JOURNAL OF COMMUNICATION AND INFORMATION SYSTEMS, VOL. 35, NO.1, 2020.

300

Bias-Compensated Estimator for Intrinsic Dimension and Differential Entropy: A Visual Multiscale Approach

Jugurta Montalvão, Jânio Canuto, and Luiz Miranda

Abstract—**Intrinsic dimension and differential entropy estimators are studied in this paper, including their systematic bias. A pragmatic approach for joint estimation and bias correction of these two fundamental measures is proposed. Shared steps on both estimators are highlighted, along with their useful consequences to data analysis. It is shown that both estimators can be complementary parts of a single approach, and that the simultaneous estimation of differential entropy and intrinsic dimension give meaning to each other, where estimates at different observation scales convey different perspectives of underlying manifolds. Experiments with synthetic and real datasets are presented to illustrate how to extract meaning from visual inspections, and how to compensate for biases.**

Index Terms—**Manifold analysis, Bias correction, Intrinsic dimension, Collision entropy, Correlation dimension.**

I. INTRODUCTION

INTRINSIC dimension (ID) estimation is a useful tool whenever patterns presented in D-dimensional spaces are supposed to form structures (manifolds) in d-dimensional subspaces, with $d < D$. Examples of such lower dimensional structures are: projections of a rigid objects whose pictures, with D pixels, are taken under d degrees of freedom [1], [2], or D-dimensional representations of vowel sounds, whereas the vocal tract that generates the sound has only d mechanical degrees of freedom [3].

In all those applications, if probabilistic models are used to represent the source of observations (i.e., the underlying d-dimensional structures) then entropy, differential entropy (DE) and entropy rate [4] can reveal relevant attributes of the corresponding structures. In pattern recognition, estimating both ID and DE is tantamount to analysing shape attributes of manifolds, as explained in Section II, thus suggesting tools for proper design and analysis of classifiers, in special those based on autoencoders. Indeed, while the number of deep neural network applications increases at an astonishing pace, some attempts to explain this success seem to suggest that most answers come from the study of physical restrictions [5] and consequent formation of data manifolds [6], [7], [8], [9].

Although ID does not impose a probabilistic model to be estimated, many published ID estimators are based on probabilistic reasoning [10], [11], [12], [13], [14], [15], [16].

J. Montalvão, Jânio Canuto, and Luiz Miranda are with the Federal University of Sergipe (UFS), São Cristóvão, Brazil e-mail: jmontalvao(at)ufs.br. Luiz Miranda is also with the *Institut Mines-Télécom Lille-Douai* (IMT Lille-Douai), France, since Nov. 2020.

Indeed, even the well known Grassberger-Procaccia (GP) estimator [14], whose aim is to characterize strange attractors in dissipative (deterministic) dynamical systems, also uses the information-theoretic framework to better explain the kind of *dimension* their method is able to estimate (also referred to as *information dimension*).

The formulation proposed in [14] includes the use of random variables (RV) as the source model for observations, and explicitly shows a link between intrinsic dimension and differential entropy. Some subsequent works followed this same path, such as [3] and the series of publications by Costa and Hero [12], [17], [18]. However, most published works deal either with DE, under the assumption that ID is known, or with ID estimation, regardless the manifold's volume (thus its DE, as explained in Section II). Indeed, in [3] it is stated that

> *The existence of manifold structures in the data is often overlooked in entropy estimations, with the result that classical methods, assuming the wrong intrinsic dimension (manifold dimension) provide erroneous estimates of the entropy.*

On the other hand, in [19], the problem of DE estimation in high-dimensional spaces was tackled through a simple but data-efficient approach, referred to as the Coincidence Method (CM), originally applied in Physics. In [20] this method was extended to differential entropy estimation in the pattern recognition context, which clearly shows that the *correlation dimension* in [14] uses the same empirical coincidence ratio as the entropy estimation method proposed in [19].

More specifically, the correlation integral defined in [14] is equivalent to the inverse of the number of coincidences defined in [19]. This equivalence is even more striking in non-redundant reformulations of the correlation integral, as in [21]. These equations are the same presented in Section III, in a unified notation, which suggests a link between works from different domains, developed in this paper to yield a visual method where ID and DE are regarded as complementary parameters of the same estimation problem.

Unfortunately, both methods [14], [19] yield biased estimates, a distortion whose source is also shared by them, which is explained by their common theoretical ground. Concerning the bias in the GP method, a theoretical model was first proposed in [22], where it was shown that ID bias can be predicted on average if the actual ID is known. In this paper, the theoretical model proposed in [22] is developed to the

point of predicting and compensating for both ID and DE biases, even if the actual ID is unknown.

This paper is organized as follows: In Section II, a brief recall of ID and DE, and their complementary meanings is presented, whereas in Section III the theoretical foundations of the joint estimator proposed in this paper are presented. Finally, in Section IV, the method is presented, along with a bias compensation approach. Experiments with both real and artificial data are presented in Section V. The main contributions of this work are discussed in Section VI.

II. Intrinsic dimension and differential entropy in a nutshell

According to [23], the ID of a given set of observations is *"the minimum number of free variables needed to represent the data without information loss"*, which agrees with the definition in [24], where *"the intrinsic dimensionality of a collection of signals is defined to be equal to the number of free parameters required in a hypothetical signal generator capable of producing a close approximation to each signal in the collection"*.

DE, on the other hand, is defined as the entropy of a continuous random variable [25]. Besides, [26] presents entropy as an effective cardinality in logarithmic scale. Likewise, DE can be regarded as an effective volume (in logarithmic scale) [20], [25].

For a brief recall on ID and DE, consider the data sources labeled 'Sinusoid' and 'Circle', borrowed from [27]. Although experiments there just consider ID, these datasets can also be used to address DE as well. These sources are defined respectively as

$$\mathbf{X}_{Sin} = [\sin(2\pi U), \ \cos(2\pi U), \ 0.1 \sin(300\pi U)]$$

and

$$\mathbf{X}_{Cir} = [\sin(2\pi U), \ \cos(2\pi U), \ 0.1 V]$$

where U and V are independent random variables uniformly distributed between 0 and 1.

Figures 1 and 2 show 3000 instances of $\mathbf{X}_{Cir}$ and $\mathbf{X}_{Sin}$ respectively. Their ID are 2 and 1, for the domain of $\mathbf{X}_{Sin}$ can be cut and straightened to a line segment (1D) of length slightly greater than 60, whereas the domain of $\mathbf{X}_{Cir}$ can also be cut and unbent to a rectangle (2D) of area 0.2π.

If the probability density function (pdf) of an RV is known, its Rényi α-entropy [25] can be obtained as

$$h_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \log_2 \int_{\mathcal{R}^D} (f_\mathbf{X}(\mathbf{x}))^\alpha d\mathbf{x}$$

where $d\mathbf{x}$ is a differential hypervolume in $\mathcal{R}^D$, only taken where the pdf $f_\mathbf{X}(\mathbf{x})$ is not null. Therefore, if the pdf is not null in d-dimensional manifolds ($d < D$) the integral must be restricted to it (therefore the local dimensions of the manifolds must be known). This definition encompasses both Shannon DE, for $\alpha \to 1$, and quadratic (or collision) entropy, for $\alpha = 2$. In both cases, $h_\alpha(\mathbf{X})$ can also be regarded as a proportion between volumes, suggesting that DE is a measure of *effective volume* for non-uniform distributions, as much as entropy is presented as an *effective cardinality* for discrete RVs [26], [20].

