MMM
MODERN MATHEMATICAL METHODS

Research Article

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

CINZIA BANDIZIOL, STEFANO DE MARCHI*, AND MICHELE ALLEGRA

ABSTRACT. We present new results on estimating the intrinsic dimension (ID) of point clouds using 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.

2020 Mathematics Subject Classification: 14F45, 55-04.

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 that reduce the number of variables (coordinates) in the data with little or no loss of information [8, 33, 60]. This step is justified because, although the data are initially defined in a high-dimensional space, they typically lie on a lower-dimensional manifold. 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 of low-dimensional data visualization [61] and the validity of dimensionality reduction as a data preprocessing step [43]. In addition, the ID is often a valuable metric in its own right, allowing the analyst to capture key information about the data geometry [2, 6, 39, 46], compare data and models [23], and track temporal variations in complexity [4, 9]. 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 [13, 14]. A majority of the proposed methods fall into one of two categories: projective and geometric-statistical. 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 [62]. The prototype for these methods is principal component analysis (PCA) [34], 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 (e.g., by retaining the D components that explain 95% of the variance in PCA).

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 dynamical systems. For instance, the correlation dimension [28] 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 in this class follow a similar logic, but make specific assumptions about the probability distribution of the data, such as local uniformity [22, 37] or isotropy [14, 21]. 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 [15, 16, 17], several authors have explored topological approaches to ID estimation [1, 63], 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 difficulties previously listed. Topological ID estimators are based on persistent homology, a popular method for computing topological features of a space at different resolutions [18, 19, 20].

In this work, we compare topological ID estimators with alternative methods to provide 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 [32]. While artificial neural networks trained to replicate real brains often exhibit low-dimensional activity [56], biological neural activity is typically high-dimensional [5, 55], which has sparked a wide debate. Here, we will consider an artificial network trained on simple tasks mimicking those performed by macaques in decision-making experiments [64], 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, and we introduce and explain the actual estimators of ID using PH Section 4 collects all numerical tests that we have run. Finally, in Section 5, we draw some conclusions and discuss 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 [16, 17]. 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 $\mathbb{Z}$,*

$$(2.1) \qquad c = \sum_i a_i \Delta_i, \quad \Delta_i \in \mathcal{S}_k(\mathcal{K}), \quad a_i \in \mathbb{Z}.$$

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*

$$(2.2) \qquad \partial_k \Delta = \sum_{i=0}^{k} (-1)^k \sigma_i,$$

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 $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_k\} \in \mathcal{X}$ spans a k-simplex if and only if $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant 2\epsilon$ 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*

$$(2.3) \qquad \emptyset \subset \mathcal{K}_1 \subset K_2 \subset \cdots \subset \mathcal{K}_l$$

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. [3]). 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 associated with noise) while features away from the diagonal are stable topological features.
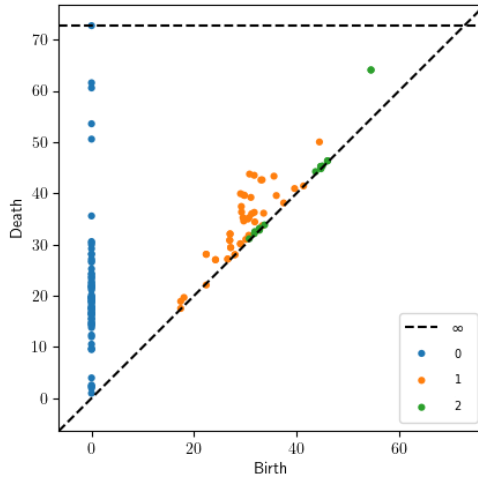

FIGURE 1. A persistence diagram with classes of dimension 0, 1 and 2. At the top of the plot, there is a dashed line that indicates infinity and allows for plotting also couples as $(i, +\infty)$

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 [13]. Estimating the ID, which counts the number of essential and fundamental features $d \ll D$, is crucial for uncovering the data structure and simplifying analysis, enabling data compression. Following [13], we can define the ID of a dataset as the minimal dimension of a manifold preserving the information contained in the dataset.

