Estimating dimensionality by means of topological data analysis

Carles Casacuberta

Universidad de Barcelona

Universidad Autónoma de Madrid

8 de noviembre de 2023

Summary

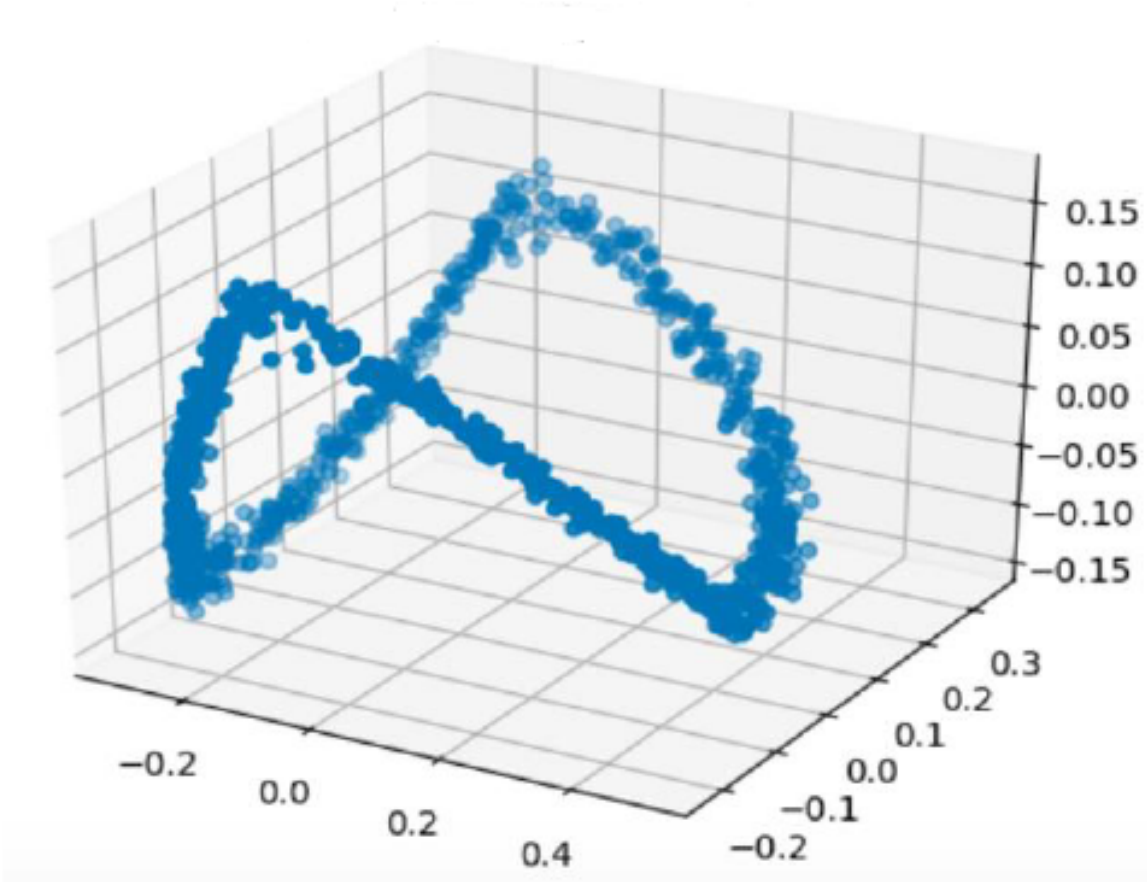
We describe two studies in which topological data analysis was used for dimensionality estimates.

- **Estimating the dimensionality of complex networks using network geometry and persistent homology,** with Meritxell Vila, Aina Ferrà, and María Ángeles Serrano, in preparation

- **A topological classifier to characterize brain states: When shape matters more than variance,** with Aina Ferrà, Gloria Cecchini, Fritz-Pere Nobbe Fisas, and Ignasi Cos, **PLoS ONE (2023) 18(10):e0292049**

Dimensionality of Data

Intrinsic dimension is the minimum number of variables needed to accurately describe the main features of a system.



Difficulties

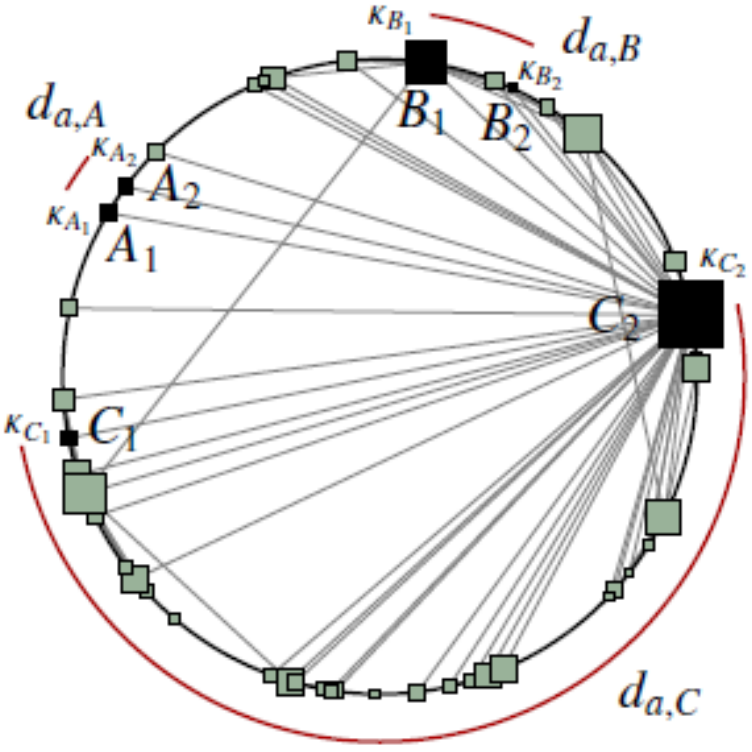
- **Failure of the manifold hypothesis:** Although data often distribute across a subset of smaller dimension within an ambient space, it need not adjust well to a manifold.
- **Curse of dimensionality:** When the ambient dimension increases, the volume of the space increases so fast that the available data become sparse.
- **Peaking phenomenon:** The average predictive power of a classifier first increases as the number of features used is increased, but beyond a certain dimension it starts deteriorating instead of improving steadily.
- **Existence of noise:** While the features of a dataset change remarkably in certain directions, small variations along other directions may be irrelevant for analysis.

Complex Networks

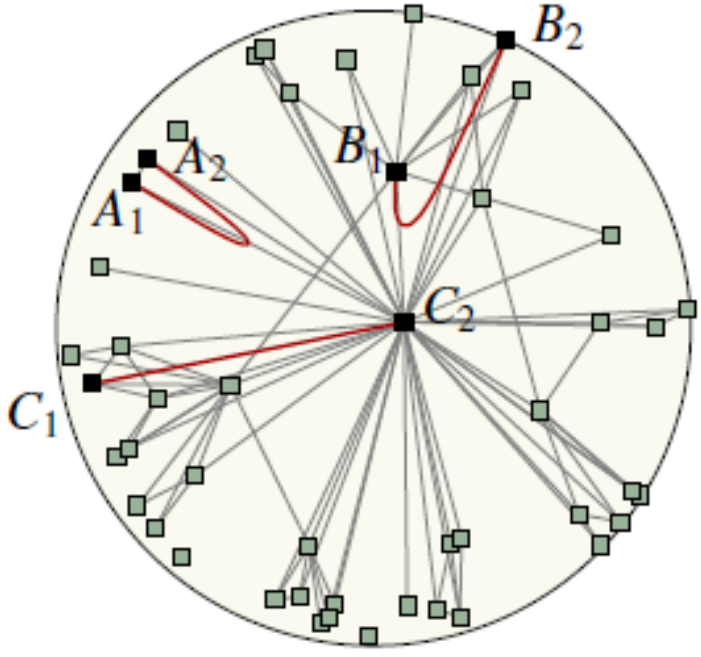
Main characteristics of **complex networks:**

- Small world phenomenon: Graph distance is useless.
- Scale-free (power-law) distribution of nodes: $P(\kappa) \sim \kappa^{-\gamma}$ with $\gamma > 2$, where κ denotes hidden degree (*popularity*).
- Degree heterogeneity: Average degree is denoted by μ.
- Clustering coefficient β (*inverse temperature*).