This intuitive perception of DE can be better explained with the notion of effective length, area or volume, as follows: an RV defined as $Z = \lambda U$ ($\lambda \in \mathcal{R}$ and $\lambda > 0$) is uniformly distributed along an 1D domain of length λ, then its DE is given by this length λ measured in logarithm scale, $h(Z) = \log(\lambda)$. In general, for non-uniform RV, the effective hypervolume is given by the hypervolume associated to another uniformly distributed RV whose observation removes the same amount of uncertainty about the outcome [4].

Both formal and intuitive points of view reveal a tricky aspect of DE estimation, that the DE is meaningless before the ID is known. Figures 1 and 2 can be used to further illustrate this point, because both RV $\mathbf{X}_{Cir}$ and $\mathbf{X}_{Sin}$ are defined in a 3D space, but they have ID equal to 2 and 1, respectively. Therefore, the DE associated to $\mathbf{X}_{Cir}$ must take a unit square as area reference to yield $h(\mathbf{X}_{Cir}) = \log_2(0.2\pi)$ bits, whereas $\mathbf{X}_{Sin}$ must take a unit line segment as length reference to yield $h(\mathbf{X}_{Sin}) \approx \log_2(60)$ bits. In both cases, an observer unaware of these IDs would fail to estimate the DE, because both datasets are presented as 3D patterns, but their underlying structures have null volume.
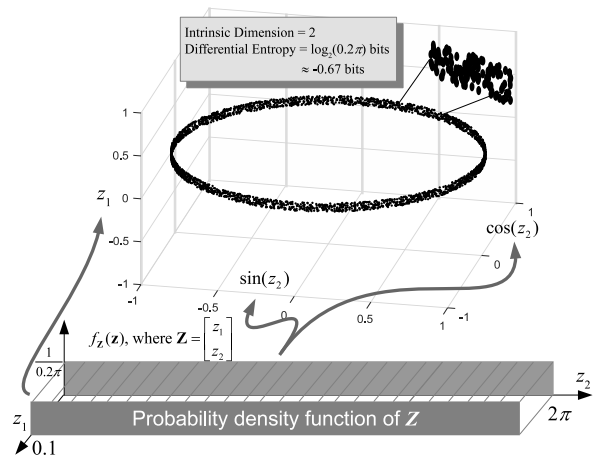

Fig. 1. Dots represent instances of $\mathbf{X}_{Cir}$, which are generated from instances of a 2-D uniform latent random variable $\mathbf{Z} = [0.1V; \ 2\pi U]$. Thus the intrinsic structure of $\mathbf{X}_{Cir}$ is planar ($d = 2$), in spite of its 3D ($D = 3$) representation. The DE of $\mathbf{X}_{Cir}$ is given by the surface area (0.2π), thus $h(\mathbf{X}_{Cir}) = \log_2(0.2\pi) \approx -0.67$ bits.

III. Joint analysis approach

In this Section, two known approaches for ID and DE estimation are briefly recalled. When put side by side, they reveal their striking equivalences, which are then articulated to yield a joint visual analysis for ID and DE.

A. Intrinsic dimension estimation

Given a set of N observations, $\{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)\}$, and a threshold r, the "information dimension" (also known as correlation dimension) is defined in [14], and can be obtained from the proportionality
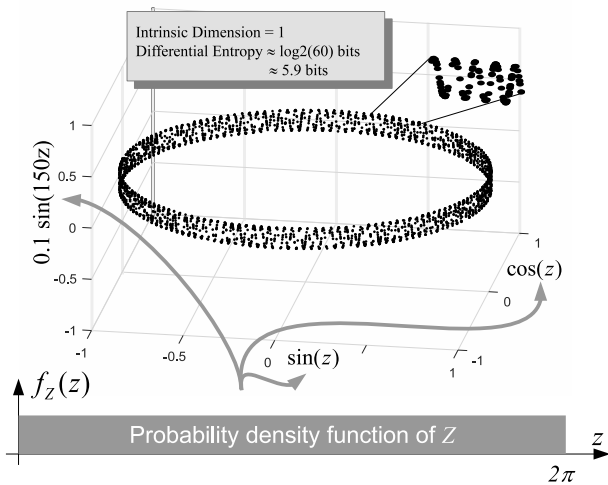
$$C(r) \propto r^d \tag{1}$$

Fig. 2. Dots represent instances of $\mathbf{X}_{Sin}$, which are generated from instances of a single uniform latent random variable $Z = 2\pi U$. Thus the intrinsic structure of $\mathbf{X}_{Sin}$ is 1D ($d = 1$), in spite of its 3D ($D = 3$) representation. The DE of $\mathbf{X}_{Sin}$ is given by the effective structure length (approximately 60), thus $h(\mathbf{X}_{Sin}) \approx 5.9$ bits.

Fig. 3. Illustration of the ID estimation method by Grassberger and Procaccia, for 3000 random instances of $\mathbf{X}_{Cir}$. Depending on the value of r, two main linear trends are noticed, thus suggesting two possible ID estimates.

as $r \to 0$, where the non-redundant [21], [22] definition of the correlation integral $C(r)$ is

$$C(r) = \frac{1}{N(N-1)} \sum_{i<j}^{N} I(\|\mathbf{x}(i) - \mathbf{x}(j)\| \le r) \qquad (2)$$

where I is an indicator function, i.e. $I(\lambda) = 1$ if λ is true, and $I(\lambda) = 0$ otherwise.

Function $I(\cdot)$ is a coincidence detection function that allows for the use of any pattern matching measure, or even mean opinion scores, which can be particularly useful for ID estimations in psychometrics or econometrics, for instance. In [22] the supremum norm is used instead of the original Euclidean norm [14], thus easing theoretical calculations regarding correlation dimension limits.

For both definitions, since the volume where coincidence occurs in the manifold scales with r^d (instead of r^D), then the number of observation pairs coinciding in this volume should scale at a known rate, if the observation volume is such that the probability density of observations is almost constant inside it.

From Eq. 1, it follows that

$$\log C(r) \approx d\log(r) - h, \qquad (3)$$

where, as $r \to 0$, h is the logarithm of the proportionality constant.

To estimate d from Eq. 3, a common approach is to use the angular coefficient of the line that best fits points $(\log r, \log C(r))$ in a given range for r. Therefore, a single best fit is expected. However, Figure 3 illustrates a case where this expectation is frustrated. This visual result was obtained with $N = 3000$ independent observations of $\mathbf{X}_{Cir}$, and r ranging from 0.01 to 1 (points were interpolated to improve visualization). It may be seen that there are two almost linear intervals with angular coefficients close to either 1 or 2, depending on the range of r.
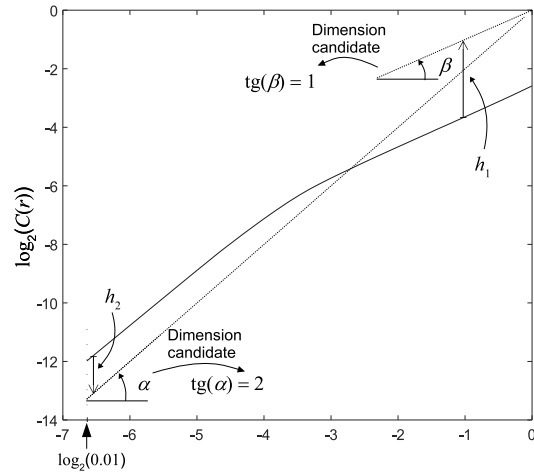
Because GP method is based on results for vanishing values of r, one should assume that the estimated ID is 2, corresponding to the lower part of the curved line in Figure 3. Indeed, the detail presented in Figure 1 clearly shows a 2D local structure. But the estimation for higher values of r is also meaningful, revealing that in a larger scale the 2D structure becomes negligible, whereas an 1D structure emerges.