Definition 3.9. *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 [13], the literature now includes a wide variety of estimators (cf. e.g. [14, 21, 22]). Among popular methods, we recall linear estimators, as the PCA, and the non-linear ones, such as the kernel PCA. While linear estimators are based on linear mappings to a lower-dimensional space of the data, they are fast and suitable for data that are "quasi" linearly distributed, but they often overestimate the ID of the manifold. The non-linear ones work better with more complex data, providing learning techniques that are particularly efficient because they preserve the geometric structure of the original feature space. Unfortunately, they are usually more computationally expensive (see, for instance, the recent paper [44]). A new class is 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 [1]. 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 power-law scaling of $N(\epsilon)$ [1]:

$$(3.4) \qquad N(\epsilon) \propto \epsilon^{-d_{BC}}.$$

Definition 3.10. *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*

$$(3.5) \qquad d_{BC} = \lim_{\epsilon \to 0} \frac{\log(N_\epsilon)}{\log(1/\epsilon)}$$

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 [1]:

Definition 3.11. *Let $\Omega \subseteq \mathbb{R}^D$ be equipped with a measure μ. Given $\epsilon > 0$, the correlation integral of μ is defined as:*

$$(3.6) \qquad C(\epsilon) := \mathbb{E}_{X \sim \mu, Y \sim \mu} \left[\int_0^\epsilon dr \delta(r - ||X - Y||) \right] = \mathbb{E}_{X \sim \mu, Y \sim \mu} \left[H(\epsilon - ||X - Y||) \right].$$

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:

$$(3.7) \qquad d_C := \lim_{\epsilon \to 0} \frac{\log(C(\epsilon))}{\log(\epsilon)}.$$

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

$$(3.8) \qquad C(\epsilon) \propto \epsilon^{d_C}$$

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,

$$\text{(3.9)} \qquad \tilde{C}(N,\epsilon) := \sum_{\substack{x_i, x_j \in \mathcal{X}_N \\ x_i \neq x_j}} H(\epsilon - \|x_i - x_j\|).$$

We then have (cf. e.g. [57]):

$$\text{(3.10)} \qquad C(\epsilon) = \lim_{N \to \infty} \tilde{C}(N,\epsilon).$$

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 [1] introduced the i-Dimensional Persistent Homology Fractal Dimension. In contrast, in [31] 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 3.12. *For any $\alpha > 0$,*

$$\text{(3.11)} \qquad E_\alpha^k(\mathcal{X}_N) := \sum_{\exists i, \, \gamma \in H_k^i(\mathcal{K}(\mathcal{X}_N))} |\gamma|^\alpha.$$

The first PH-based ID estimator proposed in the literature is the k-dimensional Persistent Homology Fractal Dimension (cf. [1]):

Definition 3.13. *The k-dimensional Persistent Homology Fractal Dimension (k-PHFD) of μ is given by*
(3.12)
$$dim_{PH}^k(\mu) := \inf_{d > 0} \left\{ \exists \, C(k, \mu, d) \, : \, E_1^k(\mathcal{X}_N) \leq CN^{(d-1)/d} \text{ with probability 1 as } N \to +\infty \right\}.$$

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 3.1. *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*

$$\text{(3.13)} \qquad E_1^k(\mathcal{X}_N) = CN^{(d-1)/d}$$

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

Assuming the validity of this conjecture, taking the logarithm in (3.13), we get

$$\text{(3.14)} \qquad \log(E_1^i(\mathcal{X}_N)) = \log(C) + \frac{d-1}{d} \log(N),$$

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 [31] has been introduced the Persistent Homology Dimension (PHD).