Hyperbolic Embeddings

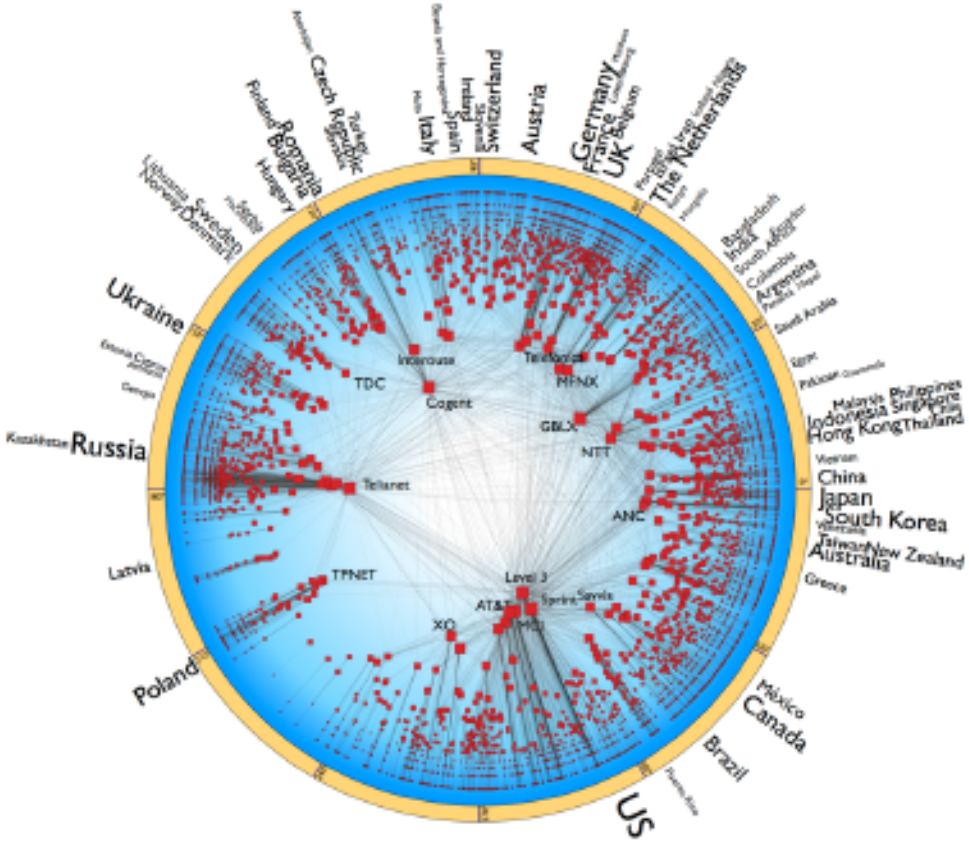


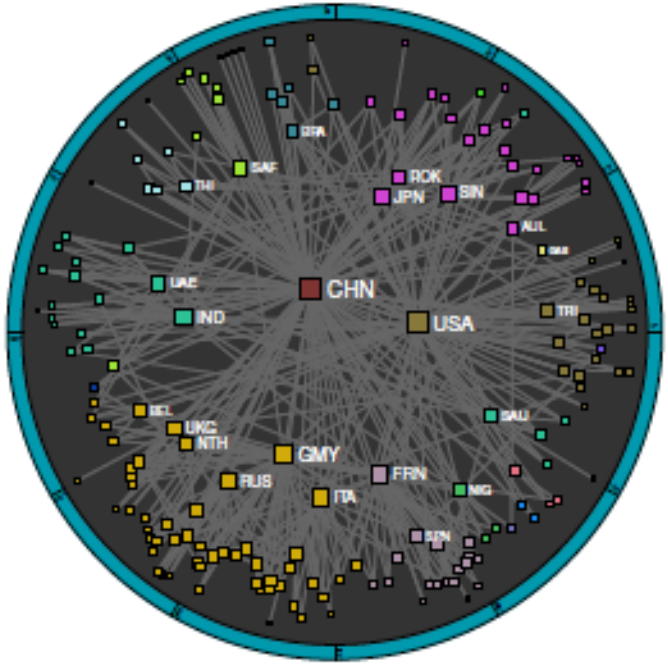
$\mathbb{S}^1$ model



$\mathbb{H}^2$ model

Hyperbolic Embeddings





Internet

World Trade

Surrogates

Synthetic graphs are generated through hyperbolic embeddings in each dimension D using an S^D model with the following probability of connection between nodes i and j:

$$p_{ij} = \frac{1}{1 + \chi_{ij}^{\beta}}, \qquad \chi_{ij} = \frac{R\Delta\theta_{ij}}{(\mu\,\kappa_i\,\kappa_j)^{1/D}}.$$

where $R = \left[\Gamma\left(\frac{D+1}{2}\right) N/(2\pi)^{\frac{D+1}{2}}\right]^{1/D}$ is the sphere radius.

N: number of nodes of the original network;

κ_i: hidden degree of node i in the original network;

$\Delta\theta_{ij}$: angular distance between nodes i and j in the S^D model.

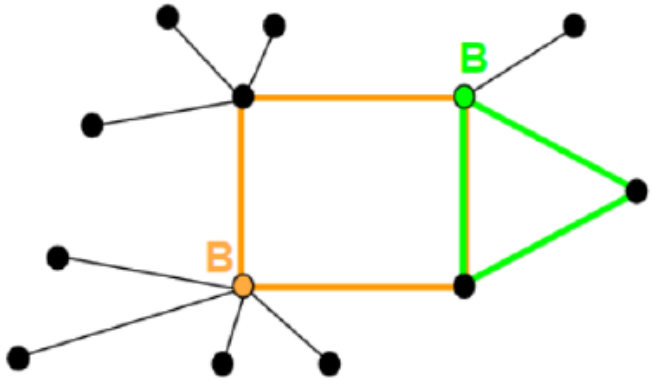
Extended Persistence of Graphs

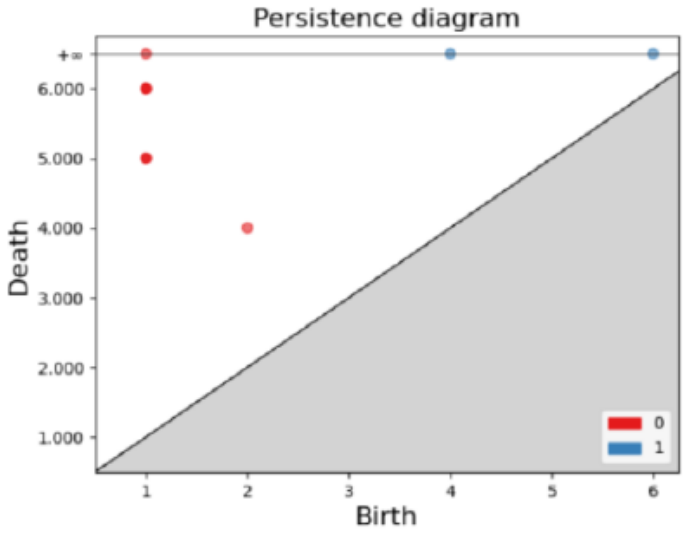
Persistence is calculated by means of a degree-based filtration:

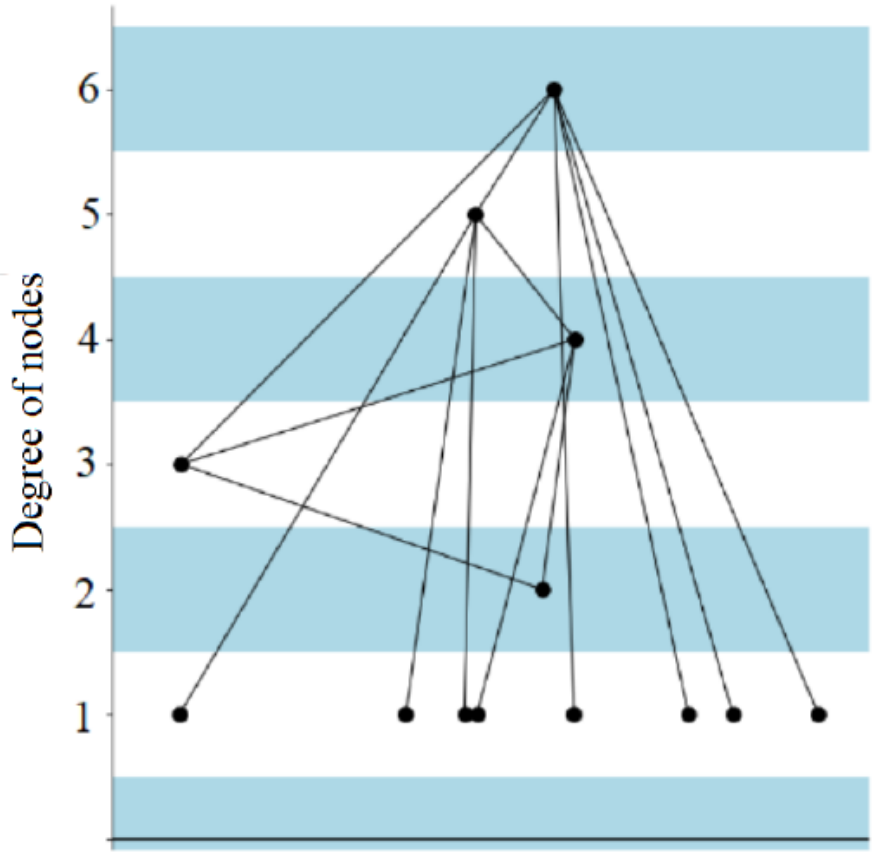
(a) Example graph (b) $f^{-1}((-\infty, 0])$ (c) $f^{-1}((-\infty, 1])$ (d) $f^{-1}((-\infty, 2])$ (e) $f^{-1}((-\infty, 3])$

