On intrinsic dimension of point clouds by a persistent homology approach: computational tips

Cinzia Bandiziol[*], Stefano De Marchi[†], Michele Allegra[‡]

November 3, 2025

Abstract

We present new results concerning estimating the *intrinsic dimension (ID)* of point clouds based on persistent homology. In particular, we compare topological ID estimators with different approaches, comprehensively assessing their strengths and weaknesses. We show that a combination of the so-called *i-dimensional persistent homology fractal dimension* estimator and the persistent homology dimension, which we termed *i-dimensional α persistent homology fractal dimension*, is a suitable choice for obtaining an effective estimation of the ID in many benchmark datasets.

Keywords: topological data analysis, persistent homology, intrinsic dimension.

1 Introduction

Due to the increasing availability of large datasets, developing principled and effective approaches to compress the information they contain is becoming a central issue in all applied sciences. In this context, a key role is played by dimensionality reduction approaches aimed at reducing the number of variables (coordinates) in the data with little or no loss of information [61, 21]. What justifies this step is that the data, despite being originally defined in a space with many coordinates, typically lie on a manifold of lower dimension. This dimension, corresponding to the minimum number of local coordinates needed to describe the data, is called *intrinsic dimension* (shortly, ID). Knowing the value of the ID is critical to ensure the reliability low-dimensional data visualization [62] and the validity of dimensionality reduction as a data preprocessing step [35]. In addition, the ID is often a very useful metric per se, allowing the analyst to capture key information about the geometry of the data [2, 39, 1, 20], compare data and models [36], and track temporal variations of complexity [6, 4]. Yet, the ID is generally not known a priori, calling for methods to obtain ID estimates directly from the data.

A wide variety of ID estimation techniques have been advanced in the literature [11, 12]. A majority of the proposed methods fall into one of two categories, that of *projective* methods and that of *geometric-statistical* methods. Projective methods try to project the data onto a space of dimension D, and assess the quality of the projection (in terms of its ability to retain key characteristics of the original dataset) as a function of D [63]. The prototype for these methods is principal component analysis (PCA) [22], which projects the data onto the linear subspace spanned by the first D eigenvectors of the covariance matrix, and uses the fraction of variance within this subspace as a quality metric. A major limitation of projective methods is that they yield a clear ID estimate only when the quality metric exhibits a clear drop below a given D. Typically, this does not occur, and the ID is fixed (rather than estimated) by searching for an optimal compromise between quality and compression (such as retaining the D components explaining 95% of the variance in PCA).

[*]Department of Mathematics, cinzia.bandiziol@phd.unipd.it
[†]Department of Mathematics, stefano.demarchi@unipd.it
[‡]Department of Physics and Astronomy, michele.allegra@unipd.it

Geometric-statistical methods build on the fact that, under broad assumptions, distances between points in the dataset follow statistical relations that depend parametrically on the ID. The prototypes of this class are methods developed in the 1980s to characterize the dimension of strange attractors in the field of dynamical systems. For instance, the correlation dimension [19] is based on the fact that the number of points within small neighborhoods of radius r around a given point scales as $N_r \sim r^D$ for $r \ll 1$, with D being the intrinsic dimension, so that D can be estimated as $\lim_{r\to 0} \frac{N_r}{r}$. More advanced methods is this class follow a similar logic, but make specific assumptions about the probability distribution of the data, such as local uniformity [18, 17] or isotropy [12, 16]. The main limitation of these methods is that they typically require a large number of data points when the ID is high (a facet of the well-known curse of dimensionality problem), and may fail in the presence of highly non-uniform or non-isotropic distributions.
Following the recent surge of topological data analysis [23, 24, 25], several authors have explored *topological* approaches to ID estimation [5], that do not fall within the two classes defined above as they are insensitive to the particular metric chosen to define distances in the data set. In principle, these methods might provide more robust estimates by overcoming some of the previously listed difficulties. Topological ID estimators are based on persistent homology, a popular method for computing topological features of a space at different resolutions [26, 27, 28].
In this work, we set out to compare topological ID estimators with alternative methods, providing a comprehensive assessment of their relative strengths and weaknesses. Upon comparing well-known and well-characterized benchmark datasets, we will focus on real data. One of the fields where ID estimation has risen to prominence in recent years is neuroscience, where it can be used to estimate the dimension of neural activity. The responses of N recorded neurons across time span a *neural manifold* embedded in the $N-$dimensional configuration space of all neurons [15]. While artificial neural networks trained to replicate real brains often display low-dimensional activity [42], biological neural activity is typically high-dimensional activity [13, 14], which has opened a wide debate. Here, we will consider an artificial network trained on simple tasks mimicking those performed by macaques in decision-making experiments [46], and compare the ID of the neural manifold as assessed by traditional methods and topological methods.
The paper is organized as follows: in Section 2 we introduce the basic definitions related to persistent homology, in Section 3 we describe the meaning of Intrinsic Dimension (ID), its importance, and common ways to compute, Section 4 introduces and explains the actual estimators of ID using PH and Section 5 collects all numerical tests that we have run. Finally, in Section 6 we make some conclusions and some future developments.

2 Persistent Homology: basic definitions

Persistent homology (PH) is now widely known and used. Comprehensive treatments are covered in recent textbooks on topological data analysis, such as [24, 25]. Here, we limit ourselves to a brief recapitulation.
Let X be a topological space (for all practical purposes, this can be assumed to be a manifold). The k-th homology group X, $H_k(X)$, consists of the k-dimensional holes of X. The number of connected components (0-dimensional holes), cycles (1-dimensional holes), cavities/voids (2-dimensional holes), and higher-order holes characterize the intrinsic topology space X, providing a qualitative summary of it. In practical applications, one does have access to X, but only to a set of points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1,\dots,N} \subset X$. Persistent homology tries to characterize the X by analyzing the homology of simplicial complexes built on $\mathcal{X}$. Let us briefly recall the basic concepts of *simplicial homology*.