Definition 3.14. *Let X and μ be as above. For each $k \in \mathbb{N}$ and $\alpha > 0$, the Persistent Homology Dimension (PHD) is*

$$(3.15) \qquad \qquad dim_{PH_k^\alpha}(\mu) = \frac{\alpha}{1-\beta}$$

where

$$(3.16) \qquad \qquad \beta = \limsup_{N \to +\infty} \frac{\log(\mathbb{E}(E_\alpha^k(\mathcal{X}_n)))}{\log(N)}.$$

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 [31], 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

$$(3.17)$$
$$dim_{PH}^{i,\alpha}(\mu) = \inf_{d>0} \left\{ d \mid \exists C(i,\mu,D) : E_\alpha^i(X_N) \leq CN^{(d-\alpha)/d} \text{ with probability 1 as N} \to +\infty \right\}.$$

Finally, we recall the definition of the complexity of data, known as Persistent Homology Complexity (cf. [31]). This quantity indicates when the dimension is difficult to compute using the methods presented above. The starting point is the cumulative PH_i curve F_i

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

Hence, the PH_i complexity of X is

$$(3.18) \qquad \qquad comp_{PH_i}(X) = \lim_{\epsilon \to 0} \frac{-\log(F_i(X,\epsilon))}{\log \epsilon}.$$

We notice that $comp_{PH_i}(\mathbb{R}^n) = 0$ for all i. For more details, see the survey [31].

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 across benchmark datasets and a dataset 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. [14]).

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 [31]. Because of the computational burden of computing topological features, we computed 4000 points for the Sierpinski Triangle and 5000 for the Ikeda attractor.
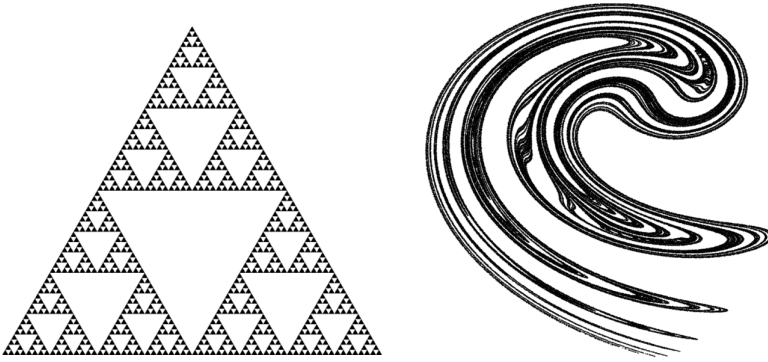

FIGURE 2. Sierpinski triangle (left) and Ikeda map (right)

- **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. In [64], 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. 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 illustrates 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, the ID is a priori completely unknown. To have only a probable decent idea of which ID is expected, we have decided to compute the related Correlation Dimension that, nowadays, is indeed such a good indicator
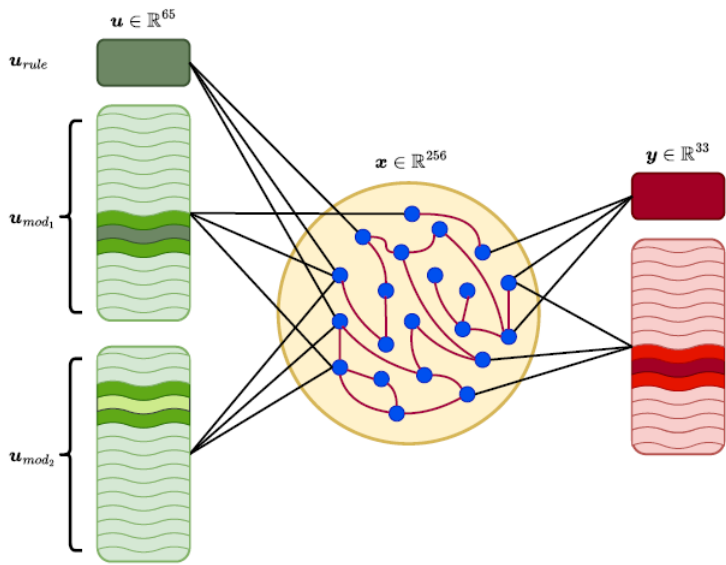
FIGURE 3. Scheme of the source of data

