Data Dimensionality Estimation: Achievements and Challanges

Francesco Camastra(✉)

Department of Science and Technology, University of Naples Parthenope,
Centro Direzionale Isola C4, 80143 Naples, Italy
camastra@ieee.org

Abstract. Dimensionality Reduction methods are effective preprocessing techniques that clustering algorithms can use for coping with high dimensionality. Dimensionality Reduction methods have the aim of projecting the original data set of dimensionality d, minimizing information loss, onto a lower M-dimensional submanifold. Since the value of M is unknown, techniques that allow knowing in advance the value of M, called intrinsic dimension (ID), are quite useful. The aim of the paper is to make the state-of-art of the methods of intrinsic dimensionality estimation, underlining the achievements and the challanges.

1 Introduction

Dimensionality Reduction methods are effective preprocessing techniques that clustering algorithms can use for coping with high dimensionality. Dimensionality Reduction methods have the aim of projecting the original data set $\Omega \subset \mathbb{R}^d$, minimizing information loss, onto a lower M-dimensional submanifold of $\mathbb{R}^d$. Since the value of M is unknown, techniques that allow knowing in advance the value of M, are quite useful. Following Fukunaga, a data set $\Omega \subset \mathbb{R}^d$ is said to have *intrinsic dimension* (ID) [16] equal to M if its elements lie entirely within a M-dimensional submanifold of $\mathbb{R}^d$, where $M < d$. It is important to observe that ID depends on the scale of data. In order to show this, it considers a two-dimensional data set, e.g., a K-Möbius strip [20], adding to a data set a three-dimensional gaussian noise. The data set, obtained in this way, has ID equal to 2 at a coarse scale, since the two-dimensional set is dominant. But if we change scale and observe the data set at fine scale, the noise becomes dominant and the ID of data set is three. ID estimation of a data set is a classical problem of pattern recognition and machine learning. The first algorithm of data dimensionality estimation, by Bennett, dates back to 1969 [3]. ID estimation is relevant in machine learning not only for dimensionality reduction methods but also for other several reasons. Firstly, using more dimensions than the necessary leads to several problems, such as an increase of the space required to store data, a decrease in the algorithm speed, since it generally depends on the data dimensionality. Besides, building reliable classifiers becomes harder and harder when the dimensionality grows (*curse of dimensionality* [2]). To this purpose,

F. Masulli et al. (Eds.): CHDD 2012, LNCS 7627, pp. 87–101, 2015.
DOI: 10.1007/978-3-662-48577-4_6

we recall that the capacity (*VC-dimension*) [57] of the linear classifiers, that determines their generalization capability, may depend on ID. Finally, ID is relevant for some prototype-based clustering algorithms. For example, ID affects the *magnification factor* [59] of a trained Neural Gas, that expresses the relation between the data density and the density of the neural gas weight vectors[1].

The aim of the paper is to make the state-of-art of the methods of the intrinsic dimensionality estimation, underlining the advances and the open problems. Extending the taxonomy proposed by Jain and Dubes [25], we group the algorithms for estimating ID in three disjoint categories, i.e., *local*, *global*, *mixed*. In the local category, there are the algorithms that provide an ID estimation using the information contained in sample neighborhoods. To the global category belong the algorithms that make use of the whole data set providing a unique and global ID estimate for the data set. Finally, in the mixed category, there are the algorithms that can produce both a global ID estimate of the whole data set and local ID estimate of particular subsets of the data set. In the paper the most relevant algorithms for each category, underlining their weak points, will be presented. In particular, it will be discussed the robustness of each method with respect to the high dimensionality. The paper is organized as follows: Sects. 2, 3, 4 describe global. local and mixed methods, respectively; the benchmarking of ID estimation method is discussed in Sects. 5 and 6 open problems are analyzed and some conclusion are drawn.

2 Global Methods

In the global category, the algorithms unfold the data set in the d-dimensional manifold. Unlike local methods that use only the information contained in the neighborhood of each data sample, global algorithms make use of the whole data set. These methods make implicitly the assumption that the data lie on a unique manifold of a fixed dimensionality. Global methods can be grouped in four families: *projection techniques*, *fractal-based algorithms*, *multidimensional scaling methods* and *other techniques*, where in the last category are collected all the methods that cannot be assigned to the first three categories.

2.1 Projection Techniques

Projection techniques search for the best subspace to project the data by minimizing the projection error. *Principal Component Analysis* (*PCA*) [26,30] is the simplest and the most widely used projection method. PCA is a linear projection method since projects the data along the directions of maximal variance. PCA algorithm for ID estimation has the following steps:

1. Compute the N eigenvalues of the covariance matrix. Order them in decreasing way, such that $\lambda_1 \geq \lambda_2, \cdots \geq \lambda_N$.