Definition 2.1. A *simplicial complex* $\mathcal{K}$ consists of a set of simplices of different dimensions so that every face of a simplex $\Delta \in \mathcal{K}$ belongs to $\mathcal{K}$ and the non-empty intersection of any two simplices $\Delta_1, \Delta_2 \in \mathcal{K}$ is a face of both Δ_1 and Δ_2.

Given a simplicial complex $\mathcal{K}$, let $\mathcal{S}_k(\mathcal{K})$ denote the set of k-dimensional simplices of $\mathcal{K}$.

Definition 2.2. An *integer valued k-dimensional chain* is a linear combination of k-simplices of $\mathcal{K}$ with coefficients in $\mathbb{Z}$,

$$c = \sum_i a_i \Delta_i \,, \quad \Delta_i \in \mathcal{S}_k(\mathcal{K}), \quad a_i \in \mathbb{Z} \,. \tag{1}$$

Let $\mathcal{C}_k(\mathcal{K})$ denote the set of integer-valued k-dimensional chains of $\mathcal{K}$, which is a group under the operation of addition.

Definition 2.3. The *boundary operator* $\partial_k : \mathcal{S}_k \to \mathcal{C}_{k-1}$ maps an oriented simplex $\Delta \in S_k(K)$ into the $(k-1)$-dimensional chain

$$\partial_k \Delta = \sum_{i=0}^{k} (-1)^k \sigma_i. \tag{2}$$

where σ_i is the $(k-1)$-face obtained by removing the i-the vertex of the simplex (with vertex order fixed by the orientation. The boundary operator can be extended by linearity to a general element of $C_k(\mathcal{K})$, obtaining a map $\partial_k : C_k(\mathcal{K}) \to C_{k-1}(\mathcal{K})$.

Definition 2.4. The kernel of ∂_k is the *group of k-cycles*, $Z_k(\mathcal{K}) := \ker(\partial_k)$. The image of ∂_{k+1} is the group of k-dimensional boundaries, $B_k(\mathcal{K}) := \mathrm{im}(\partial_{k+1})$. Finally, the quotient group $H_k(\mathcal{K}) = Z_k(\mathcal{K})/B_k(\mathcal{K})$ is the k-homology group of $\mathcal{K}$. The generators of H_k are called *homology classes.*

Simplicial complexes can be built on $\mathcal{X}$ by forming simplexes with all points below a certain distance, as follows.

Definition 2.5. Let $(\mathcal{X}, d)$ denote a finite metric space. The *Vietoris-Rips complex* for $\mathcal{X}$, associated to a parameter ϵ and denoted by $\mathcal{V}_\epsilon(\mathcal{X})$, is the simplicial complex where the following hold.

i) $\mathcal{X}$ forms the vertex set;

ii) any subset $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\} \in \mathcal{X}$ spans a k-simplex if and only if $d(\mathbf{x}_i, \mathbf{x}_j) \leqslant 2\epsilon$ for all $0 \leqslant i, j \leqslant k$.

Persistent homology analyzes nested sequences of simplicial complexes arising at increasing values of ϵ, trying to identify the topological features (homology classes) that persist across a wide range of values of ϵ.

Definition 2.6. Let $0 < \epsilon_1 < \cdots < \epsilon_l$ be an increasing sequence of real numbers. A *filtration* is the sequence of sets

$$\emptyset \subset \mathcal{K}_1 \subset K_2 \subset \cdots \subset \mathcal{K}_l \tag{3}$$

with $\mathcal{K}_i = \mathcal{V}_{\epsilon_i}(\mathcal{X})$.

Definition 2.7. The *p-persistent homology group* of K_i is the group

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i)$$

This group contains all stable homology classes in the interval $[i, i+p]$, that is, classes born before step i which are still alive after p steps.

Definition 2.8. Let γ be a homology class in $H_p(\mathcal{K}_i)$. We say that γ is *born* at the instant i if $\gamma \notin H_k^{i-1,1}$, i.e. it cannot be identified with a previously existing class in $H_p(\mathcal{K}_{i-1})$. We say that a homology class born at $\mathcal{K}_i$ *dies* at $\mathcal{K}_{i+p}$ if $\gamma \in H_k^{i,p-1}$ and $\gamma \notin H_k^{i,p}$. Then p is called the *persistence* of γ.

Notice that, highly persistent homology classes typically correspond to topological features of $\mathcal{X}$ (cf. [7]). Hence, during the filtration process, homology classes thus appear and disappear. We can represent classes in $\mathbb{R}^2_+ = \mathbb{R}_{\geq 0} \times \{\mathbb{R}_{\geq 0} \cup \{+\infty\}\}$ by assigning the point (i, j) to a class born at $\mathcal{K}_i$ and died at $\mathcal{K}_j$ (j can take the value $+\infty$, since some features can be alive up to the end of the filtration). Since there can be several independent classes born at $\mathcal{K}_i$ and died at $\mathcal{K}_j$, then each point (i, j) has a *multiplicity*, say $\mu_{i,j}$.

The collections of points (i, j) together with their multiplicity is called a *Persistence Diagram (PD)*.
Figure 1 is an example of a PD collecting features (homology classes) of dimension 0 (in blue), 1 (in orange), and 2 (in green). Points close to the diagonal represent features with a short lifetime (usually associaed with noise) while features away from the diagonal are stable topological features.