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

in practice. For completeness, we have finally computed it for all datasets and collected the results 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

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 [31], 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 [31] 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 $(\log(N_k), \log(E_\alpha^i(\mathcal{X}_N)))$ 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 IDs, and we treat 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.
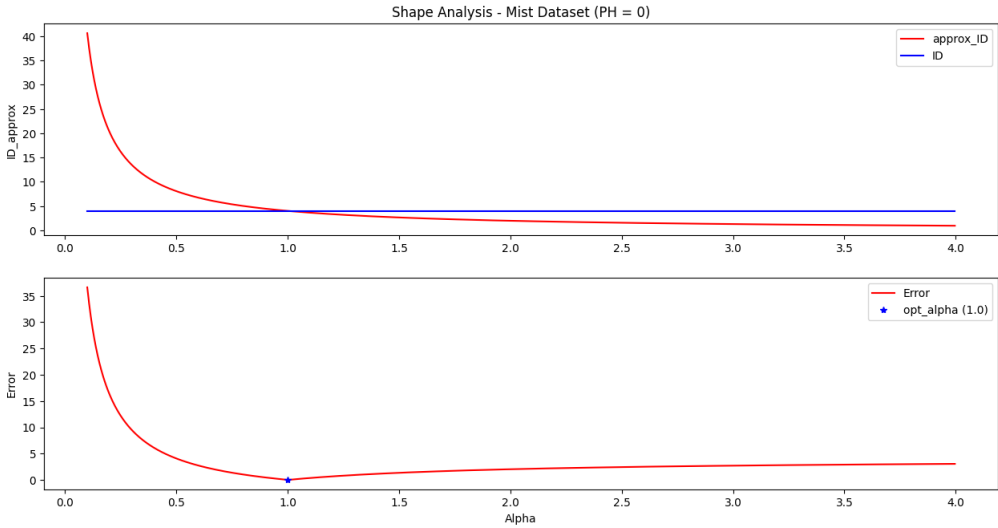

FIGURE 4. Shape Analysis for the dataset Mist with PH_0

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); on the bottom, we plot the errors, and we mark in blue the optimal point. Notably, this behavior is observed across all benchmark datasets.

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

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.*

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

- *Finally, concerning the estimator $comp_{PH}$, we have tried to replicate the results obtained in [31]. 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 exhibit clear linear behavior over a well-defined interval. Concretely, in our application, we are not able to see it. These conclusions, of course, have sparked our curiosity to investigate the definition in depth, which represents a promising direction for future research.*

5. CONCLUSION

In this paper, we have considered the persistent homology approach as an effective tool for estimating the Intrinsic Dimension of a cloud of points. The estimation of Intrinsic Dimension is essential, for example, for quantifying the complexity of the data in terms of the minimal number of dimensions required to capture its variance. At the same time, in applications, methods that can transform the original high-dimensional data into a lower-dimensional representation are 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, using Persistent Homology (PH) and inspired by the promising results of UMAP, we compared various estimators for the ID. In particular, we focused on i-Dimensional Persistent Homology Fractal Dimension [1], Persistent Homology Dimension and Persistent Homology Complexity [31].

We then decided to combine both definitions, to hopefully obtain another good and interesting estimator that we called i-dimensional α Persistent Homology Fractal Dimension. The choice of $\alpha = 1$ proved optimal for estimating the corresponding ID across almost all 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

REFERENCES

[1] H. Adams, M. Aminian, E. Farnell, M. Kirby, C. Peterson, J. Mirth, R. Neville, P. Shipman and C. Shonkwiler: *A fractal dimension for measures via persistent homology*, Topological Data Analysis: Abel Symposia Springer, Switzerland (2020).

[2] A. Aghajanyan, S. Gupta and L. Zettlemoyer: *Intrinsic dimensionality explains the effectiveness of language model fine-tuning*, In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (2021), 7319–7328.