Sublevel graphs

Persistence Diagrams

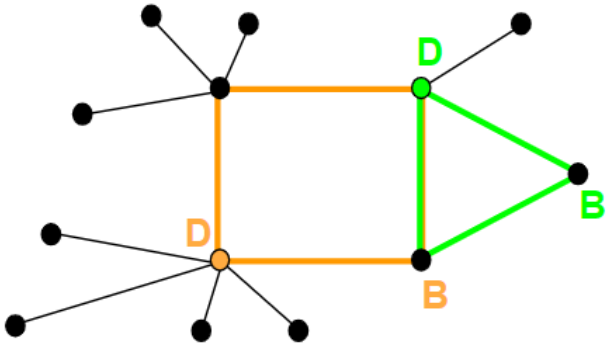


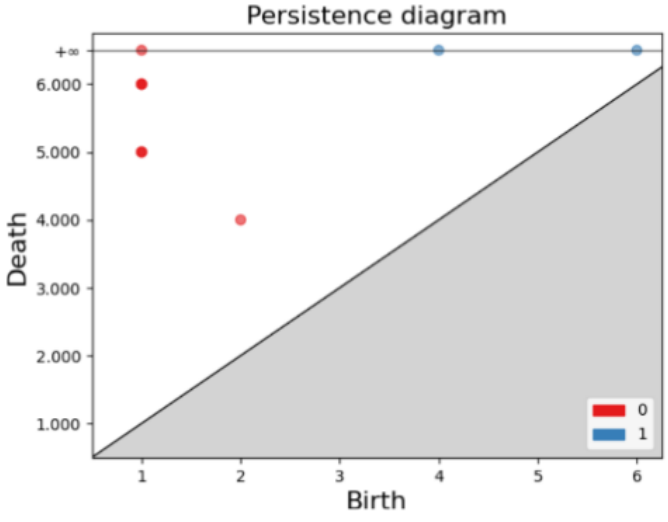


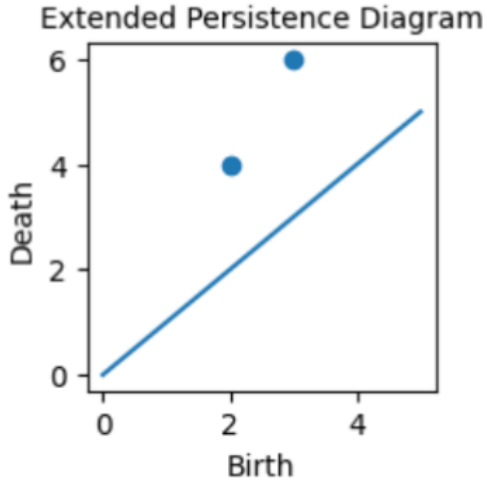


Merging of connected components

Extended Persistence







Extended persistence in homological dimension 1 combines sublevel graphs with superlevel graphs along the degree filtration.

Persistence descriptors

We use **total persistence** as numerical descriptor of a persistence diagram:

$$\mathrm{TP} = \sum_{i=1}^{n} (d_i - b_i),$$

where $\{(b_1, d_1), \ldots, (b_n, d_n)\}$ are points in the persistence diagram of a graph in a given homological dimension.

In this study we focus on homological dimension 1. Thus b_i is the birth degree of the ith cycle and d_i is the death degree, so TP is **cumulative lifetime of cycles.**

Dissimilarity Metrics

Dissimilarity between persistence diagrams can be measured with the **bottleneck distance** or using **kernels.**

The **Reininghaus kernel** or **scale-space kernel** is defined as

$$K(D, D') = \frac{1}{8\pi\sigma} \sum_{p \in D,\, q \in D'} e^{-|p-q|^2/8\sigma} - e^{-|p-\overline{q}|^2/8\sigma},$$

where $\overline{q} = (d, b)$ if $q = (b, d)$, and σ is a scale parameter.

Then a distance between D and D' is computed as

$$d(D, D') = \sqrt{K(D, D) - 2K(D, D') + K(D', D')}.$$

Results

We studied two real-world networks:

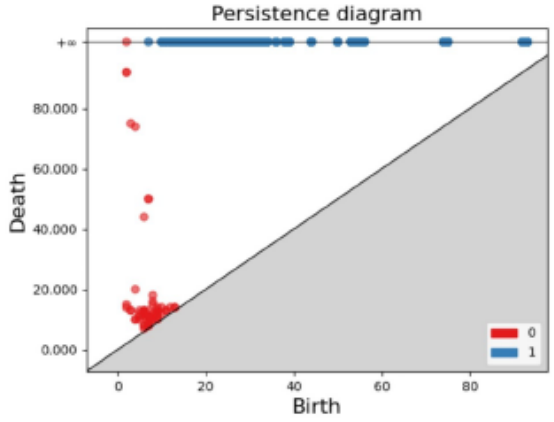
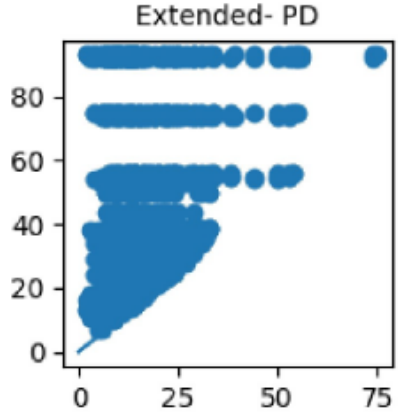
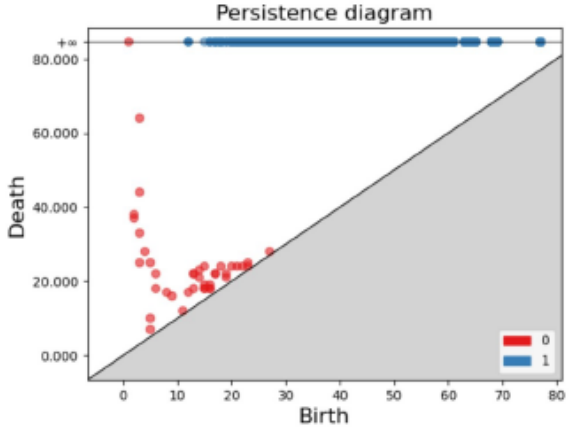
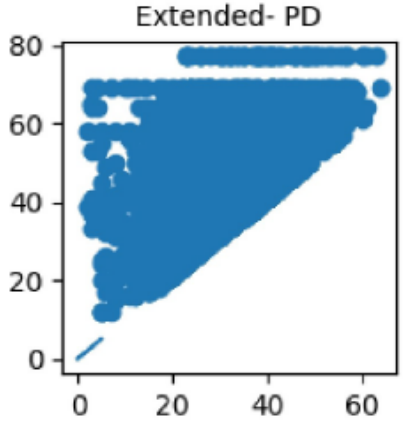
Network	Type	$\|V\|$	$\|E\|$	av. deg.	C	D
CElegans-C	Biological - Brain	279	2287	16.39	0.34	1
Human1	Biological - Brain	493	7773	31.53	0.49	3

CElegans-C: Nervous system of *Caenorhabditis elegans*.