[1] If we denote with P the relation between the data density P and the density ρ of the weight vectors, then $\rho \propto P^{\alpha}$ where $\alpha = \frac{ID}{ID+2}$.

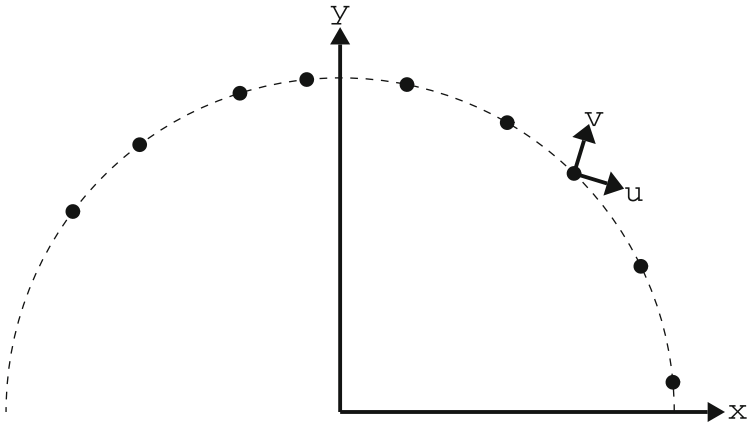

Fig. 1. Ω Data set. The data set is formed by points lying on the upper semicirconference of equation $x^2 + y^2 = 1$. The ID of Ω is *1*. Neverthless PCA yields *two* non-null eigenvalues. The principal components are indicated by u and v.

2. Normalize the eigenvalues dividing each eigenvalue by the largest one λ_1.
3. Choose a threshold value θ and compute the integer K such that $\lambda_K \geq \theta$ and $\lambda_{K+1} < \theta$.
4. return (ID=K).

It is easy to show that the loss of the information due to the discarding the lowest (N-K) eigenvectors is equal to the sum of the lowest (N-K) eigenvalues [4]. PCA is a poor estimator, since it tends to overestimate the ID. Consider a data set formed by datapoints lying on a circumference, (Fig. 1) PCA yields an ID estimate equal to 2 instead of the correct value of 1. Therefore we can assess that, since PCA overestimates ID, PCA provides can be an upper bound of the actual ID value of a dataset. Nonlinear projection methods have been designed in order to overcome the PCA limitations. In order to cope with these problems, some algorithms have been proposed to get Nonlinear PCAs. A widely used approach to get a Nonlinear PCA is the autoassociative approach [28]. Nonlinear PCA is performed by means of a five-layers neural network. The neural net has a typical bottleneck structure. The first (*input*) 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. Though autoassociative neural networks (ANNs) outperforms linear PCA, as ID estimators, in some contexts, ANNs present some drawbacks. ANNs cannot model curves or surfaces that intersect themselves. Moreover, ANN projections onto curves and surfaces are suboptimal [37].

2.2 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 [27]. Unless other global methods, they can provide as ID estimation a non-integer value. Since fractals are generally[2] characterized by a non-integer dimensionality, for instance the dimension of Cantor's set and Koch's curve [38] 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 [13] have been proposed. The *Box-Counting* and the *Correlation* dimension are the most popular. The first definition of dimension (*Hausdorff dimension*) [13,41] is due to Hausdorff [19]. 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*) [41].

Kégl's Algorithm. Let $\Omega = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell\}$ be a set of points in $\mathbb{R}^n$ of cardinality ℓ. We denote with $\nu(r)$ the number of the boxes (i.e., hypercubes) of size r required to cover Ω. It can be proven [41] that $\nu(r)$ is proportional to $(\frac{1}{r})^d$, where d is the *dimension* of the set Ω. This motivates the following definition. The Box-Counting dimension (or *Kolmogorov capacity*) D_B of the set Ω [41] is defined by

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

where the limit is assumed to exist. Recently Kégl [29], has proposed a fast algorithm (*Kégl's algorithm*) to estimate the Box-Counting dimension. Kégl's algorithm is based on the observation that $\nu(r)$ is equivalent to the cardinality of the maximum independent vertex set $MI(G_r)$ of the graph $G_r(V, E)$ with vertex set $V = \Omega$ and edge set $E = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \,|\, d(\boldsymbol{x}_i, \boldsymbol{x}_j) < r\}$. Kégl has proposed to estimate $MI(G)$ using the following greedy approximation. Given a data set Ω, we start with an empty set $\mathcal{C}$. In an iteration over Ω, we add to $\mathcal{C}$ data points that are at distance of at least r from all elements of $\mathcal{C}$. The cardinality of $\mathcal{C}$, after every point in Ω has been visited, is the estimate of $\nu(r)$. The Box-Counting dimension estimate is given by:

$$D_B = -\frac{\ln \nu(r_2) - \ln \nu(r_1)}{\ln r_2 - \ln r_1} \tag{2}$$

where r_2 and r_1 are values that can be set up heuristically. It can be proven [29] that the complexity of Kegl's algorithm is given by $O(D_B \ell^2)$, where ℓ and D_B are the cardinality and the dimensionality of the data set, respectively.