[3] D. Ali, A. Asaad, M. Jimenez, V. Nanda, E. Paluzo-Hidalgo and M. Soriano-Trigueros: *A survey of vectorization methods in topological data analysis*, IEEE Trans. Pattern Anal. Mach. Intell., **45** (2023), 14069–14080.

[4] M. Allegra, E. Facco, F. Denti, A. Laio and A. Mira: *Data segmentation based on the local intrinsic dimension*, Sci. Rep., **10** (2020), Article ID: 116449.

[5] E. Altan, S. A. Solla, L. E. Miller and E. J. Perreault: *Estimating the dimensionality of the manifold underlying multi-electrode neural recordings*, PLoS Comput. Biol., **17** (11) (2021), Article ID: e1008591.

[6] A. Ansuini, A. Laio, J. H. Macke and D. Zoccolan: *Intrinsic dimension of data representations in deep neural networks*, Adv. Neural Inf. Process. Syst., **32** (2019).

[7] A. Asaad, D. Ali, T. Majeed and R. Rashid: *Persistent homology for breast tumor classification using mammogram scans*, Mathematics, **10** (21) (2022), Article ID: 4039.

[8] J. A. D. Binnie, P. Dłotko, J. Harvey, J. Malinowski and K. M. Yim: *A survey of dimension estimation methods*, (2025), arXiv preprint arXiv:2507.13887.

[9] M. Biondo, N. Cirone, F. Valle, S. Lazzardi, M. Caselle and M. Osella: *The intrinsic dimension of gene expression during cell differentiation*, Nucleic Acids Res., **53** (16) (2025), 1–11.

[10] T. Birdal, L. Guibas, A. Lou and U. Simsekli: *Intrinsic dimension, persistent homology and generalization in neural networks*, Adv. Neural Inf. Process. Syst. (NeurIPS 2021), **34** (2021), 6776–6789.

[11] R. Brüel-Gabrielsson, V. Ganapathi-Subramanian, P. Skraba and L. J. Guibas: *Topology-aware surface reconstruction for point clouds*, Comput. Graph. Forum., **39** (2020), 197–207.

[12] A. Bukkuri, N. Andor and I. K Darcy: *Applications of topological data analysis in oncology*, Front. Artif. Intell., **4** (2021), Article ID: 659037.

[13] F. Camastra, A. Staiano: *Intrinsic dimension estimation: Advances and open problems*, Inf. Sci., **328** (2016), 26–41.

[14] P. Campadelli, E. Casiraghi, C. Ceruti and A. Rozza: *Intrinsic dimension estimation: Relevant techniques and a benchmark framework*, Math. Probl. Eng., **2015** (2015), Article ID: 759567.

[15] G. Carlsson: *Topology and data*, Bull. Amer. Math. Soc., **46** (2009), 255–308.

[16] G. Carlsson, M. Vejdemo-Johansson: *Topological data analysis with applications*, Cambridge University Press, Cambridge (2022).

[17] F. Chazal, B. Michel: *An introduction to topological data analysis: Fundamental and practical aspects for data scientists*, Front. Artif. Intell., **4** (2021), Article ID: 667963.

[18] D. Cohen-Steiner, H. Edelsbrunner and J. Harer: *Stability of persistence diagrams*, Discrete Comput. Geom., **37** (2007), 103–120.

[19] H. Edelsbrunner, J. Harer: *Persistent homology-a survey*, Contemp. Math., **453** (2008), 257–282.

[20] H. Edelsbrunner, J. Harer: *Computational topology: An introduction*, American Mathematical Society, USA (2010).

[21] V. Erba, M. Gherardi and P. Rotondo: *Intrinsic dimension estimation for locally undersampled data*, Sci. Rep., **9** (1) (2019), Article ID: 17133.

[22] E. Facco, M. d'Errico, A. Rodriguez and A. Laio: *Estimating the intrinsic dimension of datasets by a minimal neighborhood information*, Sci. Rep., **7** (1) (2017), Article ID: 12140.