That is to say that, on one hand this ambiguity is a drawback of this ID estimator, because bad choices for r may yield inconsistent estimates, whereas good choices remain an open problem [10]. On the other hand, this sensitivity to r can be carefully crafted as a tool for multiscale analysis, as discussed in Subsection III-B.

B. Differential entropy estimation

As for differential entropy, our starting point is the estimator proposed by S. Ma in the context of Statistical Mechanics [19]. This method was motivated by the huge number of reachable physical states in the original problem S. Ma addressed. By replacing states with multivariate random observations, or vectors in an abstract signal space [28] a DE estimator is obtained, which is well suited for pattern recognition problems where the amount of observations is small, as compared to the effective size (effective in the sense of [26]) of the observation domain [29].

To estimate the diferential entropy, $h_{\mathbf{X}}(\mathbf{x})$, of a random source modelled as $\mathbf{X}$, Ma's method can be summarized in the following steps:

1. Arbitrarily set a small hypercube volume r^d. It is to be noticed that in the original formulation no intrinsic dimension is considered. Here, however, one should consider the possibility of data lying in a manifold of dimension $d \le D$, which yields an actual hypervolume $r^d \le r^D$.

2. Compare all $N_t = N(N-1)/2$ instance pairs $\mathbf{x}(i)$ and $\mathbf{x}(j)$, $i < j$, and compute $n_c(r)$ as the number of detected coincidences. A coincidence occurs when $\|\mathbf{x}(m) - \mathbf{x}(n)\|_{\infty} < r/2$.

3. Compute the ratio between the number of comparisons and the number of coincidences: $Q(r) = \frac{N_t}{n_c(r)}$.

4. Estimate the *effective volume* [25] of an equivalent uniform pdf as $\hat{V}_{Ma} = r^d Q(r)$.

5. Estimate the differential entropy as the logarithm of the estimated volume:

$$\hat{h}_{Ma} = d \log_2(r) + \log_2(Q(r)). \tag{4}$$

Note that, according to the definition of $C(r)$, in Eq. 2, it can be related to $Q(r)$ as $C(r) = \frac{1}{Q(r)}$, and Eq. 4 can be rewritten as

$$\hat{h}_{Ma} = d \log_2(r) - \log_2(C(r)). \tag{5}$$

Comparing Eq. 5 to Eq. 3 one may conclude that the h in Eq. 3 is the Ma's entropy estimate, $\hat{h}_{Ma}$. As a consequence, the line fitting procedure explained for the ID estimation can also be used for DE estimation, where slope and y-intercept parameters play the role of ID and DE estimates, respectively. On the other hand, the ambiguity problem mentioned in Section III-A is crafted into a tool that allows for multiscale analysis through a perspective similar to that proposed in [16], where almost linear segments with different angular and linear coefficients give clues regarding the structure of the underlying manifold.

As an illustration of this multiscale analysis, consider again results shown in Figure 3, with two almost linear intervals. The estimated line segments have angular coefficients close to 1 and 2, respectively, associated to DE estimates $h_1 \approx -0.65$, thus close to the theoretical DE of the source, -0.67 bits, and $h_2 \approx 2.7$ bits, which is close to the logarithm of the *ring* length in Figure 1, $\log_2(2\pi) \approx 2.65$ bits.

In other words, the two almost linear segments suggest that (a) at small scales the dominant structure is 2D, with an effective area close to 2^{h_2}, whereas (b) at larger scales the dominant structure becomes roughly 1D, with effective length close to 2^{h_1}.

IV. A METHOD FOR VISUAL ANALYSIS OF ID AND DE

The method proposed here is a straightforward recombination of the approaches explained in Sections III-A and III-B, chosen for their simplicity and data efficiency (for both methods consider all possible pairs of observations). In this recombination, it is assumed that:

- ID is constant over the variable domain.
- Probability density function is locally uniform.

The method is organized in 7 steps. The first 5 steps are presented below, whereas the remaining ones are presented in Subsection IV-A, where the bias problem is addressed.

(S1) Compute the supremum norm for each vector $\mathbf{x}(i) - \mathbf{x}(j)$, $i < j$. Double each norm and store the results in an array $\mathbf{r}$.

(S2) Sort $\mathbf{r}$. Now $\mathbf{r}(k)$ is the edge size of the hypercube that yields k coincidences.

(S3) Plot $\log_2(k/L_r)$ versus $\log_2(\mathbf{r}(k))$, where L_r is the length of the array $\mathbf{r}$ and k ranges from 1 to L_r (optionally, points can be resampled and interpolated for better visualization).

(S4) Plot ID hypotheses $\log_2(\mathbf{r}(k))$ versus $d \log_2(\mathbf{r}(k))$ for some arbitrary $d < D$.

(S5) Visually chose IDs, $\hat{d}$, and DEs, $\hat{h}$, of selected segments of the plot (segments where the slope can be approximated by a constant).

Example: Let $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \mathbf{x}(3), \mathbf{x}(4), \mathbf{x}(5)\}$ be a set of $N = 5$ independent observations of a random source, namely:

$\mathbf{x}(1) = [92, 46, 138]$,
$\mathbf{x}(2) = [4, 2, 7]$,
$\mathbf{x}(3) = [48, 24, 72]$,
$\mathbf{x}(4) = [26, 13, 40]$,
$\mathbf{x}(5) = [41, 21, 62]$.

Supremum norms for all 10 non-redundant observation pairs are computed and multiplied by 2, yielding $\mathbf{r} = [262, 132, 196, 152, 130, 66, 110, 64, 20, 44]$.

These values are sorted in ascending order as:

$\mathbf{r} = [20, 44, 64, 66, 110, 130, 132, 152, 196, 262]$.

Thus, $\mathbf{r}(4) = 66$, for instance, means that a cube of edge 66 around each observation yields 4 coincidences. For this particular value one can compute $C(66)$ as the number of coincidences (4) divided by the total number of pairs (10), yielding the pair $(\log_2(66), \log_2(4/10)) \approx (6.0, -1.3)$ to be plotted.

Proceeding likewise for all values in vector $\mathbf{r}$, the plot in Figure 4 is obtained. Through visual inspection, it is possible to infer that observations roughly lie in an 1D structure, for the candidate with most similar slope in Figure 4 equals one. In other words, although observation are given in $D = 3$, we are able to infer that they lie in manifold whose intrinsic dimension is $d = 1$.

Besides, once d is estimated, the DE can be estimated as the average value of differences $d \log_2(\mathbf{r}(k)) - \log_2 C(\mathbf{r}(k))$. In this example, the differences for three arbitrarily chosen points are 7.4, 7.8 and 7.6, thus yielding an average DE estimate of $\hat{h} = 7.6$ bits.

These estimates for ID and DE suggest that the five observations in this example were sampled from an 1D structure of length $2^{\hat{h}} \approx 194$. Indeed, the $N = 5$ points were uniformly drawn from a noisy linear segment with length $\sqrt{100^2 + 50^2 + 150^2} \approx 187$. Therefore, the ID of the underlying 1D manifold was correctly inferred, while its length was roughly guessed through the estimated DE.

A. Bias compensation

Both ID and DE estimators combined in this work are based on the exponentially growing fraction of patterns randomly coinciding, on average, inside small hypercubes of growing edge. Ideally, this edge should be vanishingly small, but in practice the number of observations is finite, what yields two antagonistic restrictions, namely: that the hypercube size should be as small as possible, thus containing a small fraction of observations, and that this fraction should be as large as possible, for statistical reasons.