Grassberger-Procaccia Algorithm. A good substitute for the Box-Counting dimension can be the *Correlation dimension* [18]. Due to its computational simplicity, the Correlation dimension is successfully used to estimate the dimension

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

of attractors of dynamical systems. The *Correlation dimension* [18] of a set Ω is defined as follows. If the *correlation integral* $C_m(r)$ is defined as:

$$C_m(r) = \lim_{\ell \to \infty} \frac{2}{\ell(\ell-1)} \sum_{i=1}^{\ell} \sum_{j=i+1}^{\ell} I(\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \leq r) \tag{3}$$

where I is an *indicator function*[3], then the Correlation dimension D of Ω is:

$$D = \lim_{r \to 0} \frac{\ln(C_m(r))}{\ln(r)} \tag{4}$$

It can be proven that the Correlation Dimension is a lower bound of the Box-Counting Dimension. The most popular method to estimate Correlation dimension is the *Grassberger-Procaccia algorithm* [18]. 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 (see Fig. 2b). The computational complexity of the Grassberger-Procaccia algorithm is $O(\ell^2 s)$ where ℓ is the cardinality of the data set and s is the number of different times that the integral correlation is evaluated, respectively. However, there are efficient implementations of the Grassberger-Procaccia algorithm whose complexity does not depend on s. For these implementations, the computational complexity is $O(\ell^2)$.

Takens' Method. Takens [50] has proposed a method, based on *Fisher's method of Maximum Likelihood* [12], that allows to estimate the correlation dimension with a standard error. Let Q be the following set $Q = \{q_k \mid q_k < r\}$ where q_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:

$$\langle D_c \rangle = -\left(\frac{1}{|Q|} \sum_{k=1}^{|Q|} q_k \right)^{-1} \tag{5}$$

where $|Q|$ stands for the cardinality of Q. Takens' method presents some drawbacks. It requires some heuristics to set the radius [53]. 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 it can perform poorly [52]. Finally, Hein and Audibert [20] proposed a generalization of the correlation integral, in term of U-statistics [22], defined as follows:

$$U_{n,h}(K) = \frac{2}{\ell(\ell-1)} \sum_{i=1}^{\ell} \sum_{j=i+1}^{\ell} \frac{1}{h^m} K(\|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2 / h^2) \tag{6}$$

where $K(\cdot)$ is a generic kernel of band width h and m is the dimensionality of the manifold where the data are assumed that lie. On the basis of the Hoeffding

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

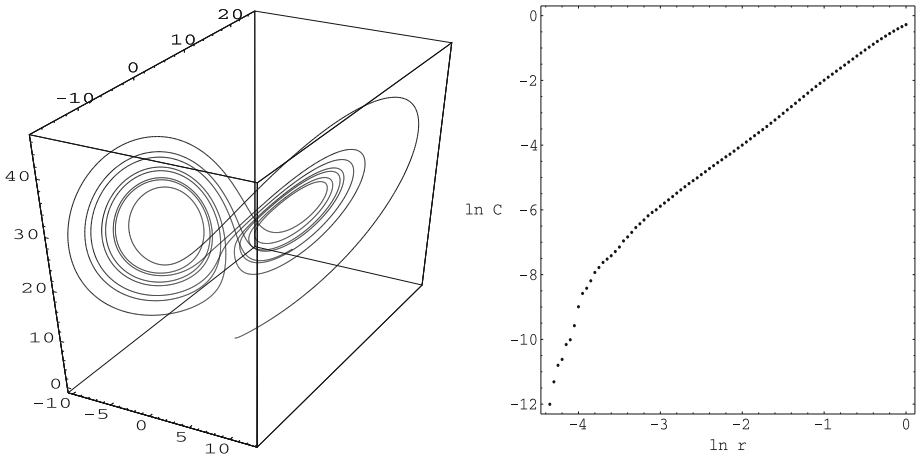

Fig. 2. (a) The attractor of the Lorentz system. (b) The log-log plot on Data set A. Data set A is a real data time series generated by a Lorentz-like system, implemented by NH_3-FIR lasers.

Theorem [23], to guarantee the convergence of the U-statistics the bandwidth h must fulfill $\ell h^m \to \infty$. Hein and Audibert used this property by fixing aconvergence rate for each dimension, that means weare fixing h as a function of the data set cardinality ℓ. Then the Eq. (6) is computed for subsamples of different cardinalities, where h varies according to the function we have fixed. ID is determined by the U-statistic which has the smallest slope as a function of h. It is worth to remark that, Hein and Audibert's algorithm tries, even if partially, to address the problem of ID dependence on the data scale.