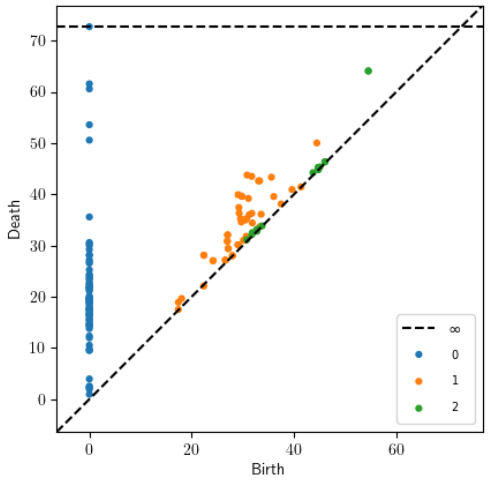

Figure 1: A persistence diagram with classes of dimension 0,1 and 2. At the top of the plot, there is a dashed line that indicates infinity and allows for plotting also couples as $(i, +\infty)$.

3 Intrinsic Dimension Estimation

The minimum number of parameters required to account for the observed properties of data is the *intrinsic (or effective) dimension.* of the data set. The concept of *Intrinsic Dimension (ID)* has become central in all fields dealing with a large amount of data characterized by a high number D of features, such as machine learning [11]. Estimating the ID, which counts the number of essential and fundamental features $d \ll D$, is crucial to uncover the data structure and simplify analysis, enabling data compression.
Following [11], we can define the ID of a dataset as as the minimal dimension of a manifold preserving the information contained in the data set.

Definition 3.1. A dataset $\Omega \subset \mathbb{R}^D$ is said to have *Intrinsic Dimension (ID)* d if its elements lie entirely, without information loss, within a d - dimensional manifold $\mathcal{M}$ of $\mathbb{R}^D$, where $d \ll D$.

3.1 Common fractal ID estimators

Estimating the ID from data is a challenging task. While the first ID estimation algorithm dates back to 1969 [11], the literature now includes a wide variety of estimators (cf. e.g. [12, 16, 17]). Among popular methods we recall *linear* estimators, as the PCA, and the *non linear* ones, such as the kernel PCA. While the linear estimators are based on linear mappings to lower dimensional space of the data, they are particular fast and suitable with data that are "quasi" linearly distributed, but often they overestimate the ID of the manifold. The non linear ones work better with more complex data, providing learning techniques

which are particularly efficient because are based on the preservation of the geometric structure of the original feature space. Unfortunately they are usually more computationally expensive (see for instance the recent paper [37]). A new class are the so-called *fractal-based estimators* which rely on the key idea that distances between points lying on a fractal or a low-dimensional manifold follow scaling laws that depend on the ID [5]. That is the volume of the d-dimensional balls of radius ϵ grows proportional to ϵ^d. The most famous fractal methods are the *box counting* and the *correlation dimension* .
The *box-counting* dimension is based on the number of boxes needed to cover a data set. Let $\mathcal{X}_N \subset \mathbb{R}^D$ be a set of N points of $\mathbb{R}^D$, considered as a metric space. Let $N(\epsilon)$ be the number of boxes (hypercubes) of size ϵ needed to cover $\mathcal{X}_N$. By scaling the box size ϵ, one generally obtains a a power-law scaling of $N(\epsilon)$ [5]:

$$N(\epsilon) \propto \epsilon^{-d_{BC}} \tag{4}$$

Definition 3.2. Let $\mathcal{X}_N$ be a subset of $\mathbb{R}^D$, and let N_ϵ denote the infimum of the number of closed balls of radius ϵ required to cover $\mathcal{X}_N$. The *Box-Counting Dimension* of Ω is

$$d_{BC} = \lim_{\epsilon \to 0} \frac{\log(N_\epsilon)}{\log(1/\epsilon)} \tag{5}$$

provided this limit exists.

If $\mathcal{X}_N$ is an I.I.D. sample of N points from a regular metric μ on $\mathbb{R}^D$, then $\lim_{N\to\infty} d_{BC} = d$.

Another fractal ID estimator that has become one of the most common and widely used is the *correlation dimension (CD)*. It describes how the number of points within a certain distance (or radius) scales with the radius as it increases. Mathematically, it is defined using the *correlation integral* of a given measure μ, which is the probability that pairs of points in a dataset are within a certain distance, say ϵ, of each other [5]:

Definition 3.3. Let $\Omega \subseteq \mathbb{R}^D$ be equipped with a measure μ. Given $\epsilon > 0$, the correlation integral of μ is defined as:

$$C(\epsilon) := \mathbb{E}_{X\sim\mu, Y\sim\mu}\Big[\int_0^\epsilon dr \delta(r - ||X - Y||)\Big] = \mathbb{E}_{X\sim\mu, Y\sim\mu}\big[H(\epsilon - ||X - Y||)\big]. \tag{6}$$

where $||\cdot||$ is a norm (typically, the Euclidean norm), $\delta(x)$ is the delta function, and H is the Heaviside step function ($H(x) = 0$ for $x < 0$, $H(x) = 1$ for $x \geq 0$). The CD is defined as:

$$d_C := \lim_{\epsilon \to 0} \frac{\log(C(\epsilon))}{\log(\epsilon)} \tag{7}$$

The idea behind this definition is that, as $\epsilon \to 0$, the correlation integral scales as

$$C(\epsilon) \propto \epsilon^{d_C} \tag{8}$$

If μ is an absolutely continuous measure on Ω, then $d_C \approx d$.
The CD can be estimated from a finite set of points. Let $\mathcal{X}_N$ be an I.I.D. sample of N points from μ. Let us count the number of pairs of points within distance ϵ,

$$\tilde{C}(N, \epsilon) := \sum_{\substack{x_i, x_j \in \mathcal{X}_N \\ x_i \neq x_j}} H(\epsilon - \|x_i - x_j\|). \tag{9}$$

We then have (cf. e.g. [60]):

$$C(\epsilon) = \lim_{N\to\infty} \tilde{C}(N, \epsilon) \tag{10}$$

In applications, given a finite (large) set of points χ_N, the CD can be estimated as the slope of the log-log plot of $\tilde{C}(N, \epsilon)$ versus ϵ in the limit of small ϵ.

