Dubes, An intrinsic dimensionality estimator from near-neighbor information, IEEE Transaction on Pattern Analysis and Machine Intelligence 1 (1) (1979) 25–37.

[9] T. Martinetz, K. Schulten, Topology representing networks, Neural Networks 3 (1994) 507–522.

[10] Y. Linde, A. Buzo, R. Gray, An algorithm for vector quantizer design, IEEE Transaction on Communications 28 (1) (1980) 84–95.

[11] G. V. Trunk, Statistical estimation of the intrinsic dimensionality of a noisy signal collection, IEEE Transaction on Computers 25 (1976) 165–171.

[12] P. J. Verveer, R. Duin, An evaluation of intrinsic dimensionality estimators, IEEE Transaction on Pattern Analysis and Machine Intelligence 17 (1) (1995) 81–86.

[13] J. Bruske, G. Sommer, Intrinsic dimensionality estimation with optimally topology preserving maps, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (5) (1998) 572–575.

[14] F. Frisone, F. Firenze, P. Morasso, L. Ricciardiello, Application of topological-representing networks to the estimation of the intrinsic dimensionality of data, in: Proceedings of International Conference on Artificial Neural Networks, 1995, pp. 323–329.

[15] J. H. Conway, N. J. A. Sloane, Sphere Packings, Lattices and Groups, Springer-Verlag, 1988.

[16] I. T. Jollife, Principal Component Analysis, Springer-Verlag, 1986.

[17] C. Bishop, Neural Networks for Pattern Recognition, Cambridge University Press, 1995.

[18] J. Karhunen, J. Joutsensalo, Representations and separation of signals using nonlinear pca type learning, Neural Networks 7 (1) (1994) 113–127.

[19] B. Schölkopf, A. Smola, K. R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319.

[20] W. H. Press, B. P. Flannery, S. A. Teulkosky, W. T. Vetterling, Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, 1989.

[21] D. Fotheringhame, R. J. Baddeley, Nonlinear principal component analysis of neuronal spike train data, Biological Cybernetics 77 (1997) 282–288.

[22] E. C. Malthouse, Limitations of nonlinear pca as performed with generic neural networks, IEEE Transaction on Neural Networks 9 (1) (1998) 165–173.

[23] C. Berg, J. P. R. Christensen, P. Ressel, Harmonic analysis on semigroups, Springer-Verlag, 1984.

[24] F. Camastra, Kernel methods for unsupervised learning, Phd Thesis Progress Report, University of Genova, 2002 .

[25] D. S. Broomhead, M. Kirby, A new approach to dimensionality reduction: Theory and algorithms, SIAM Journal of Applied Mathematics 60 (6) (2000) 2114–2142.

[26] H. Whitney, Differentiable manifolds, Annals of Math. 37 (1936) 645–680.

[27] A. K. Romney, R. N. Shepard, S. B. Nerlove, Multidimensionaling Scaling, vol. I, Theory, Seminar Press, 1972.

[28] A. K. Romney, R. N. Shepard, S. B. Nerlove, Multidimensionaling Scaling, vol. 2, Applications, Seminar Press, 1972.

[29] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–27.

[30] R. N. Shepard, The analysis of proximities: Multimensional scaling with an unknown distance function, Psychometrika 27 (1962) 125–140.

[31] R. S. Bennett, The intrinsic dimensionality of signal collections, IEEE Transactions on Information Theory 15 (1969) 517–525.

[32] J. W. J. Sammon, A nonlinear mapping for data structure analysis, IEEE Transaction on Computers C-18 (1969) 401–409.

[33] C. K. Chen, H. C. Andrews, Nonlinear intrinsic dimensionality computations, IEEE Transactions on System Man and Cybernetics SMC-3 (1973) 197–200.

[34] J. B. Kruskal, A nonlinear mapping for data structure analysis, IEEE Transaction on Computers C-20 (1971) 1614.

[35] C. L. Chang, R. C. T. Lee, A heuristic relaxation method for nonlinear mapping in cluster analysis, IEEE Transactions on Computers C-23 (1974) 178–184.

[36] R. N. Shepard, J. D. Carroll, Parametric representation of nonlinear data structures, in: Multivariate Analysis, Academic Press, 1969, pp. 561–592.

[37] J. B. Kruskal, Linear transformation of multivariate data to reveal clustering, in: Multidimensional Scaling, vol. I, Academic Press, 1972, pp. 101–115.

[38] J. B. Kruskal, J. D. Carroll, Geometrical models and badness-of-fit functions, in: Multivariate Analisys, vol. 2, Academic Press, 1969, pp. 639–671.