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 [14] that in order to get an accurate estimate of the dimension D, the set cardinality ℓ has to satisfy the so-called *Eckmann-Ruelle's inequality*, $D < 2\log_{10}\ell$.

The inequality shows that the number ℓ of data points required 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 ℓ. In order to cope with this problem and to improve the reliability of the measure for low values of ℓ, the *method of surrogate data* [54] has been proposed. The method of surrogate data is an application of *bootstrap* [15]. Given a data set Ω, the method of surrogate data consists in creating a new synthetic data set Ω', with larger 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 arbitrarily, the method of surrogate data is infeasible when the dimensionality of the data set is high.

In-fact a 18-dimensional data set to be estimated must have at least, on the base of the Eckmann-Ruelle' inequality, 10^9 points. Camastra and Vinciarelli [7,8] proposed a procedure to power Grassberger and Procaccia method (GP method), establishing empirically how much GP method underestimates the dimensionality of a data set when data set cardinality is unadequate. Consider a set Ω of cardinality ℓ. The procedure is the following:

1. Create a set Ω', whose ID d is known, with the same cardinality ℓ of Ω. For instance, Ω' could be composed of ℓ data points randomly generated in a d-dimensional hypercube.
2. Measure the correlation dimension D of Ω' with the GP method.
3. Repeat the two previous steps for T different values of d, obtaining the set $C = \{(d_i, D_i) : i = 1, 2, \ldots, T)$.
4. Perform a best-fitting to the data points in C. A plot (*reference curve*) Γ of D versus d is generated. The reference curve allows to infer the value of D when d is known.
5. The correlation dimension D of Ω is computed by GP method and, using Γ, the intrinsic dimension of Ω can be estimated.

The procedure assumes implicitly that the curve Γ depends on ℓ and the dependence of Γ on the Ω' sets are negligible. It is worth to mention that Oganov and Valle [56] used GP method in conjunction to Camastra and Vinciarelli procedure's to estimate ID of Crystal Fingerprint spaces.

2.3 Multidimensional Scaling and Other Methods

Multidimensional Scaling (MDS) [44] 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. To each projection is associated an index, usually defined *stress*, that measures the goodness of the projection. The best projection is the one whose stress is minimal. Examples of the Multidimensional scaling methods are Bennett's algorithm [3], that now has only historical interest, *MDSCAL* [31], *Sammon's mapping* [47]. In the Other Methods category, are collected the methods that do not belong to fractal, projection and MDS categories. To Other Methods category belong Costa-Hero [11] algorithm and the algorithms recently proposed by Rozza et al. [46] and Lombardi et al. [35]. For the sake of brevity, we only describe the first algorithm. Costa-Hero's algorithm assumes that data lie on a manifold. The algorithm exploits entropic graphs on in order to estimate the ID dimensionality and the entropy of the manifold. The algorithm is founded on the fact that the length function, computed on the whole graph, depends on ID.

3 Local Methods

Local methods are algorithms that provide an ID estimation using the information contained in sample neighborhoods, avoiding the projection of the data onto

a lower-dimensional manifold. In this case, data do not lie on a unique manifold of constant dimensionality but on multiple manifolds of different dimensionalities. Since a unique ID estimate for the whole data is clearly not meaningful, it prefers to provide an ID estimate for each small subset of data, assuming that it lies on a manifold of constant dimensionality. More formally, local (or *topological*) methods try to estimate the topological dimension of the data manifold. The definition of topological dimension was given by Brouwer [21] in 1913. Topological dimension is the basis dimension of the local linear approximation of the hypersurface where the data reside, 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. Algorithms that belong to this category are Fukunaga-Olsen [17], Bruske-Sommer [5], Trunk [55], Pettis et al. [42] and Verveer and Duin [58] ones.

3.1 Fukunaga-Olsen's Algorithm

Fukunaga-Olsen's algorithm is based on the observation that for data 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 [33]. 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.

3.2 TRN-Based and Local MDS Methods

Topology Representing Network (TRN) is a unsupervised neural network proposed by Martinetz and Schulten [39]. 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 [5] 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, we conclude the section on local methods quoting the local MDS methods. As the global MDS methods discussed in Sect. 2.3, local MDS methods are projection techniques that tend to preserve, as much as possible, the distances among data. In local MDS, in an analogous manner to global MDS, to each projection is associated an index or a cost that measures the goodness of the projection. Unlike MDS methods, where the whole data set is considered, local MDS methods work only on a small subset of data. Examples of Local MDS methods are ISOMAP [51] and Local Linear Embedding (LLE) [45]. The method for estimating ID is the same of global MDS. Compute several MDS projection considering different dimensionality for the output space. Pick the MDS projection with the best index or the minimum cost. The ID is given by the dimensionality of the output space of the MDS projection selected.