3.2 ID estimators using persistent homology

Several PH-based estimators have been proposed in the literature. For instance, the authors in [5] introduced the *i-Dimensional Persistent Homology Fractal Dimension*, while in [32] they introduced the *Persistent Homology Dimension*, and the *Persistent Homology Complexity*. Assume we have a probability measure μ with support on $\mathcal{M} \subset \mathbb{R}^D$. Let $\mathcal{X}_N$ denote an I.I.D. sample from μ. Let $\mathcal{K}(\mathcal{X}_N)$ denote a Vietoris-Rips complex on $\mathcal{X}_N$. Let γ be a k-dimensional topological feature (persistent homology group generator), i.e., $\exists i,\ \gamma \in H_k^i(\mathcal{K}(\mathcal{X}_N)$, and let $|\gamma|$ denote its persistence.

Definition 3.4. For any $\alpha > 0$,

$$E_\alpha^k(\mathcal{X}_N) := \sum_{\exists i,\ \gamma \in H_k^i(\mathcal{K}(\mathcal{X}_N))} |\gamma|^\alpha \tag{11}$$

The first PH-based ID estimator proposed in the literature is the *k-dimensional Persistent homology Fractal Dimension* (cf. [5]):

Definition 3.5. The *k-dimensional Persistent Homology Fractal Dimension (k-PHFD)* of μ is given by

$$dim_{PH}^k(\mu) := \inf_{d>0} \left\{ \exists\, C(k,\mu,d) \,:\, E_1^k(\mathcal{X}_N) \leq C N^{(d-1)/d} \text{ with probability 1 as } N \to +\infty \right\}. \tag{12}$$

This definition says to us that the dimension may depend on the choice of the filtered simplicial complex (in our case, the Vietoris-Rips) and on the choice of the coefficients for homology computations.

Although a stringent analytical proof is still lacking, numerical tests brought the authors to the following

Conjecture 3.6. Let μ be a nonsingular probability measure on a compact set $X \subseteq \mathbb{R}^D$, $D \geq 2$. Then, for all $0 \leq k < D$, there is a constant $C(k, \mu, D) \geq 0$ such that

$$E_1^k(\mathcal{X}_N) = C N^{(d-1)/d} \tag{13}$$

with probability 1 as $N \to \infty$.

Assuming the validity of this conjecture, taking the logarithm in (13), we get

$$\log(E_1^i(\mathcal{X}_N)) = \log(C) + \frac{d-1}{d} \log(N)\,, \tag{14}$$

which suggests that we can estimate d from the scaling of $\log(E_1^i(\mathcal{X}_N))$. Practically, an estimate of d is obtained by performing a linear regression of $(\log(E_1^i(\mathcal{X}_N))$ as a function of $\log(N))$, taking the slope as an estimate of $\frac{d-1}{d}$, and finally obtaining d through a simple inversion. In applications, the value d can then be inferred from this log-log plot.

Building on this work, in [32] has been introduced the *Persistent Homology Dimension (PHD)*.

Definition 3.7. Let X and μ be as above. For each $k \in \mathbb{N}$ and $\alpha > 0$, the *Persistent Homology Dimension (PHD* is

$$dim_{PH_k^\alpha}(\mu) = \frac{\alpha}{1-\beta} \tag{15}$$

where

$$\beta = \limsup_{N \to +\infty} \frac{\log(\mathbb{E}(E_\alpha^k(\mathcal{X}_n)))}{\log(N)} \tag{16}$$

In practice, it is not so obvious how to treat the expectation value. But analyzing this estimator more carefully, it appears to be an "extension" of k-PHFD. Observing the numerical results in [32], it is clear that the parameter α generally increases the global performance and, consequently, the estimation.

Inspired by the fact that the parameter α gives to the PHD some advantages over i-PHFD, we propose to combine both definitions, providing an estimator that we call *i-dimensional α Persistent Homology Fractal Dimension*, that can easily be defined.

Precisely, the i-dimensional α Persistent Homology Fractal Dimension (i-α-PHFD) of μ is given by

$$dim_{PH}^{i,\alpha}(\mu) = \inf_{d>0} \left\{ d \mid \exists C(i,\mu,D) : E_\alpha^i(X_N) \leq C N^{(d-\alpha)/d} \text{ with probability 1 as N} \to +\infty \right\}. \tag{17}$$

Finally, we recall the definition that measures the complexity of data, known as **Persistent Homology complexity** (cf. [32]). This quantity is an indicator of when the dimension is difficult to compute with the methods presented above.

The starting point is the *cumulative PH_i curve* F_i

$$F_i(X,\epsilon) = \#\{I \in PH_i(X) || I| > \epsilon\}$$

Hence, the PH_i complexity of X is

$$comp_{PH_i}(X) = \lim_{\epsilon \to 0} \frac{-\log(F_i(X,\epsilon))}{\log \epsilon} \tag{18}$$

We notice that $comp_{PH_i}(\mathbb{R}^n) = 0$ for all i. For more details see the survey [32].

4 Numerical Tests

In the present literature, no comparison has been made of the available ID estimators based on PH. Here, we systematically test and compare them on some benchmark datasets and a dataset taken from a computational neuroscience study.

- **Benchmark datasets**. They are commonly tested in the context of ID estimators. They consist of data sampled from *"regular"* manifolds. We collect all the related information in the Table below (see e.g. [12])

Dataset	d	**Description**
Helix	2	2-dimensional helix in $\mathbb{R}^3$
Swiss	2	Swiss-Roll in $\mathbb{R}^3$
Sphere	3	3-dimensional sphere linearly embedded in $\mathbb{R}^4$
NonLinear	4	Nonlinear Manifold in $\mathbb{R}^8$
Affine3d5d	3	Affine space in $\mathbb{R}^5$
Mist	4	Conc. figure, mistakable with a 3-dim. one in $\mathbb{R}^6$
CurvedManifold	12	Nonlinear (highly curved) manifold in $\mathbb{R}^{72}$
NonLinear6d36d	6	Nonlinear manifold in $\mathbb{R}^{36}$

- **Fractal dataset**. This dataset comes from the world of fractals and dynamical systems, in particular the Sierpinski and Ikeda Attractor (see Figure 4), both taken from [32]. Because of the computational burden of computing topological features, here we have computed 4000 points for the Sierpinski Triangle and 5000 for the Ikeda attractor, respectively.