[39] R. N. Shepard, Representation of structure in similarity data problems and prospect, Psychometrika 39 (1974) 373–421.

[40] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (12) (2000) 2319–2323.

[41] S. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (12) (2000) 2323–2326.

[42] P. Demartines, J. Herault, Curvilinear component analysis: A self-organizing neural network for nonlinear mapping in cluster analysis, IEEE Transactions on Neural Networks 8 (1) (1997) 148–154.

[43] T. Kohonen, Self-Organizing Map, Springer-Verlag, 1995.

[44] D. Kaplan, L. Glass, Understanding Nonlinear Dynamics, Springer-Verlag, 1995.

[45] B. Mandelbrot, Fractals: Form, Chance and Dimension, Freeman, 1977.

[46] J. P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, Review of Modern Physics 57 (1985) 617–659.

[47] E. Ott, Chaos in Dynamical Systems, Cambridge University Press, 1993.

[48] F. Hausdorff, Dimension und äusseres mass, Math. Annalen 79 (157).

[49] P. Grassberger, An optimized box-assisted algorithm for fractal dimension, Physics Letters A148 (1990) 63–68.

[50] C. R. Tolle, T. R. Mc Junkin, D. J. Gorisch, Suboptimal minimum cluster volume cover-based method for measuring fractal dimension, IEEE Transaction on Pattern Analysis and Machine Intelligence 25 (1) (2003) 32–41.

[51] B. Kégl, Intrinsic dimension estimation using packing numbers, in: Advances in Neural Information Processing 15, MIT Press, 2003.

[52] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D9 (1983) 189–208.

[53] D. S. Broomhead, R. Jones, Time series analysis, Proc. R. Soc. Lond. A423 (1989) 103–121.

[54] R. L. Smith, Optimal estimation of fractal dimension, in: Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity vol. XII, Addison-Wesley, 1992, pp. 115–135.

[55] F. Takens, On the numerical determination of the dimension of an attractor, in: Dynamical Systems and Bifurcations, Proceedings Groningen 1984, Springer-Verlag, 1984, pp. 99–106.

[56] R. O. Duda, P. E. Hart, Pattern Classification and Scene Analysis, John Wiley, 1973.

[57] K. Fukunaga, An Introduction to Statistical Pattern Recognition, Academy Press, 1990.

[58] J. Theiler, Statistical precision of dimension estimators, Physical Review A41 (1990) 3038–3051.

[59] J. Theiler, Lacunarity in a best estimator of fractal dimension, Physics Letters A133 (1988) 195–200.

[60] J. P. Eckmann, D. Ruelle, Fundamental limitations for estimating dimensions and lyapounov exponents in dynamical systems, Physica D-56 (1992) 185–187.

[61] L. A. Smith, Intrinsic limits on dimension calculations, Physics Letters A133 (1988) 283–288.

[62] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J. D. Farmer, Testing for nonlinearity in time series: the method for surrogate date, Physica D58 (1992) 77–94.

[63] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, 1993.

[64] F. Camastra, A. Vinciarelli, Intrinsic estimation of data: An approach based on grassberger-procaccia's algorithm, Neural Processing Letters 14 (1) (2001) 27–34.

[65] F. Camastra, A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (10) (2002) 1404–1407.

[66] M. Kirby, L. Sirovich, Application of the karhunen-loève procedure for the characterization of human faces, IEEE Transaction on Pattern Analysis and Machine Intelligence 12 (1) (1990) 103–108.

[67] Q. Huang, J. R. Lorch, R. C. Dubes, Can the fractal dimension of images be measured?, Pattern Recognition 27 (3) (1994) 339–349.

[68] L. V. Meisel, M. A. Johnson, Convergence of numerical box-counting and correlation integral multifractal analysis techniques, Pattern Recognition 30 (9) (1997) 1565–1570.

[69] V. Isham, Statistical aspects of chaos: a review, in: Networks and Chaos-Statistical and Probabilistic Aspects, Chapman and Hall, 1993, pp. 124–200.

[70] D. R. Chialvo, R. F. Gilmour, J. Jalife, Low-dimensional chaos in cardiac tissue, Nature 343 (1990) 653–658.

[71] W. S. Tirsch, M. Keidel, S. Perz, H. Scherb, G. Sommer, Inverse covariation of spectral density and correlation dimension in cyclic eeg dimension of the human brain, Biological Cybernetics 82 (2000) 1–14.

[72] S. Haykin, X. Bo Li, Detection of signals in chaos, Proceedings of IEEE 83 (1) (1995) 95–122.