In [22] it is shown that ID is always underestimated by Eq. 1 in the simple case of a hypercube inside which the
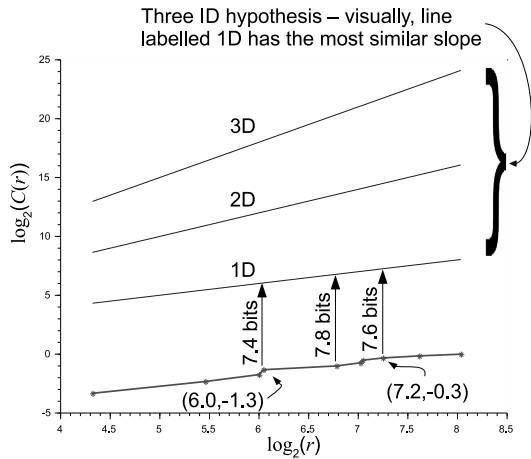
Fig. 4. Plot of ordered pairs $(\log_2(r),\ \log_2 C(r))$. The resulting plot is visually compared to 3 ID hypothesis. The best match is 1D (thus $\hat{d} = 1$), and the average vertical distance from plotted points to the corresponding line yields an estimated $\hat{h} \approx 7.6$ bits.
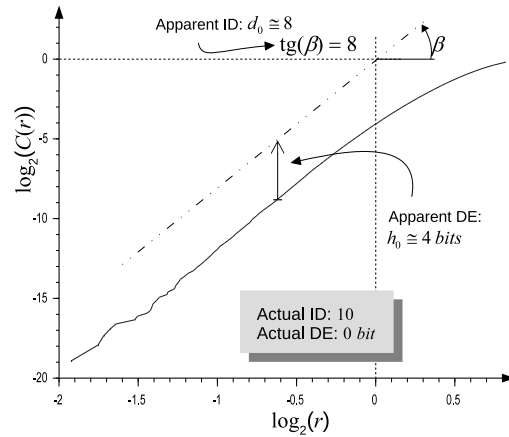

Fig. 5. An instance of biased estimates for ID and DE. The actual ID and DE are 10D and 0 bit, but the visual analysis yields estimates around 8D and 4 bits, respectively. These strong biases are caused by the smallness of the dataset, as compared to its actual ID.

probability density of a pattern being observed is uniform, even for an unlimited amount of data. The equations in [22] that explain this bias are rewritten here as Eq. 6 and 7 for the reader convenience:

$$C_0(r) = (r(2 - r))^d \tag{6}$$

$$d_0(r) = d \times \left(1 - \frac{r}{2 - r}\right) \tag{7}$$

where $C_0(r)$ and $d_0(r)$ stands for theoretical estimates of C and d for an RV uniformly distributed in a d-dimensional hyper-cube of edge r. It is noteworthy that $d_0(r)$ is the derivative of $\log C_0(r)$ with respect to $\log r$.

In [22], under the following arbitrary restrictions:

- R1: $d_0(r) \geq 0.95d$, which imposes an estimate deviation tolerance, and
- R2: the minimum r is $1/4$ of the maximum r, which allows the expected exponential proportionality of Eq. 6 to appear,

it is shown that the minimum number of observations, N_{min}, for a proper ID estimation depends upon the true ID, d, as

$$N_{min} \approx 42^d. \tag{8}$$

This requirement is impractical for most real applications. For instance, even for d as low as 5 an experimenter would need more than 100 million independent samples in order to obtain a good ID estimate. In subsequent works this result was replaced with less restrictive ones such as in [30], where a much simpler analytic model is used, yielding

$$N_{min} \approx 10^{d/2} \tag{9}$$

In spite of their differences, both works agree that small datasets yield false ID estimates, biased toward lower values. For instance, with $N = 1000$ observations independently and uniformly sampled in a 10D hypercube, the visual approach used here yields the result presented in Figure 5, suggesting a wrong ID estimate of about 8D, as well as a wrong DE estimate of about 4 bits (the actual DE is 0 bit).

To predict and compensate for both biases, an approach built upon the analytical model proposed in [22] was developed here. In practical terms, it consists of completing a table of underestimated IDs, for a given N, then using this table to infer the unbiased ID, which in turn allows the estimation of a bias compensation for the DE too.

The above mentioned table is based on Eqs. 6 and 7 and on a coarse estimation of the average supremum distance from an observation to its nearest neighbour, $\bar{r}$, where for N observations over a regular grid in a d-dimensional unit volume hypercube, one should expect

$$\bar{r}(N, d) = \frac{1}{1 + N^{1/d}}. \tag{10}$$

To obtain this average supremum distance first consider a line segment of unit length which is equally split into $n + 1$ intervals, thus allowing the placement of n equally spaced points apart from each other by $r = 1/(1 + n)$. Likewise, in a unit area square, $N = n^2$ points can be regularly arranged by keeping the same r (as a result of the same $n = N^{1/2}$) as the supremum distance between neighboring points. Through the generalization of this simple reasoning for a unit volume hypercube of dimension d, where $N = n^d$ points can be regularly arranged in the vertices of a grid, r remains the supremum distance between any neighboring points of this grid. Therefore, given N and d, there is at least one arrangement of the N points separated from nearest neighbors by $r = 1/(1 + N^{1/d})$. On the other hand, for N points randomly placed inside that same d-dimensional hypercube, the supremum distance between neighboring points is a random variable, say R, but if its underlying probability density function is uniform, one can use Eq. 10 as a coarse approximation of the expected value for R.

This approximation experimentally proved to be useful for $N << 2^d$, which tends to be the case for high ID values, were bias correction is even more relevant. For instance, if $d = 10$ and $N = 100$, the prediction is $\bar{r}(100, 10) \approx 0.38$, which is the same value experimentally obtained up to two decimal

places. Likewise, if $d = 20$ and $N = 10000$, the prediction is $\bar{r}(10000, 20) \approx 0.37$, whereas the experimental value is about 0.39. By contrast, for less sparse datasets, such as for $d = 5$ and $N = 100$, the prediction is $\bar{r}(100, 5) \approx 0.21$, whereas the experimental value is about 0.28.

Applying Eq. 10 to Eq. 7 yields

$$d_0(N, d) = d \times \left(1 - \frac{\bar{r}(N, d)}{2 - \bar{r}(N, d)}\right) \quad (11)$$

By definition [25], a random variable with uniform probability density inside a hypercube of unitary volume has null differential entropy. Therefore, given that Smith's bias is calculated precisely for this random variable, Eq. 4 should yield $h_{Ma} = 0$ in this case, and any imbalance between $\log_2(C_0(r))$ and $d_0(r)\log_2(r)$ is to be taken as an entropy bias, Δh. Therefore, for the estimated d_0 the expected DE bias is

$$\Delta h = \log_2 C_0(r) - d_0(r)\log_2(r) \quad (12)$$

Applying Eq. 6 and 7 to Eq. 12 yields

$$\Delta h = \log_2 (r(2 - r))^d - d\left(1 - \frac{r}{2 - r}\right)\log_2(r)$$

which can be simplified to

$$\Delta h = d\left(\left(\frac{r}{2 - r}\right)\log_2(r) + \log_2 (2 - r)\right) \quad (13)$$

Using Eq. 10 into Eq. 13 yields

$$\Delta h(N, d) =$$

$$d\left(\left(\frac{\bar{r}(N, d)}{2 - \bar{r}(N, d)}\right)\log_2(\bar{r}(N, d)) + \log_2 (2 - \bar{r}(N, d))\right) \quad (14)$$

Finally, to compensate for biases, Steps S1 to S5, as proposed in Section IV, are followed by two more steps, namely:

(S6) Using Eq. 11, find the compensated ID estimate, $\bar{d}$, that yields the closest $d_0(N, d)$ to the visually estimated $\hat{d}$.