- **Neural activity dataset**. To validate PH-based ID estimation methods, we test them on a real dataset that was extensively characterized with other ID estimation methods.

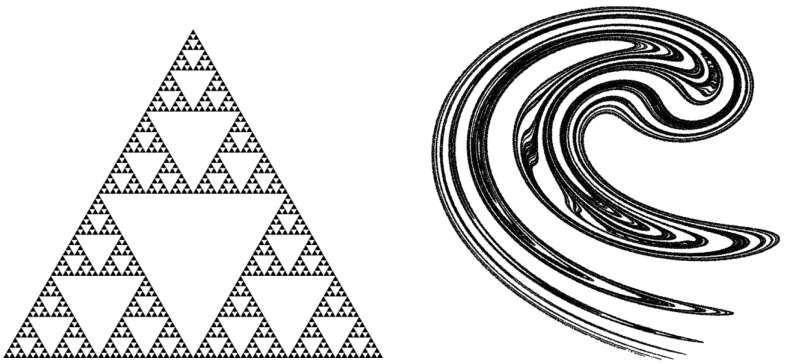

Figure 2: Sierpinski triangle (left) and Ikeda map (right)

In Ref. [cite preprint], the authors characterized the ID of manifolds generated by simulated neural activity. To this aim, they considered artificial recurrent neural networks (RNNs), which are frequently taken as simplified models of real brain networks. In particular, they trained RNNs to perform 20 interrelated tasks mimicking typical tasks in experiments with non-human primates and useful to understand basic cognitive processes such as working memory, inhibition, and context-dependent integration [46]. In each task, the network receives one or two inputs or 'stimuli', both representing an angular variable or direction, and should produce an 'output', also representing a direction. In real experiments, animals are shown dots moving in a specific direction in their left visual field (input 1) and right visual field (input 2) and they should produce a motor response, typically a gaze movement (output) in a specific direction that is a function of the received stimuli. The figure represents the idea of how to use an RNN. We have considered only 3 stimuli, whose data are stored in .csv file with the names Fdgo, Context, and Reactgo_filtered with 25200, 10400 and 5200, respectively, in $\mathbb{R}^{256}$.

comes from a research project still open. Starting from the analysis of activity trajectories of particular recurrent neural networks (RNNs), the aim is to mirror the brain functionality related to basic tasks using an NN. In particular, we considered the RNNs developed by Yang et al. [46] for the study of the properties of the network while performing cognitive tasks. In the literature, the authors trained RNNs to solve simple tasks that mimic typical stimulus-response mapping in experiments with primates. They selected 20 interrelated tasks, useful to understand basic cognitive processes such as working memory, inhibition, and context-dependent integration. In each task, the network receives one or two inputs or 'stimuli', both representing an angular variable or direction (in real experiments, animals are shown dots moving in a specific direction in their left or right eye). In addition, the network receives a 'fixation input', corresponding to the instruction to maintain The gaze is fixed on a cross in the center of the screen. When the fixation input disappears, the network should produce an 'output', representing a motor response. The correct output depends (in more or less simple ways) on the preceding stimuli (in real experiments, animals should move their gaze in a direction that is a function of the received stimuli). The figure represents the idea of how to use an RNN. We have considered only 3 stimuli, whose data are stored in .csv file with the names Fdgo, Context, and Reactgo_filtered with 25200, 10400, and 5200, respectively, in $\mathbb{R}^{256}$.

For the last examples, coming from an open field of research, a priori, the ID is completely unknown. To have only a probably decent idea of which ID is expected, we have decided to compute the related Correlation Dimension that, nowadays, is indeed such a good indicator in practice. For the sake of completeness, we have finally computed it for all datasets and collected the results here in Table 4, where d denotes the ID being approximated.

All codes have been written in Python 3.11 and are available on the GitHub page

`https://github.com/cinziabandiziol/Topological_ID_Estimator`

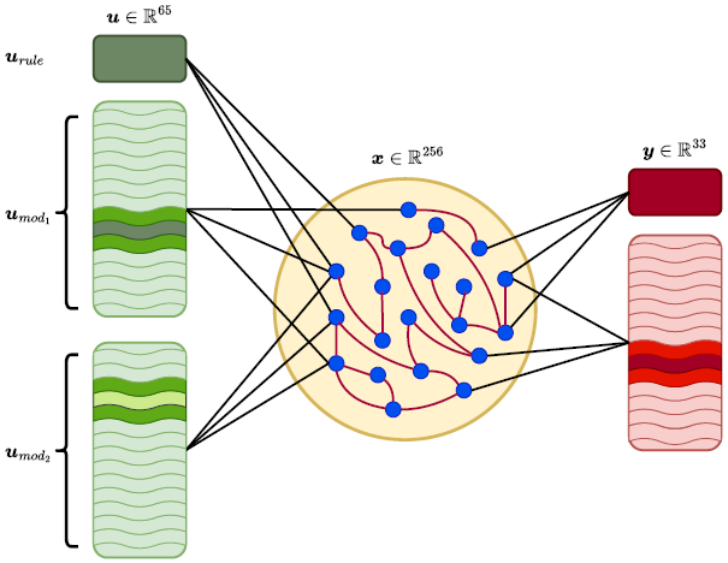

Figure 3: Scheme of the source of data

Dataset	d	Corr. dim.
Helix	2	1.99
Swiss	2	1.98
Sphere	3	2.98
NonLinear	4	3.87
Affine3d5d	3	3.01
Mist	4	3.54
CurvedManifold	12	11.66
NonLinear6d36d	6	5.82
Sierpinski Triangle	1.585	1.585
Ikeda	unk	1.68
Fdgo	unk	1.07
Contextdm1	unk	1.14
Reactgo	unk	2.15