[23] E. Facco, A. Pagnani, E. T. Russo and A. Laio: *The intrinsic dimension of protein sequence evolution*, PLoS Comput. Biol., **15** (4) (2019), Article ID: e1006767.

[24] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera: *Learning from imbalanced data sets*, Springer Switzerland (2018).

[25] M. Flammer: *Persistent homology-based classification of chaotic multi-variate time series: Application to electroencephalograms*, SN Comput. Sci., **5** (2024), Article ID: 107.

[26] A. T. Fomenko: *Visual geometry and topology*, Springer Science and Business Media, Berlin (2012).

[27] Giotto-tda 0.5.1 Documentation. Available online: https://giotto-ai.github.io/gtda-docs/0.5.1/library.html (2021) (accessed on 25 January 2021).

[28] P. Grassberger, I. Procaccia: *Measuring the strangeness of strange attractors*, Physica D: Nonlinear Phenomena, **9** (1-2) (1983), 189–208.

[29] The GUDHI Project, GUDHI User and Reference Manual, 3.5.0 Edition, GUDHI Editorial Board. (2022). Available online: https://gudhi.inria.fr/doc/3.5.0/ (accessed on 13 January 2022).

[30] M. Guillemard A. Iske: *Interactions between kernels, frames and persistent homology*, Springer: Recent Applications of harmonic analysis to function spaces, differential equations, and data science, Switzerland (2017).

[31] J. Jaquette, B Schweinhart: *Fractal dimension estimation with persistent homology: A comparative study*, Commun. Nonlinear Sci. Numer. Simul., **84** (2020), Article ID: 105163.

[32] M. Jazayeri, S. Ostojic: *Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity*, Curr. Opin. Neurobiol., **70** (2021), 113–120.

[33] W. Jia, M. Sun, J. Lian and S. Hou: *Feature dimensionality reduction: a review*, Complex Intell. Syst., **8** (3) (2022), 2663–2693.

[34] I. T. Jolliffe, J. Cadima: *Principal component analysis: a review and recent developments*, Philos. Trans. A Math. Phys. Eng. Sci., **374** (2065) (2016), Article ID: 20150202.

[35] R. Kindelan, J. Frías, M. Cerda and N. Hitschfeld: *A topological data analysis based classifier*, Adv. Data Anal. Classif., **18** (2024), 493–538.

[36] A. Lawson, Y.-M. Chung and W. Cruse: *A hybrid metric based on persistent homology and its application to signal classification*, 2020 25th International Conference on Pattern Recognition (ICPR), Milan (Italy) (2020), 9944–9950.

[37] E. Levina, P. Bickel: *Maximum likelihood estimation of intrinsic dimension*, Adv. Neural Inf. Process. Syst., **17** (2004), 777–784.

[38] D. Leykam, D. G. Angelakis: *Topological data analysis and machine learning*, Adv. Phys., **8** (1) (2023), Article ID: 2202331.

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

[40] S. Majumdar, A. K. Laha: *Clustering and classification of time series using topological data analysis with applications to finance*, Expert Syst. Appl., **162** (2020), Article ID: 113868.

[41] C. Moon, Q. Li and G. Xiao: *Using persistent homology topological features to characterize medical images: Case studies on lung and brain cancers*, Ann. Appl. Stat., **17** (3) (2023), 2192–2211.

[42] D. Moroni, M. A. Pascali: *Learning topology: bridging computational topology and machine learning*, Pattern Recognit. Image Anal., **31** (2021), 443–453.

[43] H. S. Obaid, S. A. Dheyab and S. S. Sabry: *The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning*, 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON) (2019), Jaipur (India) 279–283.

[44] K. Özçoban, M. Manguoğlu and E. F. Yetkin: *A novel approach for intrinsic dimension estimation*, (2025), arXiv preprint arXiv:2503.09485.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay: *Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res., **12** (2011), 2825–2830.