(S7) Obtain $\Delta h(N, \bar{d})$ using Eq. 14 and compute a compensated DE estimate as $\bar{h} = \hat{h} - \Delta h(N, \bar{d})$.

Illustration: An experimenter gathered $N = 1000$ multivariate observations with $D = 20$ attributes, and this observer applies the visual method (steps S1 to S5), thus obtaining the solid curve in Figure 5. A naive experimenter would believe that the ID of that data is 8, according to the angle of the dashed line (found after visual comparison between some competing slopes). Lets call it the *apparent ID*, $\hat{d} \approx 8$, associated to the *apparent DE*, $\hat{h} \approx 4$ bits. However, because N is too small as compared to 42^8 [22], or even to $10^{8/2}$ [30], one should not accept the result of this first analysis. Proceeding with step S6, a range of possible IDs near $\hat{d}$ is considered and Eq.11 is used to complete Table I, from which it is possible to infer that the *apparent ID* near 8 corresponds to a bias compensated ID of 10, which is the actual ID of the data source used in this illustration. On step S7, Eq. 14 further yields $\Delta h(1000, 10) \approx 4.2$ and a less biased DE estimate is finally obtained as $\hat{h} - \Delta h(1000, 10) \approx -0.2$ bits (the actual DE of the data source used in this illustration is zero).

TABLE I
BIAS COMPENSATION TABLE FOR FIGURE 5.

d	8	9	**10**	11	12
$d_0(1000, d)$	6.6	7.3	**7.99**	8.69	9.36

V. EXPERIMENTAL RESULTS

Two sets of experimental results are presented. First with two artificial data whose intrinsic dimensions are known, and their corresponding results are presented as evidences in favor of the proposed approach. Those results palliate the difficulty of providing statistical analysis for the method, since it depends upon visual (human) evaluation as part of the process. Two real datasets are analyzed afterwards, and despite the fact that their intrinsic dimensions were already analyzed in former published papers, our results induce some interesting questions regarding estimates consistency and the need for bias compensation.

The first artificial dataset source corresponds to a 12-dimensional manifold ($d = 12$) in 72-dimensions ($D = 72$) first proposed in [27], then reused afterwards in [1] and [31], which makes it a suitable dataset for comparison purposes. $N = 1600$ random data points were used and two results are separately presented in Figures 6 and 7 for a better visualization of an interesting aspects of this dataset. For values of $\log_2(r)$ from -0.3 to 0.1 (fine observation scale), the apparent ID is about 9.4, whereas the apparent DE is about 15 bits, as can be better observed in Figure 6. As for Figure 7, One may observe instead an apparent ID of about 12.2, whereas the apparent DE remains around 15 bits, for values of $\log_2(r)$ from 0.1 to 1.0 (coarse observation scale). The visual comparison to slopes such as 9.4 and 12.2 was induced by the values found in Table II, for integer values of compensated IDs..
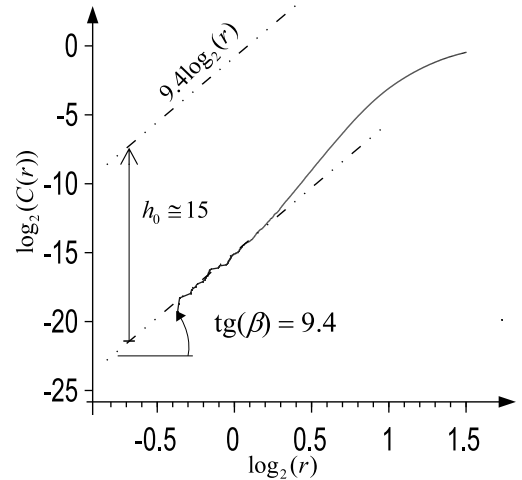

Fig. 6. ID estimation for a 12-dimensional manifold in 72-dimensions proposed in [27]. For values of $\log_2(r)$ from -0.3 to 0.1 (small observation scale), the apparent ID is about 9.4, whereas the apparent DE is about 15 bits, both biased.

Again, proceeding with step S6, a range of possible IDs is considered and Eq.11 is used to complete Table II, from which one may conclude that for fine scales of observation, the actual ID of the corresponding manifold is about 12 (from
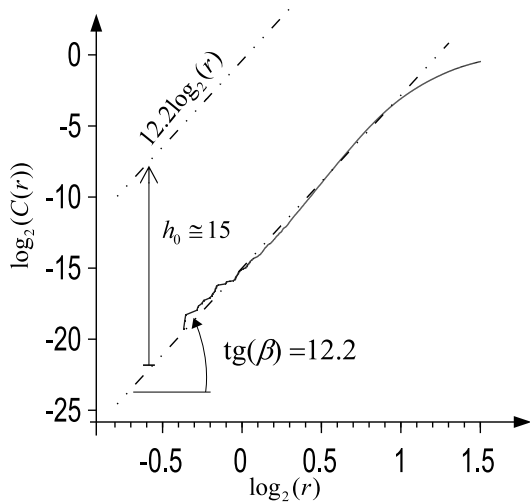
Fig. 7. ID estimation for a 12-dimensional manifold in 72-dimensions proposed in [27]. For values of $\log_2(r)$ from 0.1 to 1.0 (coarse observation scale), the apparent ID is about 12.2, whereas the apparent DE remains around 15 bits.
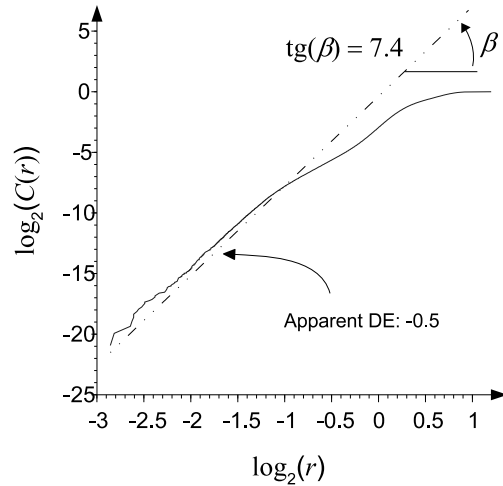

Fig. 8. ID estimation for the dataset labeled "Data Set D". For small values of $log_2(r)$ the apparent ID is about 7.4, whereas the apparent DE is close to zero, about -0.5 bits.

9.4, after bias compensation), whereas its biased DE of about 15 bits should be compensated (step S7) to $\hat{h} - \Delta h(1600, 12) \approx 15 - 4.8 = 10.2$ bits. It is noteworthy that 12 is indeed the artificially imposed ID to the manifold underlying this dataset. Moreover, in [27] it is highlighted that this manifold has a *"high curvature and nontrivial probability measure effects on the manifold"*, and we believe that the second linear trend shown in Figure 7 is a consequence of that high curvature, for the apparent ID of about 12.2 is compensated to 16, which is compatible with the idea that a 12D manifold can be curved to the point that, for a coarse observation scale, it forms a (hollow) structure of dimension higher than 12. The biased DE of such structure is compensated to $\hat{h} - \Delta h(1600, 16) \approx 15 - 5.8 = 9.2$ bits.