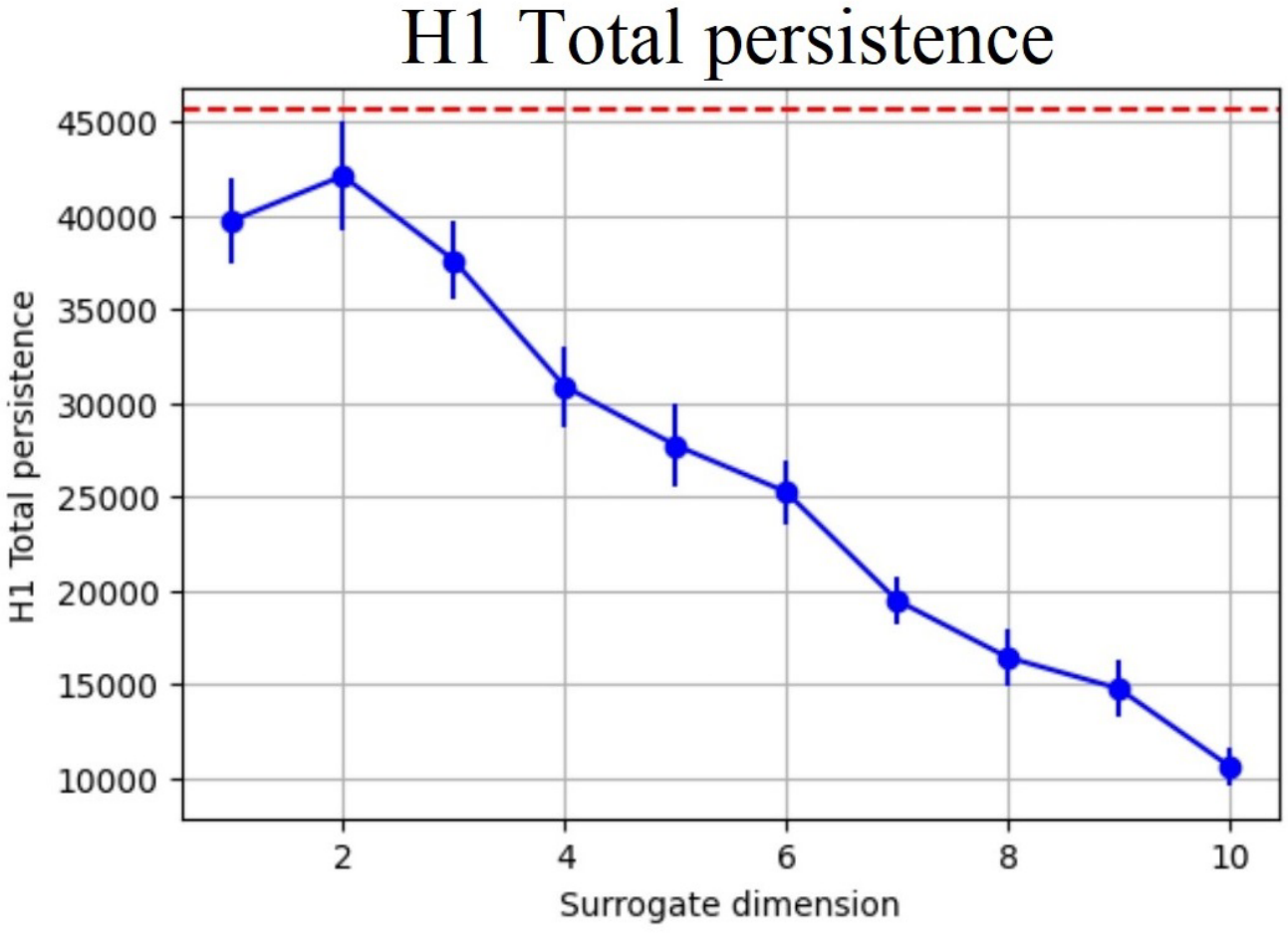
Human1: A connectome of the human brain at one hemisphere.

P. Almagro, M. Boguñá, M. A. Serrano, Detecting the ultra low dimensionality of real networks, *Nature Commun.* 13, 6096 (2022)