4 Mixed Methods

The most relevant methods that belong to this category are Levina-Bickel [32] and Carter-Raich-Hero algorithms [9]. For the sake of brevity, we only describe the former algorithm.

4.1 Levina-Bickel Algorithm

The Levina-Bickel algorithm provides a maximum likelihood ID estimate. The Levina-Bickel algorithm derives the maximum likelihood estimator (MLE) of the intrinsic dimensionality D from a data set $\Omega = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell) \in \mathbb{R}^n$. The dataset Ω represents an embedding of a lower-dimensional sample, i.e., $\boldsymbol{x}_i = g(Y_i)$ where Y_i are sampled from an unknown smooth density f on $\mathbb{R}^D$ with $D \leq n$, g is a smooth mapping. Last assumption guarantees that close data in $\mathbb{R}^D$ are mapped to close neighbors in the embedding. That being said, we fix a data point $\boldsymbol{x} \in \mathbb{R}^n$ assuming that $f(\boldsymbol{x})$ is constant in a sphere $S_{\boldsymbol{x}}(r)$ centered in $\boldsymbol{x}$ of radius r and we view Ω as a homogeneous Poisson process in $S_{\boldsymbol{x}}(r)$. Given the inhomogeneous process $\{P(t, \boldsymbol{x}), 0 \leq t \leq r\}$

$$P(t, \boldsymbol{x}) = \sum_{i=1}^{\ell} I(\boldsymbol{x}_i \in S_{\boldsymbol{x}}(t)), \tag{7}$$

which counts the data whose distance from $\boldsymbol{x}$ is less than t. If we approximate it by means a Poisson process and we neglect the dependence on $\boldsymbol{x}$, the rate $\lambda(t)$ of the process $P(t)$ is given by:

$$\lambda(t) = f(\boldsymbol{x}) V(D) D t^{D-1}, \tag{8}$$

where $V(D)$ is the volume of a D-dimensional unit hypersphere. The Eq. (8) is justified by the Poisson process properties since the surface area of the sphere $S_{\boldsymbol{x}}(t)$ is $\frac{d}{dt}[V(D)t^D] = V(D)Dt^{D-1}$. If we define $\theta = logf(\boldsymbol{x})$, the log-likelihood of the process $P(t)$ [49] is:

$$L(D,\theta) = \int_0^r log\lambda(t)dP(t) - \int_0^r \lambda(t)dt. \tag{9}$$

The equation describes an exponential family for which a maximum likelihood estimator exists with probability that tends to 1 as the number of samples ℓ tends to infinity. The maximum likelihood estimator is unique and must satisfy the following equations:

$$\frac{\partial L}{\partial \theta} = \int_0^r dP(t) - \int_0^r \lambda(t)dt = P(r) - e^{\theta}V(D)r^D = 0. \tag{10}$$

$$\begin{aligned}\frac{\partial L}{\partial D} = &\left(\frac{1}{D} + \frac{V'(D)}{V(D)}\right)P(r) + \int_0^r log\ t\ dP(t) + \\ &- e^{\theta}V(D)r^D\left(log\ r + \frac{V'(D)}{V(D)}\right) = 0.\end{aligned} \tag{11}$$

If we plug the Eq. (10) into the Eq. (11) we obtain the maximum likelihood estimate for the dimensionality D:

$$\hat{D}_r(\boldsymbol{x}) = \left[\frac{1}{P(r,\boldsymbol{x})}\sum_{j=1}^{P(r,\boldsymbol{x})} log\frac{r}{T_j(\boldsymbol{x})}\right]^{-1}, \tag{12}$$

where $T_j(\boldsymbol{x})$ denotes the Euclidean distance between $\boldsymbol{x}$ and its j-th nearest neighbor. Levina and Bickel suggest to fix the number of the neighbors k rather than the radius of the sphere r. Therefore the estimate becomes:

$$\hat{D}_k(\boldsymbol{x}) = \left[\frac{1}{k-1}\sum_{j=1}^{k-1} log\frac{T_k(\boldsymbol{x})}{T_j(\boldsymbol{x})}\right]^{-1}. \tag{13}$$

The estimate of the dimensionality is obtained averaging on all data points of the data set Ω, that is:

$$\hat{D}_k = \frac{1}{\ell}\sum_{i=1}^{\ell}\hat{D}_k(\boldsymbol{x}_i) \tag{14}$$