TABLE II
BIAS COMPENSATION TABLE FOR FIGURES 6 AND 7.

d	11	**12**	13	14	15	**16**
$d_0(1600, d)$	8.8	**9.4**	10.1	10.8	11.5	**12.2**

The second artificial dataset is labeled "Data Set D" [32], also used in [1] under label "Santa Fe dataset". As explained in [32], it corresponds to a *"relatively long series of known high-dimensional dynamics (...) with weak nonstationarity"* with 100,000 points obtained by numerical integration of the equations of motion for a damped, driven particle. The simulated values were organized in $N = 2000$ 50D patterns, as in [1], which yielded the visual result presented in Figure 8, where an apparent ID of about 7.4 is observed, along with a small apparent bias of about -0.5 bits[1].

Consulting Table III, one may infer a compensated ID of about 9 (the actual ID of this artificial data source), and a corresponding compensated DE of about $\hat{h} - \Delta h(2000, 9) \approx -0.5 - 4.1 = -4.6$ bits.

[1] The visual comparison to slopes such as 6.7, 7.4 and 8.1 was induced by the values found in Table III, for integer values of compensated IDs.

TABLE III
BIAS COMPENSATION TABLE FOR FIGURE 8.

d	7	8	**9**	10	11
$d_0(2000, d)$	6.0	6.7	**7.4**	8.1	8.8

The first real dataset used in this paper is labeled "Paris-14E Parc Montsouris" in [33], corresponding to a time series formed by daily average temperatures (in tenths of Celsius degrees) in Paris, from January 1, 1958 to December 31, 2001. The 15,706 measurements were organized in $N = 785$ patterns of $D = 20$ measurements each. In [33] three ID estimation algorithms were applied to this dataset, including GP, with which the authors of [33] estimated an ID of 4.91.

By contrast, Figure 9 presents our reproduction of the experiment with the Grassberger-Procaccia approach, where for values of $\log_2(r)$ from 5.7 to 6.3 the apparent ID is about 10.7, whereas the apparent DE is about 76.5 bits.
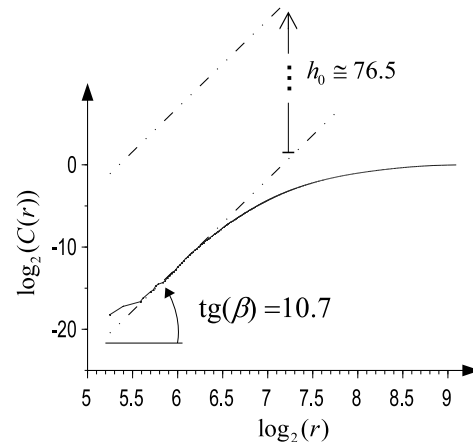

Fig. 9. ID estimation for the dataset labeled "Paris14e Parc Montsouris". For values of $\log_2(r)$ from 5.7 to 6.3 the apparent ID is about 10.7, whereas the apparent DE is about 76.5 bits.

This visual result, even before any bias compensation, suggests that an ID of about 5 is far from any ID value estimated for small values of $\log_2(r)$. We then conjecture that the authors of [33] estimated an average slope for a wide range of $\log_2(r)$, which indeed would yield an ID estimate near 5. Besides, in [1] twelve different ID estimators were applied to this same dataset, yielding inconsistent estimates ranging from 3.71 up to 13.52.

In this work, we assume that the apparent ID of 10.7 in Figure 9 as our best guess for small values of r, whose bias compensation, according to Table IV yields an ID of about 14. Likewise, the corresponding compensated DE is about $\hat{h} - \Delta h(785, 14) \approx 76.5 - 5.1 = 71.4$ bits.

To check this result, an additional analysis similar to that shown in Figure 5 was done, this time with $N = 785$ random observations of a random variable uniformly distributed in a hyper-cube of 14 dimensions, thus with actual ID of 14, and actual DE of 0 bit. In this experiment, the apparent ID and DE were found to be $d_0 = 10.7$ and $h_0 \approx 5.1$ bits, as shown in Figure 10, which seems to confirm that results shown in Figure 9 are compatible with a random source of 14D (apparent ID of about 10.7), to which a bias compensation of about 5.1 bits is necessary. In other words, Figure 10 corroborates the idea that the "Paris14e Parc Montsouris" dataset lies in a 14D (thus greater than 10.7) manifold whose DE is about 71.4 (instead of 76.5) bits.
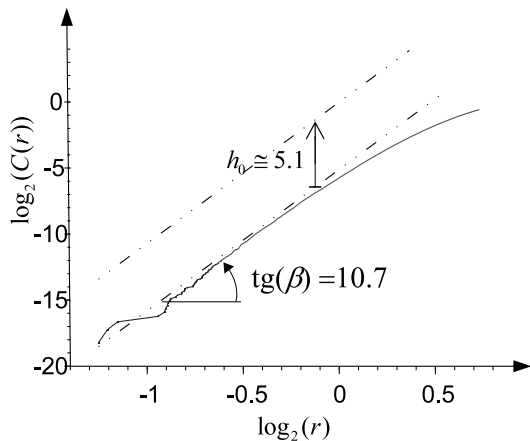

Fig. 10. ID estimation for an artificial 14D dataset of null DE. The apparent ID of about 10.7 is visually compatible with Figure 9, whereas the apparent DE of about 5.1 bits equals the DE bias compensation applied to the "Paris14e Parc Montsouris" dataset.

TABLE IV
BIAS COMPENSATION TABLE FOR FIGURE 9.

d	12	13	**14**	15	16	17
$d_0(785, d)$	9.3	10.0	**10.7**	11.4	12.0	12.7

Another experiment with real data was done with all $N = 6990$ available observations of digits labeled '2' in the MNIST dataset[34], for practical purposes, this dataset was labelled as "MNIST 2". Digit '2' was chosen to allow a comparison of our result to similar experiments reported in [12], [13] and [27]. Figure 11 corresponds to the visual analysis from

this experiment, where an apparent ID of 13 was observed, associated to an apparent DE of about 134 bits.
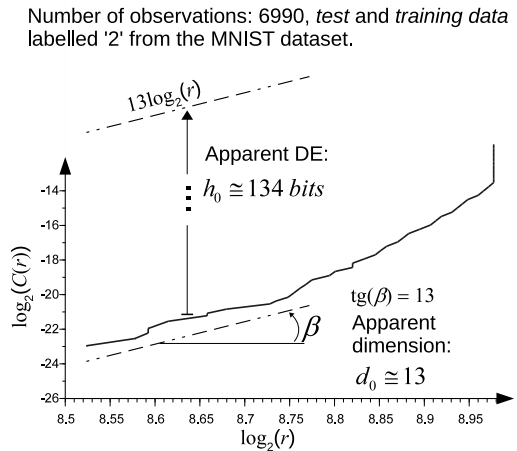

Fig. 11. ID estimation for digit 2 from the MNIST dataset. The apparent ID is about 13, whereas the apparent DE is about 134 bits, both biased.

The visually estimated ID around 13 is in agreement to results presented in [12], [13] and [27], but it seems to be a misleading observation, for the corresponding bias-compensated ID is higher than 13. Indeed, after going through steps S6 and S7, Table V suggests that, for $N = 6990$ observations, an apparent ID of about 13 is expected when the actual ID is 17.

TABLE V
BIAS COMPENSATION TABLE FOR FIGURE 11.

d	13	14	15	16	**17**	18
$d_0(6990, d)$	10.3	11.0	11.7	12.4	**13.1**	13.7

Besides, by assuming that the actual ID is 17, Equation 14 predicts a DE bias of about $\Delta h(6990, 17) = 6.4$ bits, therefore, one may estimate that the actual DE is $h_0 - \Delta h(6990, 17) = 134 - 6.4 \approx 128$ bits. This is less than the DE estimated by [12], of about 145 bits. Such a discrepancy may be partially accounted for the fact that in [12], the estimated DE is the intrinsic Rényi α-entropy for $\alpha = 1/2$, whereas the collision DE ($\alpha = 2$) is estimated here.