Results

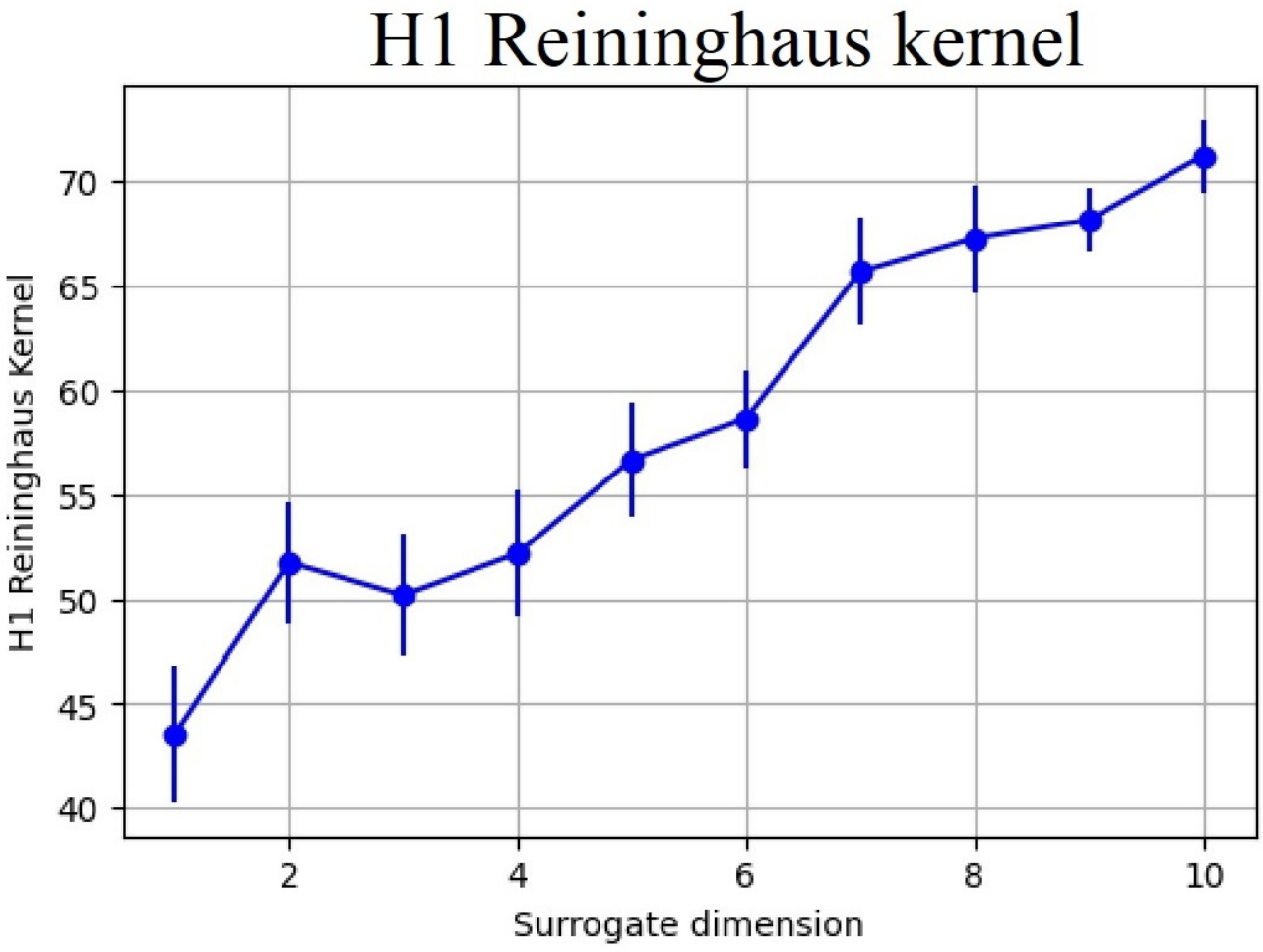
CElegans-C	Human1
 	 

Results



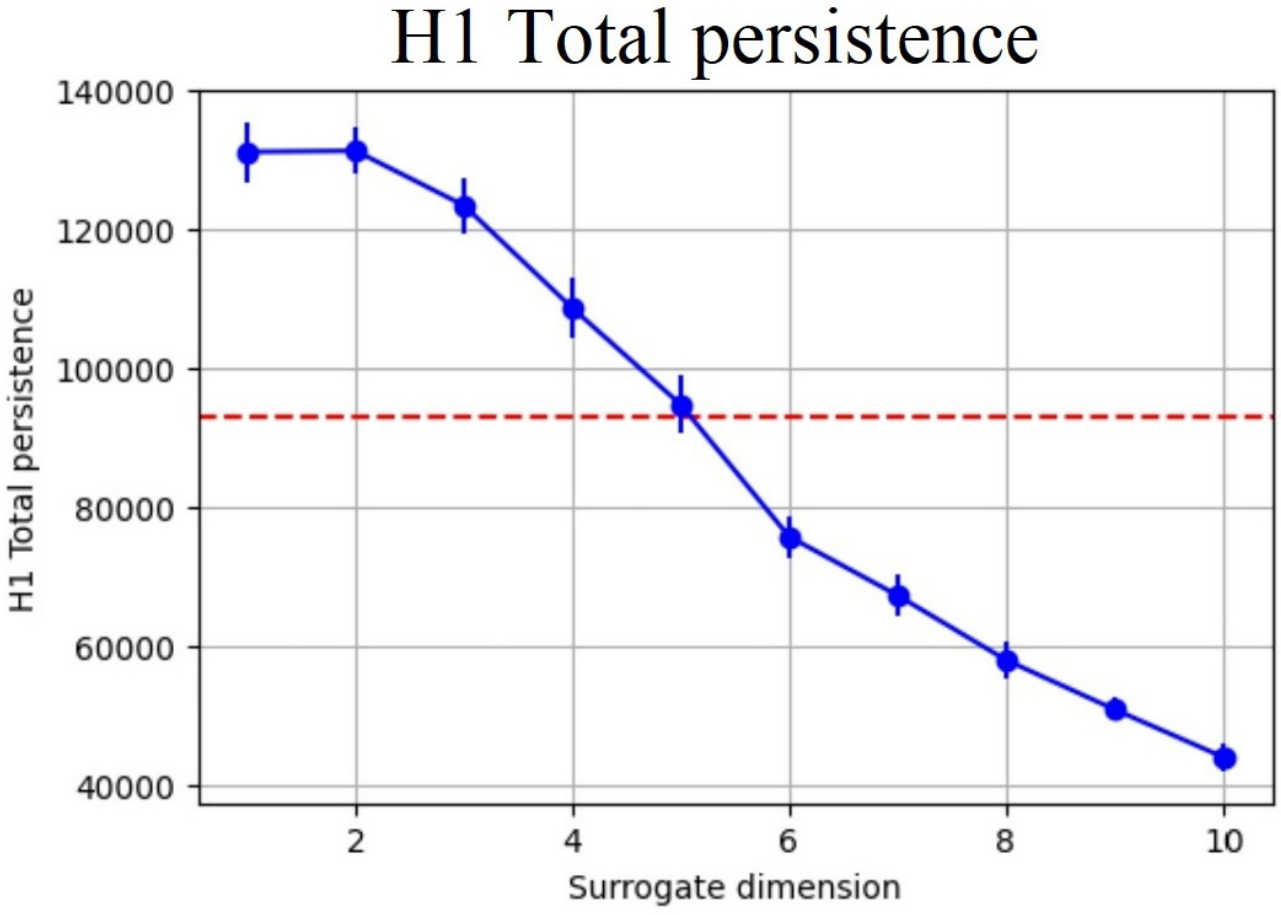
CElegans-C

Results



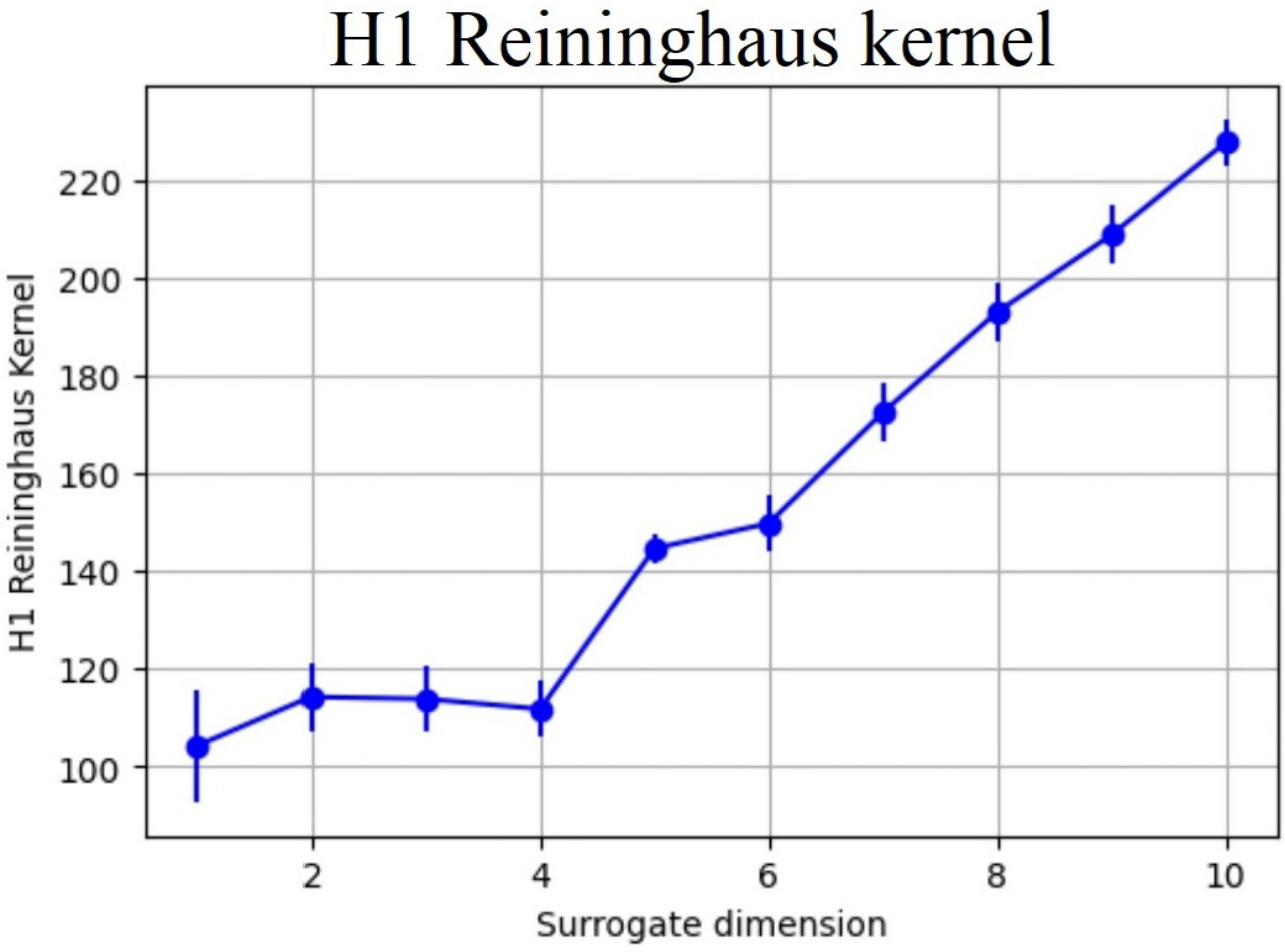
CElegans-C

Results



Human1

Results



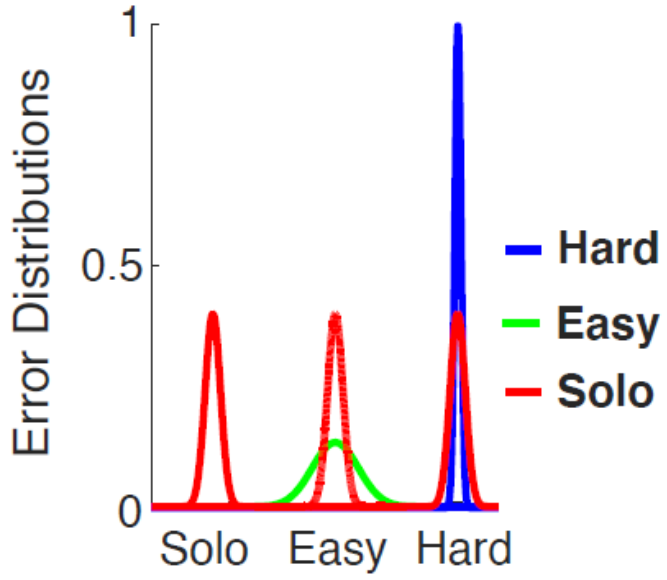
Human1

Conclusions

- Our results are compatible with a low-middle dimension (4-5) for the human brain connectome and a very low dimension (1-2) for the CElegans-C nervous system.
- Extended persistence in homological dimension 1 of surrogate graphs in hyperbolic models can estimate dimensionality of real-world complex networks.
- A degree-based filtration of graphs can be useful for topological data analysis.

Behavioral Neuroscience

In a **behavioral neuroscience study** carried out in Barcelona in 2015, each of 11 participants was offered a game with two sessions of 6 blocks of 108 trials, in which they were playing alone (4 blocks), with a virtual weak partner (4 blocks), and with a virtual strong partner (4 blocks).

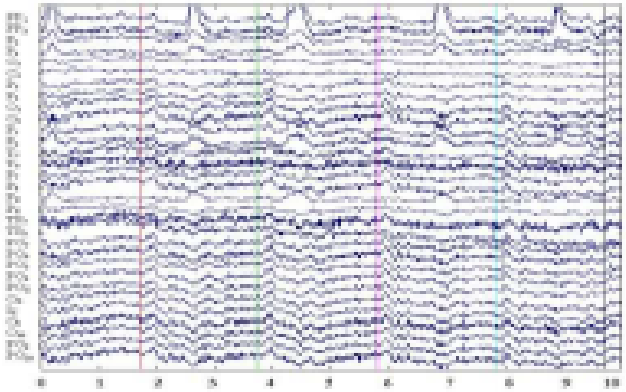


Behavioral Neuroscience

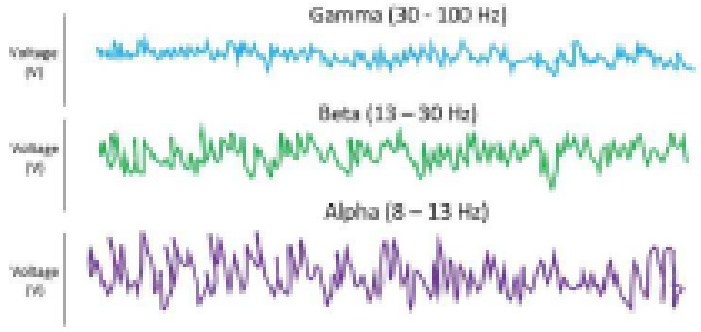
The dataset consisted of electroencephalogram fragments of 1200 ms recorded through 60 brain electrodes. The amplitude of each time series was averaged, separately into 8-15 Hz (alpha), 15-32 Hz (beta) and 32-80 Hz (gamma) frequency bands.