The estimate of the dimensionality depends on the value of k. Levina and Bickel suggest to average over a range of values of $k = k_1, \ldots, k_2$ obtaining the final estimate of the dimensionality, i.e.,

$$\hat{D} = \frac{1}{k_2 - k_1 + 1}\sum_{k=k_1}^{k=k_2}\hat{D}_k. \tag{15}$$

David Mac Kay and Zoubin Ghamarani, in an unpublished comment [36], made a strong criticism against Levina and Bickel' s procedure of the global ID estimation. Instead, they proposed to average the inverse of the estimators $\hat{D}_k(\boldsymbol{x}_i)$. In this way, the Eq. (14) has to be replaced with:

$$\hat{D}_k = \frac{\ell(k-1)}{\sum_{i=1}^{\ell}\sum_{j=1}^{k-1} log\frac{T_k(\boldsymbol{x}_i)}{T_j(\boldsymbol{x}_i)}} \tag{16}$$

Using the same Levina and Bickel's approach, the final estimate of the dimensionality has to be obtained averaging $\hat{D}_k$ over a range of values of $k = k_1, \ldots, k_2$ obtaining the final estimate of the dimensionality expressed by Eq. (15). Regarding the computational complexity, the Levina-Bickel algorithm requires a sorting algorithm[4], whose complexity is $O(\ell \log \ell)$, where ℓ denotes the cardinality of the data set. Hence the computational complexity for estimating $\hat{D}_k$ is $O(k\ell^2 \log \ell)$, where k denotes the numbers of the neighbors that have to be considered. Besides, Levina and Bickel suggest to consider an average estimate repeating the estimate D_k s times, where s is the difference between the maximum and the minimum value that k can assume, i.e., k_2 and k_1, respectively. Therefore the overall computational complexity of the Levina-Bickel algorithm is $O(k_2 s\ell^2 \log \ell)$.

5 ID Estimation Methods Benchmarking

A crucial issue in ID estimation is the experimental validation of the algorithms designed for ID estimation. The experimental validation of such an algorithm requires benchmarks, i.e., data sets, Benchmarks can be of two different types: synthetical or real data. Regarding synthetical benchmarks, it is not difficult to build synthetical data sets of given ID [20]. Moreover, the literature offer a certain number of synthetical benchmarks, both low-dimensional and high dimensional. To this purpose, it is worth to mention 2-dimensional *Swiss Roll* [51], 3-dimensional 10-Möbius strip [20], 9-dimensional data set D of Santa Fe time series competition [43], 12-dimensional manifold [20]. Unlike synthetical benchmarks, it can be cumbersome to get real data benchmarks of known ID. Firstly, it is necessary to split the benchmarks in two subfamilies: low-dimensional and high-dimensional. Regarding low-dimensional real data benchmarks, the literature offers a limited availability of benchmarks, e.g., the 3-dimensional Face Set [51] and the attractors in the phase space, of known dimensionality, generated, using *method of delays* [41], by real data time series. To this purpose, it is worth to mention the Lorentz attractor generated by the data set A[5] [24]

[4] The complexity of effective sorting algorithms (e.g., mergesort and heapsort) is $\ell \log \ell$, where ℓ is the number of elements that have to be sorted.

[5] The data set A is a real data time series generated by a Lorentz-like chaotic system, implemented by NH_3-FIR lasers.

Table 1. Chua's circuit and Data set A attractor dimension estimates by Kégl, Levina-Bickel, Grassberger-Procaccia methods.

	Data set A attractor dimension	Chua's circuit attractor dimension
Kégl estimate	2.02	2.14
Levina-Bickel estimate	2.35	2.26
Grassberger-Procaccia estimate	2.00	2.20
Theoretical value	2.06	∼2.26

and Chua's attractor generated by a real data time series, measured from a hardware realization [1] of Chua's circuit [10]. In Table 1 some experimental comparisons [6] among ID estimators, performed on Data Set A and Chua's circuit, are reported. If we pass to consider high-dimensional real data benchmarks of known ID, the situation becomes very difficult. To our best knowledge, the only high-dimensional benchmarks are the *Crystal Fingerprint spaces* (or *Crystal Fingerspaces*) [40,56] recently proposed by Oganov and Valle in Crystallography with the aim of representing crystalline structures. Crystal Fingerprint spaces are spaces built starting by the real measured distances between atoms in the crystalline structure. The theoretical ID of a Crystal Fingerspace, based on crystal degree of freedoms, is 3N+3, where N is the number of the atoms in the crystalline unitary cell. Crystal Fingerspaces have been derived for several crystal structures, e.g., 39-dimensional H_2O (crystalline cell with 8 atoms) and 147-dimensional SiO_2 (crystalline cell with 48 atoms). Crystal Fingerspace data are available at http://mariovalle.name/CrystalFp/index.php/CrystalFpLib/Data.