Table 1: Correlation Dimension of all datasets

4.1 Shape Parameter Analysis

Taking the new definition, namely that of i-dim. α PHFD, it depends on the parameter α. Now we are interested in investigating if the choice $\alpha = 1$ turns out to be equal exactly to the i-dim. PHFD, or is there a better choice of the parameter? We have conducted a numerical analysis and, inspired by paper [32], we consider taking α, the parameter to be tested, in the range $(0, 4)$. We have chosen values not so large since the analysis in [32] has taken this direction, which has been revealed to be the most meaningful. For each value of α, we computed the related i-dim. α PHFD as follows.

1. Compute the persistent feature, usually only of dimension $0, 1$.
2. To mirror the limit $n \to \infty$, consider some number n_k closer to the maximum n available (e.g. for the benchmark datasets $n = 10000$).
3. Consider the points $(\log(N_k), \log(E^i_\alpha(\mathcal{X}_N)))$ and compute the linear approximation (we did by using the Python function `numpy.polyfit`): the slope is then equal to $\frac{D-1}{D}$.
4. Make the inverse and obtain the approximated value of D.

We apply this workflow to all datasets with known ID, and we have considered the approximation errors as the differences between the real ID and the approximated one. There exists say, an α^*, such that this error is closer to or equal to zero. We call this the "optimal" one.

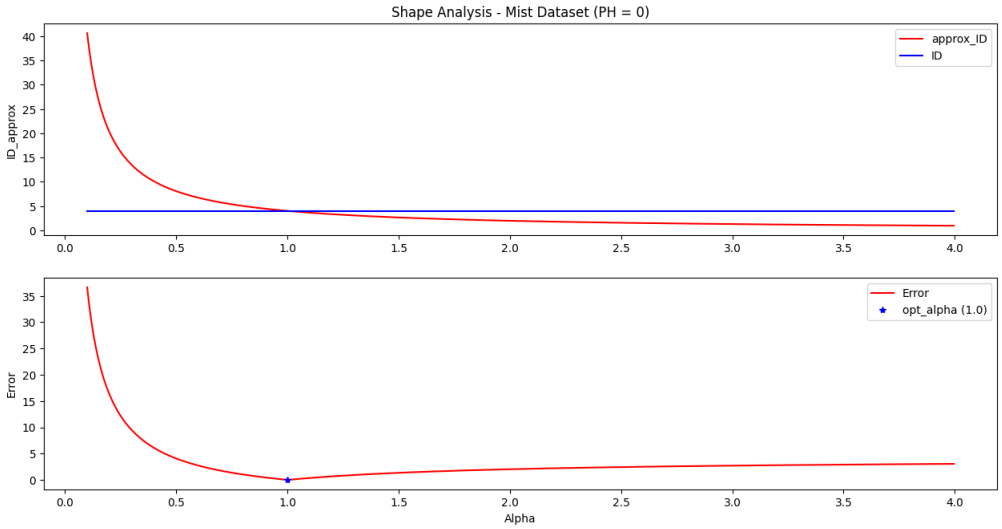

Figure 4: Shape Analysis for the dataset Mist with PH_0

Interestingly, experiments reveal without any doubt that, considering PH_0 and PH_1 both, the optimal value is exactly 1, revealing how the i-dim. PHFD is indeed optimal in α. We report in Figure 4.1 two plots: on the top, the approximation ID (red) and the real ID (blue), while on the bottom, we plot the errors, and we have marked in blue the optimal point. Notably, similar behavior can be seen for all benchmark datasets.

4.2 ID estimates

We report in Table 2 the approximation of ID of all datasets using the "optimal" parameter $\alpha^* = 1$.

Dataset	d	PH_0	PH_1
Helix	2	2.01	2.38
Swiss	2	1.93	2.16
Sphere	3	2.90	3.14
NonLinear	4	3.98	6.45
Affine3d5d	3	2.84	2.91
Mist	4	4.01	6.11
CurvedManifold	12	12.73	-
NonLinear6d36d	6	5.96	9.80
Serpinski	1.58	1.61	1.87
Ikeda Attractor	1.71	2.12	2.13
Reactgo	unk	2.47	2.54
Fdgo	unk	2.14	2.17
Contextdm1	unk	3.07	3.03

Table 2: Computations of $0, 1$-dim. PHFD for all datasets

Remark 4.1. *Some shreds of evidence.*

- *From the results, in general, considering PH_0, the estimator performs better than on estimating PH_1. Besides, if we consider the approximation of ID for the neuroscience*

dataset obtained using Correlation Dimension, again, the results are not so promising. This is a point that needs more investigation and analysis

- *Finally, for what concerns the estimator $comp_{PH}$, we have tried to replicate the results obtained in [32]. Unfortunately, the results seem to be far from the desiderata. We have some doubts about the global definition of the estimator since, as shown in the references, the corresponding function F should have a clear linear behavior in a well-defined interval. Concretely, in our application, we are not able to see it. These conclusions, of course, have given us the curiosity to investigate the definition in depth, and this represents a good direction for future research.*

5 Conclusion

In this paper, we have considered the persistent homology approach as a tool for providing an effective way to estimate the Intrinsic Dimension of a cloud of points. The estimation of Intrinsic Dimension is essential, for example, in quantifying the complexity of the data in terms of the minimal number of dimensions required to capture the data's variance. At the same time, in applications, having methods able to transform the original high-dimensional data into a lower-dimensional representation is crucial. Especially in Machine Learning and Data Analysis, it can be helpful to improve model performance, reduce computation time, and mitigate the curse of dimensionality.