Thus the dataset consisted of $12 \times 108 = 1296$ points in a 60-dimensional space for each participant.

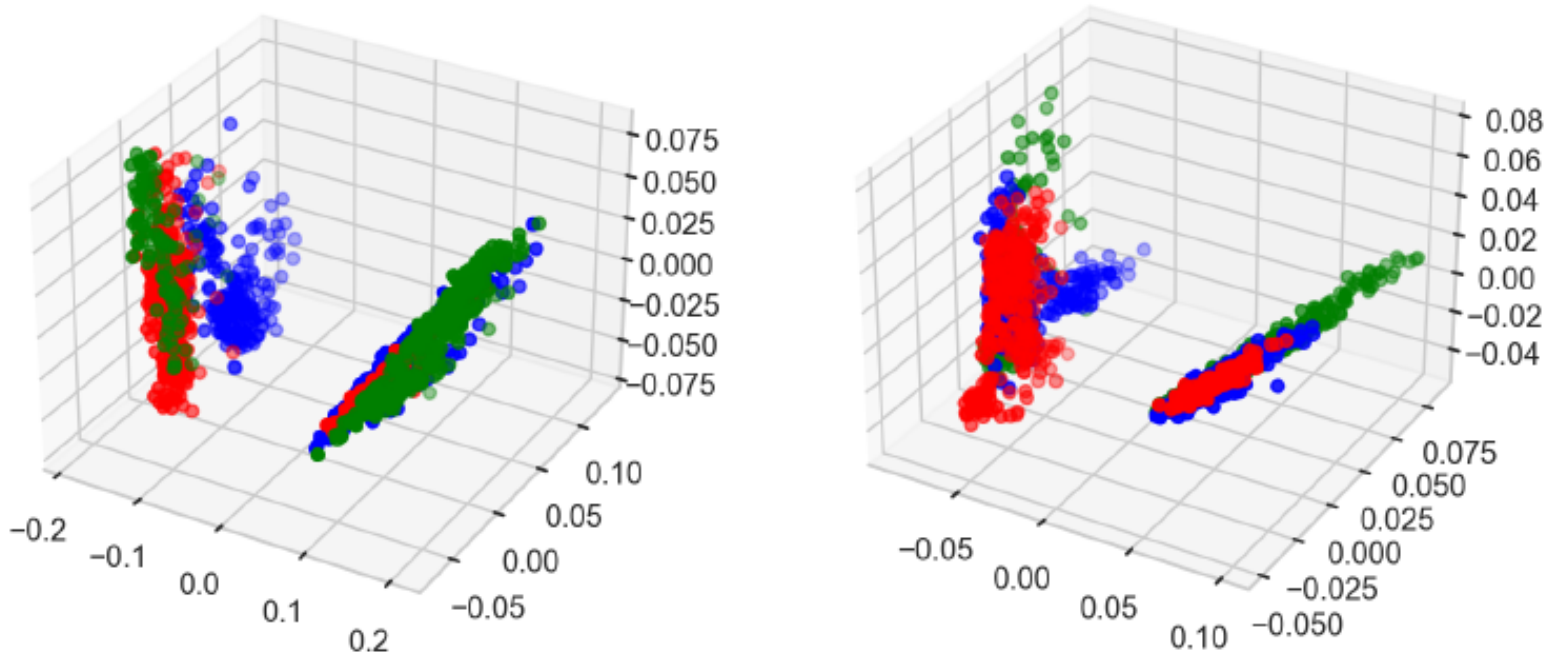
60 Channel Temporal Series



Frequency Components

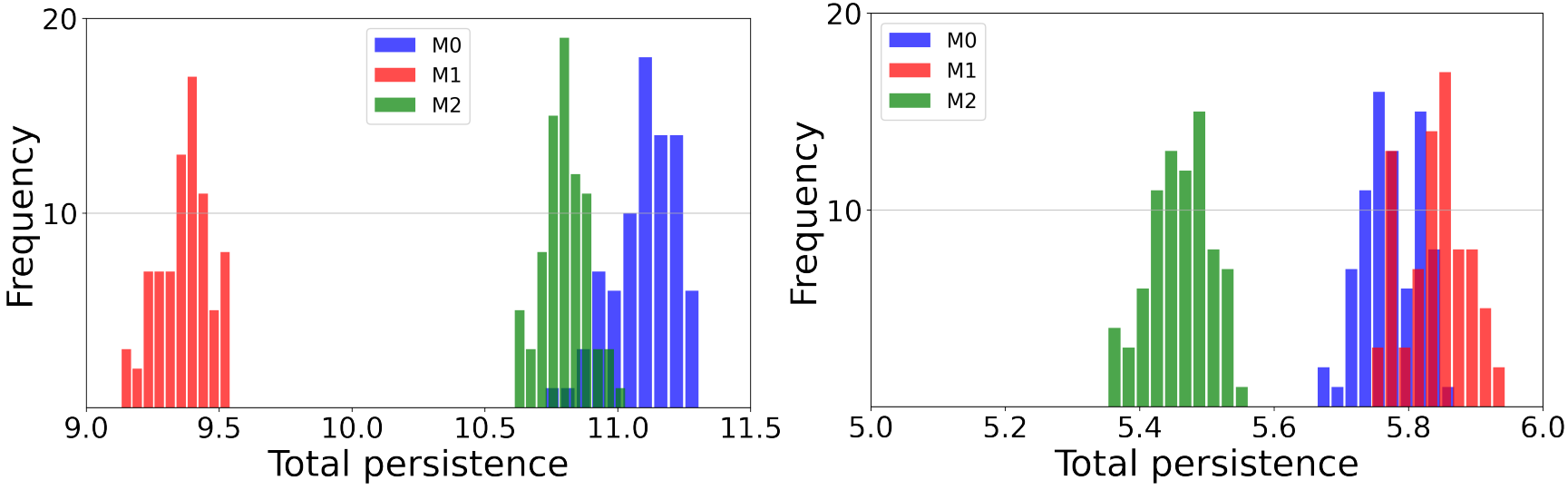


Motivational States



Point clouds corresponding to participants 1 and 8 in the γ band after applying principal component analysis (PCA) for reduction to dimension 3. The two clusters correspond to two sessions performed in each block. Blue: Solo; Red: Easy; Green: Hard.

Total Persistence



Distribution of total persistence $\Sigma_i(d_i - b_i)$ in dimension 0 of participant 1 (left) and participant 8 (right) of the point clouds corresponding to three different motivational states: M_0 blue (playing solo); M_1 red (playing against a weak opponent); M_2 green (playing against a strong opponent).

Comparing Motivational States

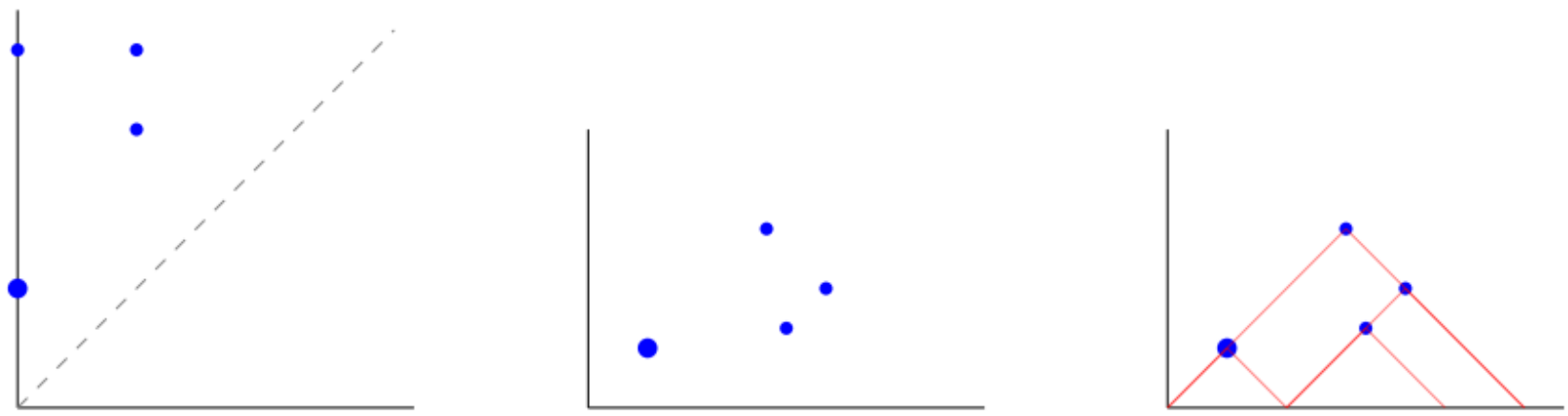
The underlying assumption is that **point clouds sampled from different classes exhibit recognizably different shapes.**

The plausibility of this claim in our study was tested by means of bootstrapping on each motivational state by repeatedly sampling 75% of each data cloud randomly with replacement 80 times.