6 Conclusions

In the paper we have reviewed the intrinsic dimension estimation methods underlining their advances. Nevertheless, some problem remain open. As remarked previously, intrinsic dimension depends on the scale of data. Although some ID estimation methods [20,34] tried to take in account, even if partially, of the data scale, a reliable multiscale ID estimator is not available, yet. The other open problems are related to the robustness of ID estimators w.r.t. the curse of dimensionality. About this topic, there are two issues that remain to be fully addressed. The former issue is the following. Each ID estimation method should provide a lower bound on the cardinality in order to guarantee an accurate ID estimation. To our best knowledge, this lower bound [14,48] is available only for Correlation Dimension estimation methods, e.g., Eckmann-Ruelle's inequality, whereas the other algorithms fully ignored the topic. The latter issue is the lack of the robustness of ID estimators w.r.t. high dimensionality. Although an empirical solution [8] was proposed, the construction of a robust ID estimators w.r.t. high dimensionality remains one of the challange of the research in machine learning.

Acknowledgements. Firstly, the author wish to thank Mario Valle for having made public Crystal Fingerspace datasets and commenting on the draft. The author thanks the anonymous referees for their useful remarks.

References

1. Aguirre, L., Rodrigues, G., Mendes, E.: Nonlinear identification and cluster analysis of chaotic attractors from a real implementation of chua's circuit. Int. J. Bifurcat. Chaos **6**(7), 1411–1423 (1997)
2. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
3. Bennett, R.S.: The intrinsic dimensionality of signal collections. IEEE Trans. Inf. Theory **15**, 517–525 (1969)
4. Bishop, C.: Neural Networks for Pattern Recognition. Cambridge University Press, Cambridge (1995)
5. Bruske, J., Sommer, G.: Intrinsic dimensionality estimation with optimally topology preserving maps. IEEE Trans. Pattern Anal. Mach. Intel. **20**(5), 572–575 (1998)
6. Camastra, F., Filippone, M.: A comparative evaluation of nonlinear dynamics methods for time series prediction. Neural Comput. Appl. **18**(8), 1021–1029 (2009)
7. Camastra, F., Vinciarelli, A.: Intrinsic dimension estimation of data: an approach based on grassberger-procaccia's algorithm. Neural Process. Lett. **14**(1), 27–34 (2001)
8. Camastra, F., Vinciarelli, A.: Estimating the intrinsic dimension of data with a fractal-based method. IEEE Trans. Pattern Anal. Mach. Intel. **24**(10), 1404–1407 (2002)
9. Carter, K., Raich, R., Hero, A.: On local intrinsic dimension estimation and its application. IEEE Trans. Sig. Process. **58**(2), 650–663 (2010)
10. Chua, L., Komuro, M., Matsumoto, T.: The double scroll. IEEE Trans. Circuits Syst. **32**(8), 797–818 (1985)
11. Costa, J., Hero, A.: Geodesic entropic graphs for dimension and entropy estimation in manifold learning. IEEE Trans. Sig. Process. **52**(8), 2210–2221 (2004)
12. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley, New York (2001)
13. Eckmann, J.P., Ruelle, D.: Ergodic theory of chaos and strange attractors. Rev. Mod. Phys. **57**, 617–659 (1985)
14. Eckmann, J.P., Ruelle, D.: Fundamental limitations for estimating dimensions and lyapounov exponents in dynamical systems. Physica **D–56**, 185–187 (1992)
15. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)
16. Fukunaga, K.: Intrinsic dimensionality extraction. In: Krishnaiah, P.R., Kanal, L.N. (eds.) Classification, Pattern Recognition and Reduction of Dimensionality. Handbook of Statistics, pp. 347–360. North Holland, Amsterdam (1982)
17. Fukunaga, K., Olsen, D.: An algorithm for finding intrinsic dimensionality of data. IEEE Trans. Comput. **C–20**(2), 176–183 (1971)
18. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica D **9**, 189–208 (1983)
19. Hausdorff, F.: Dimension und äusseres mass. Mathematische Annalen **79**, 57 (1918)
20. Hein, M., Audibert, J.Y.: Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In: ICML 2005 Proceedings of 22^{nd} International Conference on Machine Learning, pp. 289–296 (2005)