As claimed in Section §4, by Persistent Homology (PH), and inspired by the promising result of UMAP, we provided a comparison of various estimators for the ID. In particular, we focused on *i-Dimensional Persistent Homology Fractal Dimension* [5], *Persistent Homology Dimension* and *Persistent Homology Complexity* [32].

We then decided to combine both definitions, to hopefully obtain another good and interesting estimator that we called *i-dim. α Persistent Homology Fractal Dimension*. The choice of the parameter $\alpha = 1$ proved to be the optimal choice for estimating the corresponding ID for almost all the benchmark datasets and the features PH_0, PH_1. For some datasets, like Ikeda attractor, the ID estimated is far from the expected one. Also, for some datasets and using other estimators, like the CD or the $compr_{PH}$, the results are not promising. These issues should be investigated more deeply in future work.

Acknowledgments. This work has been done within the *Approximation Theory and Applications* topical group of the Italian Mathematical Union, RITA "Italian Network on Approximation", and the INdAM-GNCS group. This work has the support of PRIN 2023-2025, Computational mEthods for Medical Imaging (CEMI). We also thank the Padova Neuroscience Center for the opportunity to use neuroscience data.

References

[1] Aghajanyan, A., Zettlemoyer, L., & Gupta, S. (2020). Intrinsic dimensionality explains the effectiveness of language model fine-tuning. arXiv preprint arXiv:2012.13255.

[2] Ansuini, A., Laio, A., Macke, J. H., & Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. Advances in Neural Information Processing Systems, 32.

[3] T. Birdal, L. Guibas, A. Lou and U. Simsekli (2021), Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

[4] Biondo, M., Cirone, N., Valle, F., Lazzardi, S., Caselle, M., & Osella, M. (2024). The intrinsic dimension of gene expression during cell differentiation. bioRxiv, 2024-08.

[5] Adams H.; Aminian M.; Farnell E.; Kirby M.; Peterson C.; Mirth J.; Neville R.; Shipman P.; Shonkwiler C. A Fractal Dimension for Measures via Persistent Homology. (eds) Topological Data Analysis *Abel Symposia Springer* **2020**, *15*, ISBN: 978-3-030-43407-6

[6] Allegra, M., Facco, E., Denti, F., Laio, A., & Mira, A. (2020). Data segmentation based on the local intrinsic dimension. Scientific reports, 10(1), 16449.

[7] Ali D.; Asaad A.; Jimenez M.; Nanda V.; Paluzo-Hidalgo E.; Soriano-Trigueros M. A Survey of Vectorization Methods in Topological Data Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45, 14069–14080*

[8] Asaad A.; Ali D.; Majeed T.; Rashid R. Persistent Homology for Breast Tumor Classification Using Mammogram Scans. *Mathematics* **2022**, *10, 21*

[9] Brüel-Gabrielsson R.; Ganapathi-Subramanian V.; Skraba P.; Guibas L.J. Topology-Aware Surface Reconstruction for Point Clouds. *Computer Graphics Forum* **2020**, *39, 197–207*

[10] Bukkuri A.; Andor N.; Darcy I. K. Applications of Topological Data Analysis in Oncology. *Front. Artif. Intell.* **2021**, 4, 659037

[11] Camastra F.; Staiano A. Intrinsic dimension estimation: Advances and open problems. *Information Sciences* **2016**, *328, 26–41*

[12] Campadelli P.; Casiraghi E.; Ceruti C.; Rozza A. Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Information Analysis of High-Dimensional Data and Applications* **2015**

[13] Altan, E., Solla, S. A., Miller, L. E., & Perreault, E. J. (2021). Estimating the dimensionality of the manifold underlying multi-electrode neural recordings. PLoS computational biology, 17(11), e1008591.

[14] Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. Proceedings of the National Academy of Sciences, 119(43), e2200800119.

[15] Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Current opinion in neurobiology, 70, 113-120.

[16] Erba, V., Gherardi, M., & Rotondo, P. (2019). Intrinsic dimension estimation for locally undersampled data. Scientific reports, 9(1), 17133.

[17] Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific reports, 7(1), 12140.

[18] Levina, E., & Bickel, P. (2004). Maximum likelihood estimation of intrinsic dimension. Advances in neural information processing systems, 17.

[19] Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. Physica D: nonlinear phenomena, 9(1-2), 189-208.

[20] Li, C., Farkhoor, H., Liu, R., & Yosinski, J. (2018). Measuring the intrinsic dimension of objective landscapes. arXiv preprint arXiv:1804.08838.

[21] Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8(3), 2663-2693.

[22] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[23] Carlsson G. Topology and data. *Bulletin of the American Mathematical Society* **2009**, *46, 255–308*

[24] Carlsson G.; Vejdemo-Johansson M. Topological Data Analysis with Applications. *Publisher: Cambridge University Press*, **2022**, ISBN: 978-110-883-865-8

[25] Chazal F.; Michel B. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Front. Artif. Intell.* **2021**, *4*

[26] Cohen-Steiner D.; Edelsbrunner H.; Harer J. Stability of persistence diagrams. *Discrete & computational geometry* **2007**, *37, 103–120*

[27] Edelsbrunner H.; Harer J. Persistent homology - a survey. *Contemporary Mathematics* **2008**, *453, 257–282*

[28] Edelsbrunner H.; Harer J. Computational topology: An introduction. *Publisher: American Mathematical Society*, **2010**

[29] Flammer M. Persistent Homology-Based Classification of Chaotic Multi-variate Time Series: Application to Electroencephalograms. *SN COMPUTER SCIENCE* **2024**, *5, 107*

[30] Fomenko A.T. Visual geometry and topology *Publisher: Springer Science and Business Media*, **2012**