The statistical null hypothesis that the distributions were pairwise equal was rejected for all participants by means of a Kolmogorov–Smirnov test with *p*-values below 0.0001.

Persistence Landscapes

Landscapes yield a convenient vectorization of a persistence diagram as a sequence of piecewise linear functions with compact support:



Persistence Silhouettes

A **silhouette** of a persistence diagram with m points (b_i, d_i) is a weighted average of landscape tent functions

$$\phi(t) = \frac{\sum_{i=1}^{m} w_i \, \Lambda_{(b_i,d_i)}(t)}{\sum_{i=1}^{m} w_i}$$

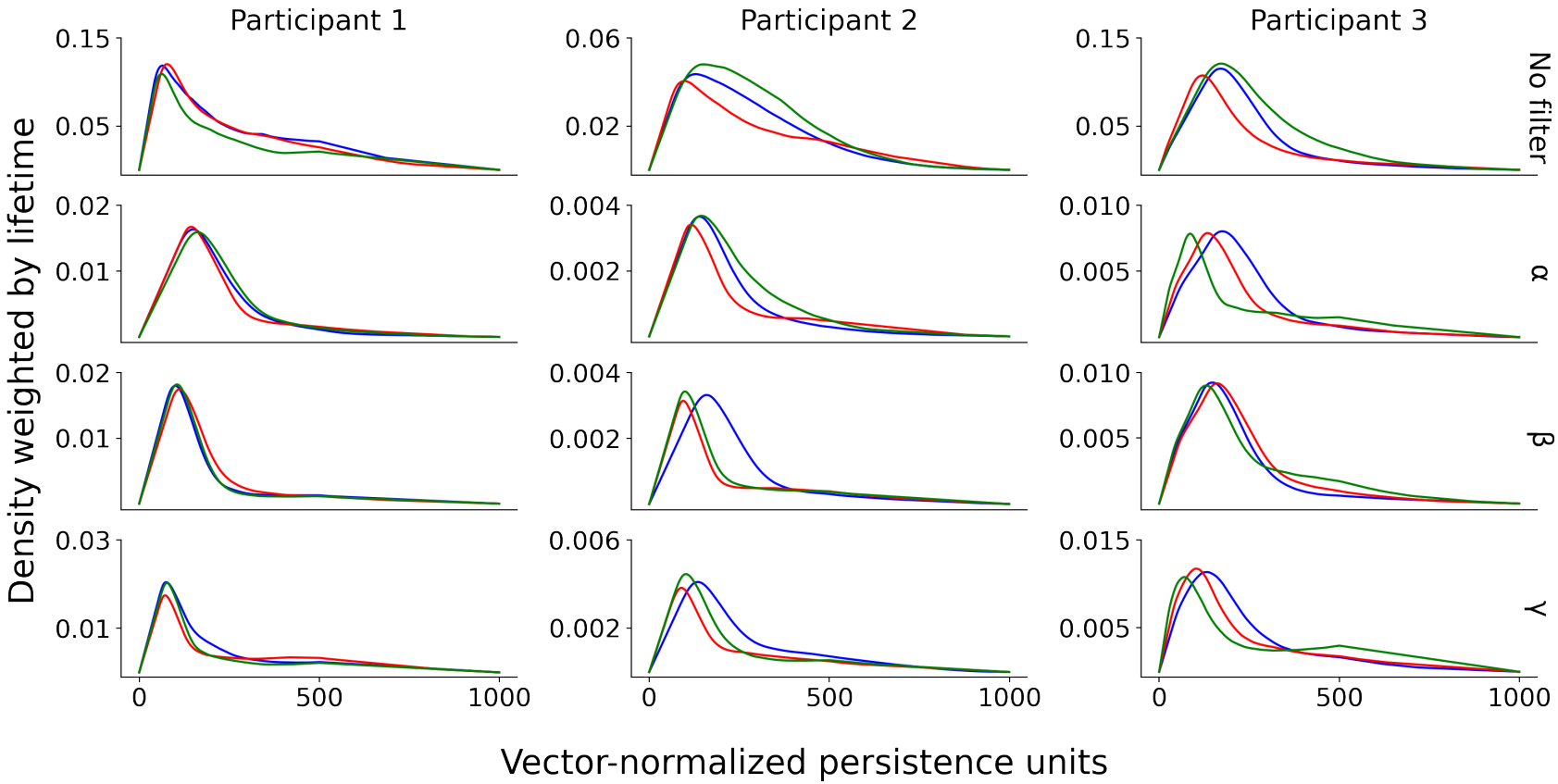
where $\{w_i\}$ are weights to be chosen, and

$$\Lambda_{(b,d)}(t) = \max\{0, \min\{t - b, d - t\}\}.$$

A frequent choice is $\boldsymbol{w_i = (d_i - b_i)^p}$ where p is optional:

- Choosing p small enhances low-persistence features.
- Choosing p large enhances highly persistent features.

Persistence Silhouettes

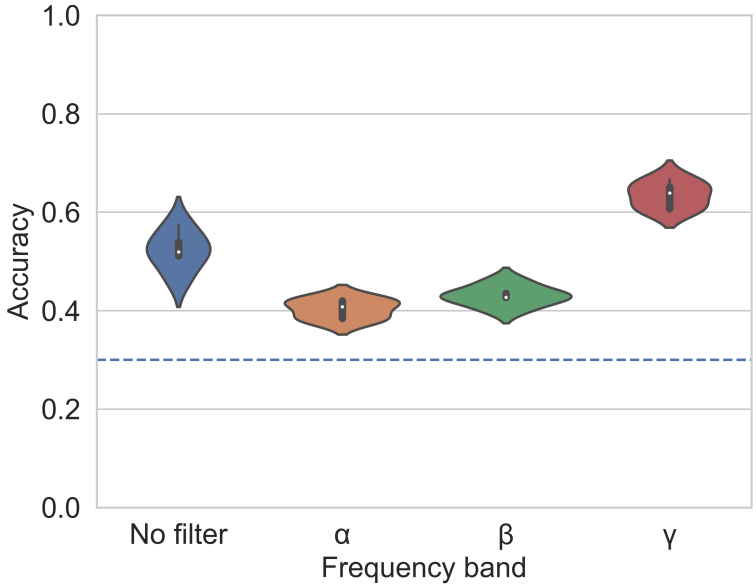


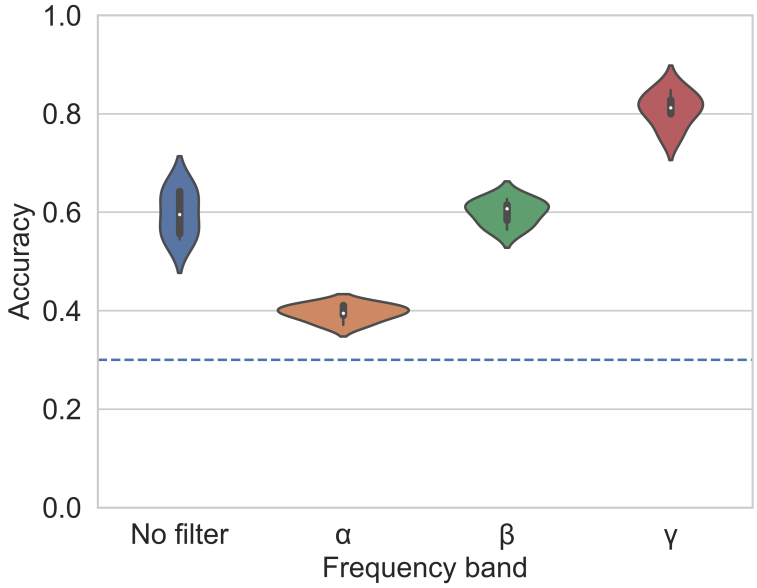
Silhouettes from persistence diagrams in dimension zero for each motivational state (M_0: blue, M_1: red, M_2: green) for each frequency band (α, β, γ) plus the unfiltered dataset.