As in the former experiment with real datasets, to check our results, an additional analysis was done similar to that shown in Figure 5 with $N = 6990$ random observations of a random variable uniformly distributed in a unit-volume hyper-cube of 17 dimensions, thus with actual ID and DE equal to 17 and 0 bits, respectively. In this experiment, the apparent dimension and entropy were found to be $d_0 = 13$ and $h_0 \approx 6.8$ bits, with a visual aspect quite similar to those presented in Figures 5 and 10. This result seems to confirm our conclusion that "MNIST 2" samples lies in a 17D manifold. However, the bias compensation prediction of about 6.4 slightly deviated from the observed bias of about 6.8 bits, for the artificial data used in the test.

Two more image datasets were included in the experiments, namely: CIFAR-10 [35] and the set of 5,000 images

publicly provided by the COCO Consortium, a subset of images referred to as "2017 Val" in the Consortium website. In both cases, images were converted into 256 gray level representations, and because images in the "COCO" dataset have irregular shapes and resolutions, all 5,000 images were cropped around their centers, thus yielding constant-sized patches of 141×141 pixel. This size corresponds to the biggest square patch that fits into all images. On the other hand, CIFAR-10 is a set of 50,000 tiny images of 32×32 pixels, from which only 5,000 were randomly drawn for the purpose of this experiment. Figures 12 and 13 illustrate some images from each dataset, along with their corresponding analyses.
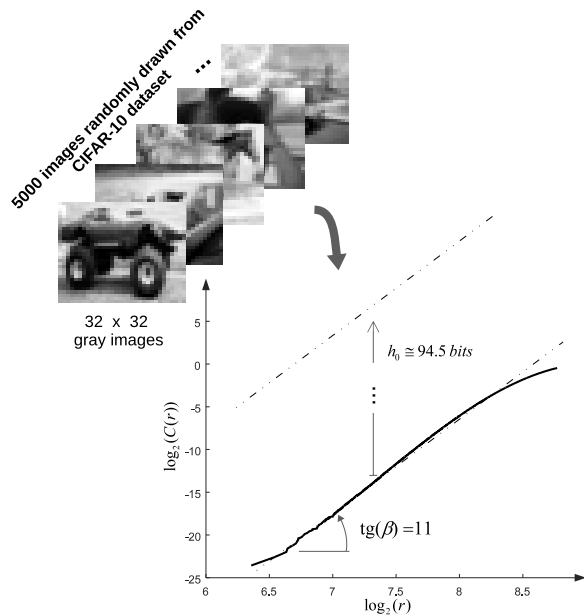

Fig. 12. ID estimation for 5000 images from the CIFAR-10 dataset. The apparent ID is about 11, whereas the apparent DE is about 94.5 bits, both biased.

Despite the much bigger image size in cropped COCO dataset, both visually estimated ID are close to each other. Moreover, the apparent ID of the smaller images from CIFAR-10 is higher than that of COCO, which is probably related to the fact that in CIFAR-10, scenes with target objects are fully represented, whereas many cropped images from COCO have few details, or just parts of objects with almost constant textures. The bias-compensated IDs for those subsets of CIFAR-10 and cropped COCO, according to Table VI, are 14D and 12D, respectively.

TABLE VI
BIAS COMPENSATION TABLE FOR BOTH FIGURES 12 AND 13.

d	10	11	**12**	13	**14**	15
$d_0(5000, d)$	8.2	8.9	**9.6**	10.3	**11.0**	11.7

As for COCO, in Figure 13 one can further notice that at mid-range analysis scale, the apparent ID seems to decrease, before it increases again. It is a curious result, suggesting the formation of low dimensional (possibly elongated) structures,
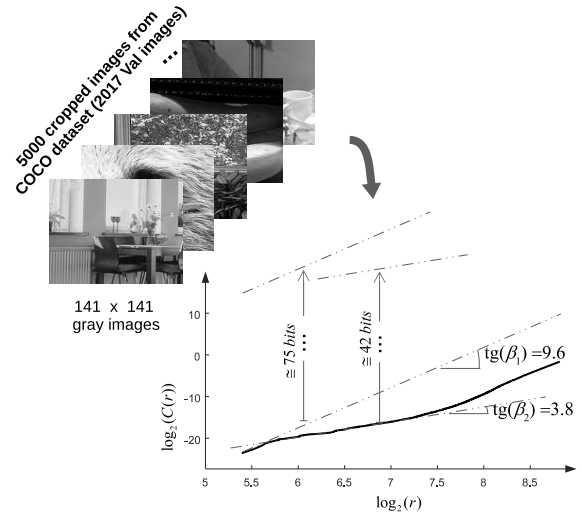

Fig. 13. ID estimation for 5000 images randomly drawn and cropped around their center from the COCO dataset. Two main apparent ID can be observed. The smallest scale yields an apparent ID of about 9.6, for an apparent DE of about 75 bits.

which is analogous to what is illustrated in Figure 3, with synthetic data, where the 2D fine structure of the manifold is masked by the 1D annular structure, in a coarse analysis scale.

VI. CONCLUSION

A new approach for bias-compensated estimation of intrinsic dimension and differential entropy was proposed in this paper. It corresponds to the natural combination of previously published estimation methods, one for *collision entropy* – or quadratic entropy –, by Ma [19], and another for *correlation dimension*, by Grassberger and Procaccia [14]. In the first part of this work it was explained why these two approaches are connected in spite of their different goals, and how ID and DE should be regarded as two complementary aspects of random observations analysis, thus yielding a joint estimation approach.

An important aspect of this approach is its dependency on scale of analysis. Although it is frequently regarded as a practical obstacle for estimators, we propose that estimates at different scales convey different perspectives of underlying manifolds. Accordingly, we propose a pragmatic visual approach, followed by some illustrations.

On the other hand, the seminal work by Smith [22] is a clear warning regarding the always present bias in the Grassberger-Procaccia estimator. Then, we built upon the theoretical model used by Smith to introduce a systematic bias compensation for both ID and DE estimation, whose use is validated through experiments with real data and further illustrated through experiments with synthetic ones.

It is worthy noticing that while this work is strongly based on Smith's analysis, which yields a quite severe restriction on the minimum number of observations for a reliable estimate of d, as pointed out in Eq. 8, that restriction does not apply to this work. Indeed, Smith's analysis imposes that the estimated dimension should not be less than 95% of the actual one,

without any kind of bias compensation. By contrast, in this work, instead of imposing a bias threshold, we use Smith's formula to compensate for that bias, even if the number of observations is much less than $N_{min} \approx 42^d$.

The proposed approach is developed under the assumptions that the ID is constant over the variable domain and that the underlying probability density function is locally uniform. If these assumptions are not verified, the proposed approach should not be applied. Unfortunately, testing whether the intrinsic dimension of observation sets is constant is a problem even harder than estimating its average ID, as proposed in this work, for it demands the partial ID estimation from subsets of observations. Therefore, for non-synthetic data, constant intrinsic dimension should remain an arbitrary assumption. Notwithstanding, thanks to the visual analysis that is an important part of this approach, and taking into account its potential for a geometrical analysis of manifolds as a whole, as proposed in [36], we believe that the study of visual patterns (of log(r) versus log(C)) even when these assumptions are violated can be a promising research subject for the future.