21. Heyting, A., Freudenthal, H.: Collected Works of L.E.J Brouwer. North Holland Elsevier, Amsterdam (1975)
22. Hoeffding, W.: A class of statistics with asymptotically normal distributions. Ann. Stat. **19**, 293–325 (1948)
23. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
24. Hubner, U., Weiss, C., Abraham, N., Tang, D.: Lorentz-like chaos in nh_3-fir lasers. In: Gershenfeld, N.A., Weigend, S.A. (eds.) Time Series Prediction: Forecasting the Future and Understanding the Past, pp. 73–104. Addison Wesley, Reading (1994)
25. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice Hall, New Jersey (1988)
26. Jollife, I.T.: Principal Component Analysis. Springer, New York (1986)
27. Kaplan, D., Glass, L.: Understanding Nonlinear Dynamics. Springer, New York (1995)
28. Karhunen, J., Joutsensalo, J.: Representations and separation of signals using nonlinear PCA type learning. Neural Netw. **7**(1), 113–127 (1994)
29. Kégl, B.: Intrinsic dimension estimation using packing numbers. In: Thrun, S., Saul, L.K., Schölkopf, B. (eds.) Advances in Neural Information Processing. MIT Press, Cambridge (2003)
30. Kirby, M.: Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns. Wiley, New York (2001)
31. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**, 1–27 (1964)
32. Levina, E., Bickel, P.: Maximum likelihood estimation of intrinsic dimension. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing, pp. 777–784. MIT Press, Cambridge (2005)
33. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Trans. Commun. **28**(1), 84–95 (1980)
34. Little, A., Jung, Y.M., Maggioni, M.: Multiscale estimation of intrinsic dimensionality of a data set. In: Manifold Learning and Its Applications: Papers from the AAAI Fall Symposium, pp. 26–33. IEEE (2009)
35. Lombardi, G., Rozza, A., Ceruti, C., Casiraghi, E., Campadelli, P.: Minimum neighbor distance estimators of intrinsic dimension. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part II. LNCS, vol. 6912, pp. 374–389. Springer, Heidelberg (2011)
36. MacKay, D., Ghamarani, Z.: Comments on 'Maximum likelihood estimation of intrinsic dimension by E. Levina and M. Bickel', University of Cambridge (2005). http://inference.phy.cam.uc.uk/mackay/dimension
37. Malthouse, E.C.: Limitations of nonlinear PCA as performed with generic neural networks. IEEE Trans. Neural Netw. **9**(1), 165–173 (1998)
38. Mandelbrot, B.: Fractals: Form, Chance and Dimension. Freeman, San Francisco (1977)
39. Martinetz, T., Schulten, K.: Topology representing networks. Neural Netw. **3**, 507–522 (1994)
40. Oganov, A., Valle, M.: How to quantify energy landscapes of solids. J. Chem. Phys. **130**, 104504 (2009)
41. Ott, E.: Chaos in Dynamical Systems. Cambridge University Press, Cambridge (1988)
42. Pettis, K., Bailey, T., Jain, T., Dubes, R.: An intrinsic dimensionality estimator from near-neighbor information. IEEE Trans. Pattern Anal. Mach. Intel. **1**(1), 25–37 (1979)

43. Pineda, F., Sommerer, J.: Estimating generalized dimensions and choosing time delays: a fast algorithm. In: Weigend, S., Gershenfeld, N.A. (eds.) Time Series Prediction: Forecasting the Future and Understanding the Past, pp. 367–385. Addison Wesley, Reading (1994)
44. Romney, A.K., Shepard, R.N., Nerlove, S.B.: Multidimensionaling Scaling, vol. I. Theory. Seminar Press, New York (1972)
45. Roweis, S., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(12), 2323–2326 (2000)
46. Rozza, A., Lombardi, G., Rosa, M., Casiraghi, E., Campadelli, P.: IDEA: intrinsic dimension estimation algorithm. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part I. LNCS, vol. 6978, pp. 433–442. Springer, Heidelberg (2011)
47. Sammon, J.W.J.: A nonlinear mapping for data structure analysis. IEEE Trans. Comput. **C–18**, 401–409 (1969)
48. Smith, R.: Optimal estimation of fractal dimension. In: Casdagli, M., Eubank, S. (eds.) Nonlinear Modeling and Forecasting, pp. 115–135. Addison Wesley, New York (1992)
49. Snyder, D.: Random Point Processes. Wiley, New York (1975)
50. Takens, F.: On the numerical determination of the dimension of an attractor. In: Dynamical Systems and Bifurcations, Proceedings Groningen 1984, pp. 99–106. Addison Wesley (1985)
51. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(12), 2319–2323 (2000)
52. Theiler, J.: Lacunarity in a best estimator of fractal dimension. Phys. Lett. A **133**, 195–200 (1988)
53. Theiler, J.: Statistical precision of dimension estimators. Phys. Rev. **A41**, 3038–3051 (1990)
54. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D.: Testing for nonlinearity in time series: the method for surrogate date. Physica D **58**, 77–94 (1992)
55. Trunk, G.V.: Statistical estimation of the intrinsic dimensionality of a noisy signal collection. IEEE Trans. Comput. **25**, 165–171 (1976)
56. Valle, M., Oganov, A.: Crystal fingerprint space- a novel paradigm for studying crystal-structure sets. Acta Crystallogr. Sect. A **A66**, 507–517 (2010)
57. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
58. Verveer, P.J., Duin, R.: An evaluation of intrinsic dimensionality estimators. IEEE Trans. Pattern Anal. Mach. Intel. **17**(1), 81–86 (1995)
59. Villmann, T., Claussen, J.C.: Magnification control in self-organizing maps and neural gas. Neural Comput. **18**(2), 446–469 (2000)