TDA Classifier

1. The training set was split into classes according to labels.
2. For each class label c in the training set, we calculated a persistence silhouette S_c in homological dimension zero with lifetimes as weights.
3. To classify an input x from the testing set, we added x to the cloud of training datapoints X_c of each class label c. Then, we recomputed the persistence silhouettes $S_{c,x}$ for the datasets $X_c \cup \{x\}$, and finally calculated the Euclidean distance between the new silhouettes and the former ones.
4. We assigned the class label $c(x) = c_*$ where c_* attained the smallest distance between silhouettes.

Results

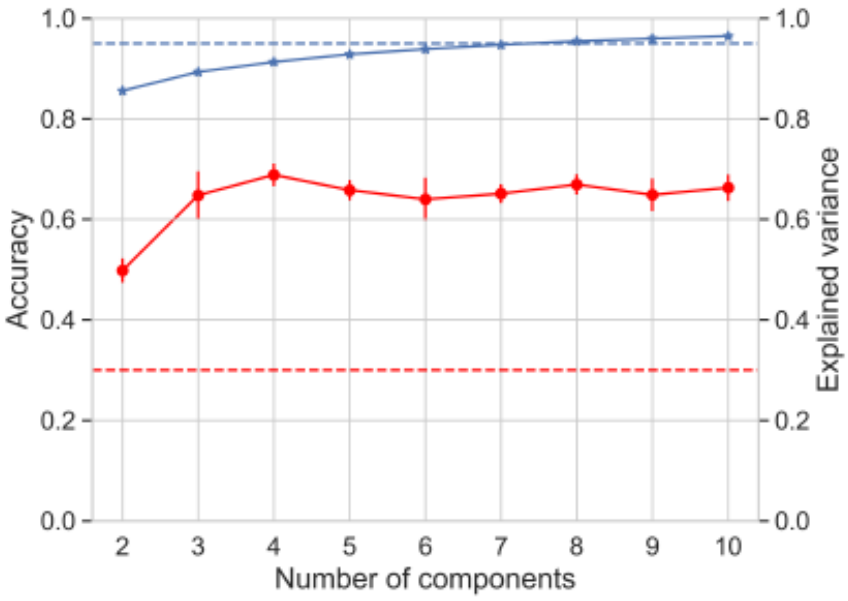


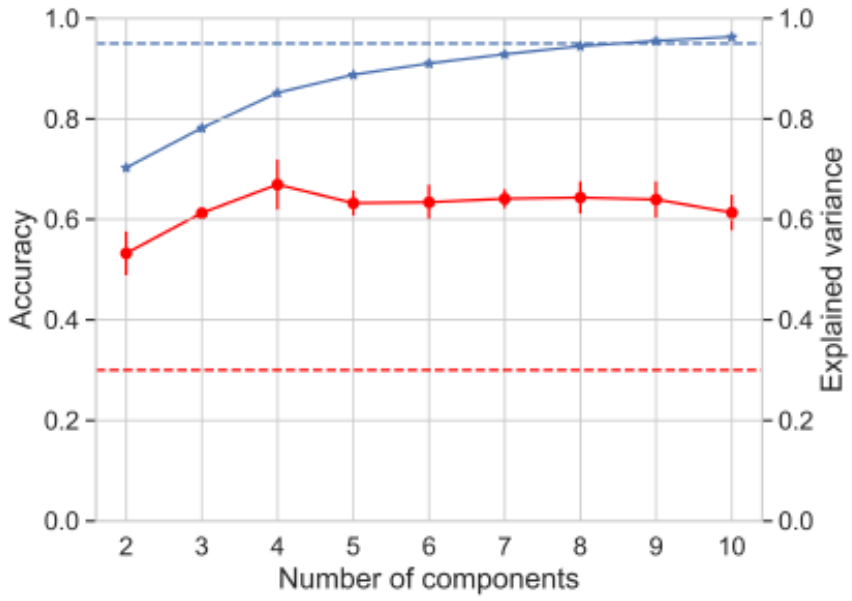


Accuracies of the topological classifier by frequency band without dimensionality reduction for participants 1 and 3.

Violin plots encode median, interquartile range, and a kernel-smoothed probability density.

Accuracy Variation





Comparison of variation of accuracy (red) with explained variance (blue) for participants 1 (left) and 8 (right) in a range of PCA dimensions. The blue dotted line indicates 95% of explained variance and the red dotted line is chance level.

Persistence Entropy

The above results indicate that our TDA classifier attains a maximum accuracy when the original dataset is reduced into a 4-dimensional ambient space by means of PCA.

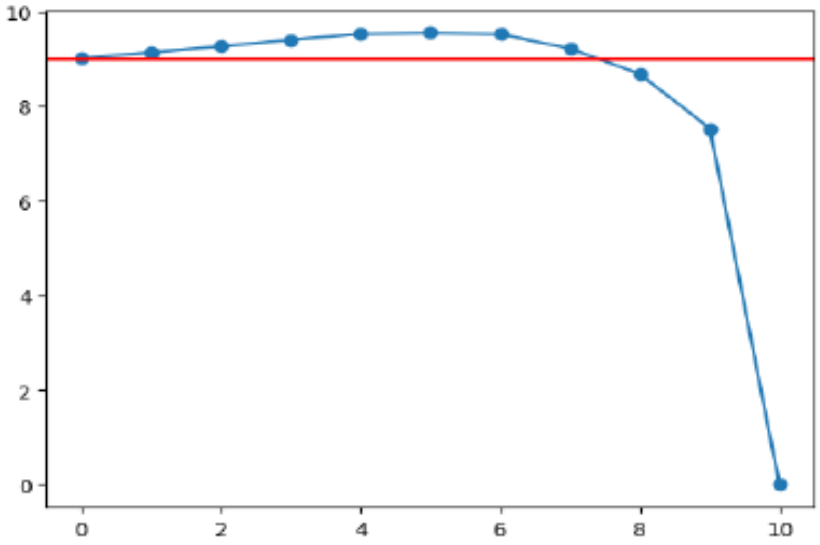
In order to test this observation further, we designed a second study in which persistence descriptors of the original point cloud were compared with the same descriptors of PCA projections in a range of dimensions from 12 to 2.

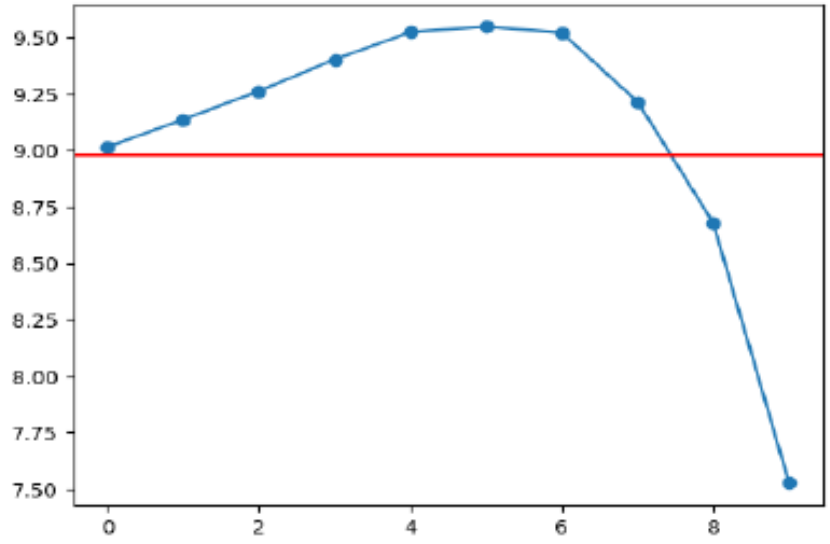
For this purpose, **persistence entropy** was used:

$$\mathrm{PE}(D) = -\sum_{(b_i,d_i)\in D} \left(\frac{d_i - b_i}{TP}\right) \log_2\left(\frac{d_i - b_i}{TP}\right),$$

where $\mathrm{TP} = \sum_i (d_i - b_i)$ denotes total persistence.

Dimensionality Estimates





Comparison of persistence entropy of the study dataset in ambient dimension 12 (red line) with persistence entropy of PCA-reduced point clouds in dimensions from 12 to 2 for two participants.

Conclusions

- Although the original data were recorded in dimension 60, our TDA classifier achieved maximum accuracy when PCA was applied to data onto dimension approximately 4.
- Persistence entropy of PCA-reduced point clouds was most similar to the original point cloud around dimension approximately 4.
- Experimental evidence from previous neuroscience studies suggests that the number of **brain sources** controling EEG signals for motivational states could be comprised between 4 and 6.