[31] Guillemard M.; Iske A. Interactions between kernels, frames and persistent homology. *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science, Springer* **2017**, *861-888*

[32] Jaquette J.; Schweinhart B. Fractal dimension estimation with persistent homology: A comparative study. *Communications in Nonlinear Science and Numerical Simulation* **2020**, *84, 105163*

[33] Lawson A.; Chung Y.-M.; Cruse W. A Hybrid Metric based on Persistent Homology and its Application to Signal Classification. *2020 25th International Conference on Pattern Recognition (ICPR)* **2020**

[34] Leykam D.; Angelakis D. G. Topological data analysis and machine learning. *Advances in Physics* **2023**, 8(1)

[35] Obaid, H. S., Dheyab, S. A., & Sabry, S. S. (2019, March). The impact of data preprocessing techniques and dimensionality reduction on the accuracy of machine learning. In 2019 9th annual information technology, electromechanical engineering and microelectronics conference (iemecon) (pp. 279-283). IEEE.

[36] Facco, E., Pagnani, A., Russo, E. T., & Laio, A. (2019). The intrinsic dimension of protein sequence evolution. PLoS computational biology, 15(4), e1006767.

[37] Kadir Özçoban and Murat Manguoğlu and Emrullah Fatih Yetkin, A Novel Approach for Intrinsic Dimension Estimation, ArXiv 2025, `https://arxiv.org/abs/2503.09485`,

[38] Pedregosa F.; Varoquaux G.; Gramfort A.; Michel V.; Thirion B.; Grisel O.; Blondel M.; Prettenhofer P.; Weiss R.; Dubourg V.; Vanderplas J.; Passos A.; Cournapeau D.; Brucher M.; Perrot M.; Duchesnay E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12, 2825-2830*

[39] Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., & Goldstein, T. (2021). The intrinsic dimension of images and its impact on learning. arXiv preprint arXiv:2104.08894.

[40] Rotman J. J. *An introduction to Algebraic Topology*; *Publisher: Springer* **1988**

[41] Saul N.; Tralie C. Scikit-tda: Topological data analysis for Python. Available online: https://doi.org/10.5281/zenodo.2533369 (**2019**)

[42] Sussillo, D., & Barak, O. (2013). Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. Neural computation, 25(3), 626-649.

[43] Skaf Y.; Laubenbacher R. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics* **2022**, 130, 104082

[44] Som A.; Choi H.; Ramamurthy K. N.; Buman M. P.; Turaga P. PI-Net: A Deep Learning Approach to Extract Topological Persistence Images. *ArXiv* **2020**, *submitted*

[45] Tralie C.; Saul N.; Bar-On R. Ripser.py: A lean persistent homology library for Python. *The Journal of Open Source Software* **2018**

[46] Yang G.R.; Joglekar M.R.; Song H.F.; Newsome W.T.; Wang X.J. Task representations in neural networks trained to perform many cognitive tasks *Nature neuroscience* **2019**, *22 (2), 297-306*

[47] Zomorodian A.; Carlsson G. Computing Persistent Homology. *Discrete Comput. Geom.* **2005** *33, 249–274*

[48] The GUDHI Project, GUDHI User and Reference Manual, 3.5.0 Edition, GUDHI Editorial Board. Available online: https://gudhi.inria.fr/doc/3.5.0/ (**2022**)

[49] Giotto-tda 0.5.1 documentation. Available online: https://giotto-ai.github.io/gtda-docs/0.5.1/library.html (**2021**)

[50] Fernández A.; García S.; Galar M.; Prati R.C.; Krawczyk B.; Herrera F. Learning from Imbalanced Data Sets. *Publisher: Springer Nature Switzerland*, **2018**, ISBN: 978-3-319-98073-7

[51] Majumdar S.; Laha A.K. Clustering and classification of time series using topological data analysis with applications to finance. *Expert Systems with Applications* **2020**, *162, 113868*

[52] Moon C.; Li Q.; Xiao G. Using persistent homology topological features to characterize medical images: Case studies on lung and brain cancers. *Ann. Appl. Stat.* **2023**, 17

[53] Moroni D.; Pascali M.A. Learning topology: bridging computational topology and machine learning. *Pattern Recognition and Image Analysis* **2021**, *31, 443-453*

[54] Pun. C. S.; Lee S. X.; Xia K. Persistent-homology-based machine learning: a survey and a comparative study. *Artif Intell Rev* **2022** 55, 5169–5213

[55] Kindelan R.; Frías J.; Cerda M.; Hitschfeld N. A topological data analysis based classifier. *Advances in Data Analysis and Classification* **2024**, Issue 2/2024

[56] Sonego P.; Pacurar M.; Dhir S.; Kertész-Farkas A.; Kocsor A.; Gáspári Z.; Leunissen J.A.M.; Pongor S. A Protein Classification Benchmark collection for machine learning. *Nucleic Acids Research* **2006**

[57] Townsend J.; Micucci C.P.; Hymel J. H.; Maroulas V.; Vogiatzis K. D. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat. Commun* **2020**, *11, 3230*

[58] Shawe-Taylor J.; Cristianini N. Kernel Methods for Pattern Analysis. *Publisher: Cambridge University Press*, **2009**, ISBN: 978-051-180-968-2

[59] Scholkopf B.; Smola A.J. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. *The MIT Press* **2002**, ISBN: 978-026-225-693-3.

[60] James Theiler. Estimating fractal dimension. *JOSA A*, 7(6):1055–1073, 1990.

[61] Van Der Maaten, L., Postma, E. O., & Van Den Herik, H. J. (2009). Dimensionality reduction: A comparative review. Journal of machine learning research, 10(66-71), 13.

[62] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[63] Verveer, P. J., & Duin, R. P. W. (1995). An evaluation of intrinsic dimensionality estimators. IEEE Transactions on pattern analysis and machine intelligence, 17(1), 81-86.