[73] P. Somervuo, Speech dimensionality analysis on hypercubical self-organizing maps, Neural Processing Letters to appear.

[74] J. A. Scheinkman, B. Le Baron, Nonlinear dynamics and stock returns, Journal of Businness 62 (1989) 311–337.

[75] H. Kanz, T. Schreiber, Nonlinear Time Series Analysis, Cambridge University Press, 1997.

[76] H. Tong, Nonlinear Time Series, Oxford University Press, 1990.

[77] N. Packard, J. Crutchfield, J. Farmer, R. Shaw, Geometry from a time series, Physical Review Letters 45 (1) (1980) 712–716.

[78] F. Takens, Detecting strange attractor in turbolence, in: Dynamical Systems and Turbolence, Warwick 1980, Springer-Verlag, 1981, pp. 366–381.

[79] R. Mañé, On the dimension of compact invariant sets of certain nonlinear maps, in: Dynamical Systems and Turbolence, Warwick 1980, Springer-Verlag, 1981, pp. 230–242.

[80] F. Camastra, A. M. Colla, Neural short-term prediction based on dynamics reconstruction, Neural Processing Letters 9 (1) (1999) 45–52.

[81] I. Dvorak, A. Holden, Mathematical Approaches to Brain Functioning Diagnostics, Manchester University Press, 1991.

[82] E. N. Lorenz, Deterministic non-periodic flow, Journal of Atmospheric Science 20 (1963) 130–141.

[83] J. Houghton, The bakerian lecture 1991: the predictability of weather and climates, Phil. Trans. R. Soc. Lond. A337 (1991) 521–572.

[84] J. Scargle, Studies in astronomical time series analysis. *iv*. modeling chaotic and random processes with linear filters, Astrophysical Journal 359 (1990) 469–482.

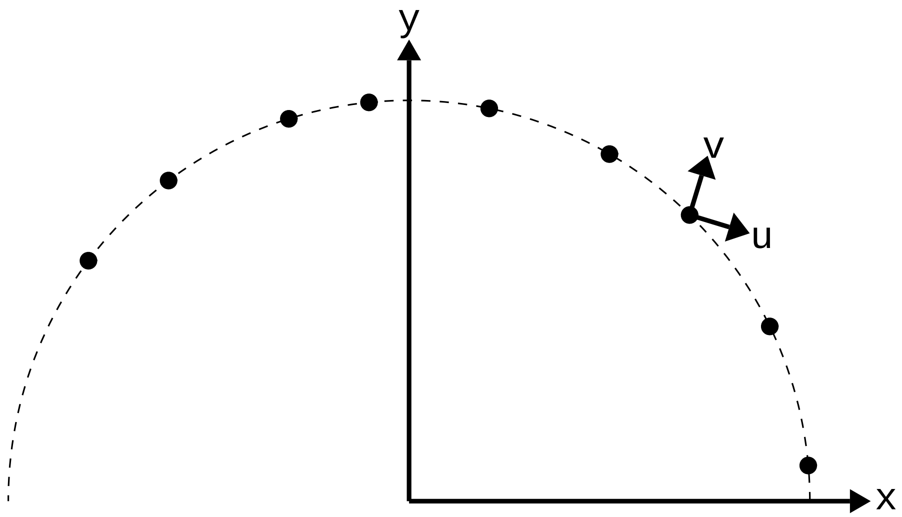

Fig. 1. Ω Data Set. The data set is formed by points lying on the upper semicirconference of equation $x^2 + y^2 = 1$. The ID of Ω is *1*. Neverthless PCA yields *two* non-null eigenvalues. The principal components are indicated by u and v.