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

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

[48] J. J. Rotman: *An introduction to algebraic topology*, Springer, (New York) (1988).

[49] N. Saul, C. Tralie: *Scikit-tda: Topological data analysis for Python*, (2019). Available online: https://doi.org/10.5281/zenodo.2533369 (accessed on 25 January 2019).

[50] B. Scholkopf, A. J. Smola: *Learning with Kernels: Support vector machines, regularization, optimization and beyond*, The MIT Press, Massachusetts (2002).

[51] J. Shawe-Taylor, N. Cristianini: *Kernel methods for pattern analysis*, Cambridge University Press, Cambridge (2009).

[52] Y. Skaf, R. Laubenbacher: *Topological data analysis in biomedicine: A review*, J. Biomed. Inform., **130** (2022), Article ID: 104082.

[53] A. Som, H. Choi, K. N. Ramamurthy, M. P. Buman and P. Turaga: *PI-Net: A deep learning approach to extract topological persistence images*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, (2020), 834–835.

[54] P. Sonego, M. Pacurar, S. Dhir, A. Kertész-Farkas, A. Kocsor, Z. Gáspári, J. A. M. Leunissen and S. Pongor: *A Protein classification benchmark collection for machine learning*, Nucleic Acids Res., **35** (2007), 232–236.

[55] B. Sorscher, S. Ganguli and H. Sompolinsky: *Neural representational geometry underlies few-shot concept learning*, Proc. Natl. Acad. Sci., **119** (43) (2022), Article ID: e2200800119.

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

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

[58] J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas and K. D. Vogiatzis: *Representation of molecular structures with persistent homology for machine learning applications in chemistry*, Nat. Commun., **11** (2020), Article ID: 3230.

[59] C. Tralie, N. Saul and R. Bar-On: *Ripser.py: A lean persistent homology library for Python*, J. Open Source Softw., **3** (29) (2018), 1–4.

[60] L. Van Der Maaten, E. O. Postma and H. J. Van Den Herik: *Dimensionality reduction: A comparative review*, J. Mach. Learn. Res., **10** (2009), 1–41.

[61] L. Van der Maaten, G. Hinton: *Visualizing data using t-SNE*, J. Mach. Learn. Res., **9** (11) (2008), 2579–2605.

[62] P. J. Verveer, R. P. W. Duin: *An evaluation of intrinsic dimensionality estimators*, IEEE Trans. Pattern Anal. Mach. Intell., **17** (1) (1995), 81–86.

[63] J. Von Rohrscheidt, B. Rieck: *Topological singularity detection at multiple scales*, Proceedings of the 40th International Conference on Machine Learning, Hawaii (USA) (2023), 35175–35197.

[64] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome and X. J. Wang: *Task representations in neural networks trained to perform many cognitive tasks*, Nat. Neurosci., **22** (2) (2019), 297–306.

[65] A. Zomorodian, G. Carlsson: *Computing persistent homology*, Discrete Comput. Geom., **33** (2005), 249–274.

CINZIA BANDIZIOL
UNIVERSITY OF PADOVA
DEPARTMENT OF MATHEMATICS "TULLIO LEVI-CIVITA"
35122, PADUA, ITALY
Email address: cinzia.bandiziol@phd.unipd.it

STEFANO DE MARCHI
UNIVERSITY OF PADOVA
DEPARTMENT OF MEDICINE
35128, PADUA, ITALY
PADOVA NEUROSCIENCE CENTER
35122, PADUA, ITALY
SELÇUK UNIVERSITY
DEPARTMENT OF MATHEMATICS
SELÇUKLU, KONYA, TÜRKIYE
Email address: stefano.demarchi@unipd.it

MICHELE ALLEGRA
UNIVERSITY OF PADOVA
DEPARTMENT OF PHYSICS AND ASTRONOMY
PADOVA NEUROSCIENCE CENTER
35122, PADUA, ITALY
Email address: michele.allegra@unipd.it