We also believe that the proposed tool for manifold analysis can be useful in pattern recognition context, specially in this renewed era of artificial neural network applications. In this work, we try to provide a pragmatic tool for the bias-compensated estimation of manifold volume and intrinsic dimension. This can be regarded as a first step in understanding how layered processing structures disentangle data manifolds, and how to eventually improve it.

REFERENCES

[1] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli, "Novel high intrinsic dimensionality estimators," *Machine learning*, vol. 89, no. 1-2, pp. 37–65, 2012, doi: 10.1007/s10994-012-5294-7.

[2] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000, doi: 10.1126/science.290.5500.2319.

[3] M. Nilsson and W. B. Kleijn, "On the estimation of differential entropy from data located on embedded manifolds," *IEEE Transactions on Information Theory*, vol. 53, no. 7, pp. 2330–2341, 2007, doi: 10.1109/TIT.2007.899533.

[4] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948, 10.1002/j.1538-7305.1948.tb01338.x.

[5] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?" *Journal of Statistical Physics*, vol. 168, no. 6, pp. 1223–1247, 2017, doi: 10.1007/s10955-017-1836-5.

[6] P. P. Brahma, D. Wu, and Y. She, "Why deep learning works: A manifold disentanglement perspective." *IEEE Trans. Neural Netw. Learning Syst.*, vol. 27, no. 10, pp. 1997–2008, 2016, doi: 10.1109/TNNLS.2015.2496947.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015, doi: 10.1038/nature14539.

[8] S. Yu and J. C. Principe, "Understanding autoencoders with information theoretic concepts," *arXiv preprint arXiv:1804.00057*, 2018, doi : 10.1016/j.neunet.2019.05.003.

[9] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Advances in neural information processing systems*, 2014, pp. 2924–2932.

[10] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, 2016, doi: 10.1016/j.ins.2015.08.029.

[11] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, "Intrinsic dimension estimation: Relevant techniques and a benchmark framework," *Mathematical Problems in Engineering*, vol. 2015, doi: 10.1155/2015/759567.

[12] J. A. Costa and A. O. Hero, "Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets," in *Signal Processing Conference, 2004 12th European*. IEEE, pp. 369–372, 2004.

[13] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," *Scientific reports*, vol. 7, no. 1, p. 12140, 2017, doi: 10.1038/s41598-017-11873-y.

[14] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical review letters*, vol. 50, no. 5, pp. 346–349, 1983, doi: 10.1103/PhysRevLett.50.346.

[15] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in neural information processing systems*, 2005, pp. 777–784.

[16] A. V. Little, Y.-M. Jung, and M. Maggioni, "Multiscale estimation of intrinsic dimensionality of data sets." in *AAAI fall symposium: manifold learning and its applications*, vol. 9, 2009, p. 04.

[17] J. A. Costa and A. O. Hero, "Determining intrinsic dimension and entropy of high-dimensional shape spaces," in *Statistics and Analysis of Shapes*. Springer, 2006, pp. 231–252, doi: 10.1007/0-8176-4481-4_9.

[18] ——, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, 2004, doi: 10.1109/TSP.2004.831130.

[19] S.-k. Ma, "Calculation of entropy from data of motion," *Journal of Statistical Physics*, vol. 26, no. 2, pp. 221–240, 1981, doi: 10.1007/BF01013169.

[20] J. Montalvão, R. Attux, and D. Silva, "A pragmatic entropy and differential entropy estimator for small datasets," *Journal of Communication and Information Systems*, vol. 29, no. 1, 2014, doi : 10.14209/jcis.2014.8.

[21] J. Golay and M. Kanevski, "A new estimator of intrinsic dimension based on the multipoint morisita index," *Pattern Recognition*, vol. 48, no. 12, pp. 4070–4081, 2015, doi: 10.1016/j.patcog.2015.06.010.

[22] L. A. Smith, "Intrinsic limits on dimension calculations," *Physics Letters A*, vol. 133, no. 6, pp. 283–288, 1988, doi: 10.1016/0375-9601(88)90445-8.

[23] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 10, pp. 1404–1407, 2002, doi: 10.1109/TPAMI.2002.1039212.

[24] R. S. Bennett, "Representation and analysis of signals part xxi. the intrinsic dimensionality of signal collections," Johns Hopkins Univ. Baltimore MD Dept. of Electrical and Computer Engineering., Tech. Rep., 1965.

[25] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012, doi: 10.1002/047174882X.

[26] M. O. Hill, "Diversity and evenness: a unifying notation and its consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, 1973, doi: 10.2307/1934352.

[27] M. Hein and Y. Audibert, "Intrinsic dimensionality estimation of submanifolds in euclidean space," in *Proceedings of the 22nd International Conference on Machine Learning*. Max-Planck-Gesellschaft, 2005, p. 289, doi: 10.1145/1102351.1102388.

[28] R. Bennett, "The intrinsic dimensionality of signal collections," *IEEE Transactions on Information Theory*, vol. 15, no. 5, pp. 517–525, 1969, doi: 10.1109/TIT.1969.1054365.

[29] I. Nemenman, "Coincidences and estimation of entropies of random variables with large cardinalities," *Entropy*, vol. 13, no. 12, pp. 2013–2023, 2011, doi: 10.3390/e13122013.

[30] J.-P. Eckmann and D. Ruelle, "Fundamental limitations for estimating dimensions and lyapunov exponents in dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 56, no. 2-3, pp. 185–187, 1992, doi: 10.1016/0167-2789(92)90023-G.

[31] J. Einbeck, Z. Kalantan, and U. Kruger, "Practical considerations on nonparametric methods for estimating intrinsic dimensions of nonlinear data structures," *International journal of pattern recognition and artificial intelligence*, vol. 34, no. 9, pp. 2 058 010–1–2 058 010–35, 2020, doi : 10.1142/S0218001420580100.

[32] N. A. Gershenfeld and A. S. Weigend, "The future of time series," Xerox Corporation, Palo Alto Research Center, No. XEROX-SPL-93-057., Tech. Rep., 1993.

[33] F. Camastra and M. Filippone, "A comparative evaluation of nonlinear dynamics methods for time series prediction," *Neural Computing and Applications*, vol. 18, no. 8, pp. 1021–1029, 2009, doi: 10.1007/s00521-009-0266-y.

[34] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[35] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[36] J. Montalvão, "Noise variance estimation through joint analysis of intrinsic dimension and differential entropy," *Signal Processing Letters*, vol. 26, no. 9, pp. 1330–1333, 2019, doi: 10.1109/LSP.2019.2929437.

Jugurta Montalvão was born in Aracaju, Brazil, in 1968. He received the title of Electrical Engineer (1992) from the University of Campina Grande (UFPB II), Master in Electrical Engineering (1995) from the University of Campinas (UNICAMP) and Doctor in "Automatique et traitement du signal" (2000) from the University Paris-Sud XI. He joined the Department of Electrical Engineering of the Federal University of Sergipe (UFS) in 2005. His main research interests are: pattern recognition and signal processing.

Jânio Canuto was born in Maceió, Brazil, in 1984. He received the title of Electrical Engineer (2007) from the Federal University of Sergipe (UFS), M.Sc. in Electrical Engineering (2010) from the State University of Campinas (UNICAMP) and Ph.D. in Computer Science (2014) from Télécom SudParis. He is currently a postdoc fellow at the Department of Computer Science of the Federal University of Sergipe (UFS). His main research interests are: pattern recognition and signal processing.

Luiz Miranda was born in Recife, Brazil, in 1991. He received the title of Electrical Engineer (2014) from the Federal University of Sergipe (UFS) and M.Sc. in Electrical Engineering (2017) from the Federal University of Sergipe (UFS). His main research interests are: pattern recognition and signal processing.