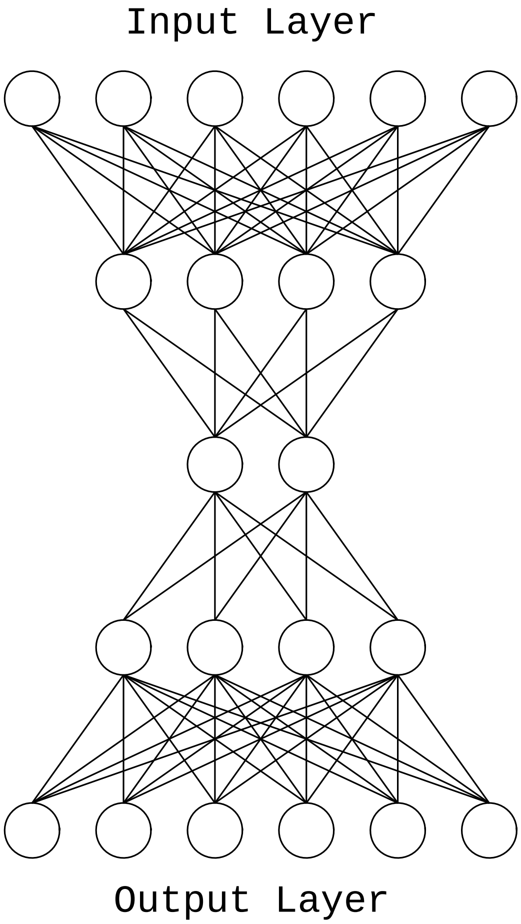

Fig. 2. A Neural Net for Nonlinear PCA

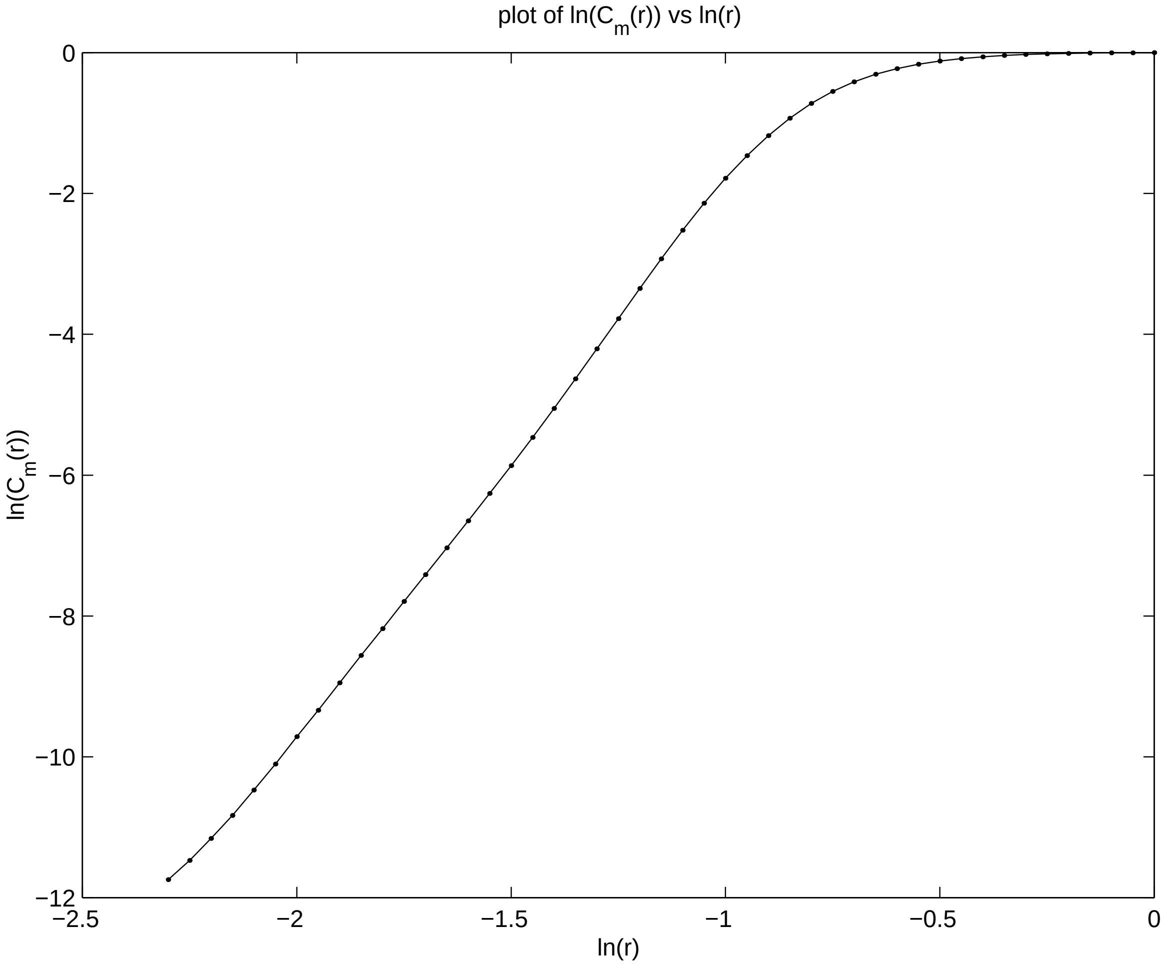

Fig. 3. Plot of $\ln(C_m(r))$ vs $\ln(r)$.