On the effectiveness of persistent homology

Renata Turkeš
Guido Montúfar
Nina Otter

Universiteit Antwerpen · imec · IDLab INTERNET & DATA LAB · UCLA · Queen Mary University of London

What is persistent homology?

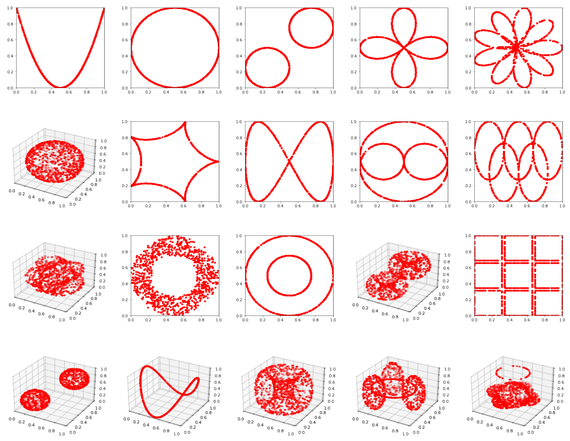
For a given data X (e.g., a point cloud or an image), persistent homology (PH) captures information about k-dimensional cycles (connected components, holes, voids, ...) in the so-called filtration, a nested family $\{K_r\}_{r\in\mathbb{R}}$ of topological spaces which approximate X at different scales $r \in \mathbb{R}$. PH can be seen as a multi-set of persistence intervals $[b, d]$, where b is the scale r where a topological feature (a k-dimensional cycle) is born, and d is the scale r when the feature dies in the filtration, and it can be represented by a variety of signatures.

What can you use persistent homology for?

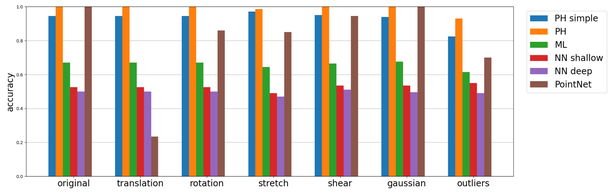
There have been numerous successful applications of PH in the last decade, but the reasons behind these successes are not yet well understood. The data used in real-world applications is complex, so that there are numerous effects at play and one is often left unsure why PH worked, i.e., what type of topological or geometric information it captured that facilitated the good performance. To initiate an investigation into the effectiveness of PH, or in other words, to investigate what is seen by PH, we set out to identify some fundamental data-analysis tasks that can be solved with PH.

Number of holes (Betti number β_1)

We consider 20 different shapes in $\mathbb{R}^2$ and $\mathbb{R}^3$, with four different shapes having the same number of holes ($0, 1, 2, 4$ or 9). For each shape, we construct 50 point clouds each consisting of $1\,000$ points sampled from a uniform distribution over the shape, resulting in a balanced dataset of $1\,000 = 20 \times 50$ point clouds.

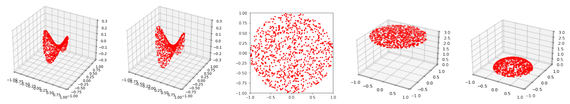


We train the classifier on 80% of the original point clouds, and test on the remaining 20% of original or noisy data. The results show that PH obtains very good test accuracy on this classification task, even in the presence of affine transformations or noise, outperforming baseline machine- and deep-learning techniques.

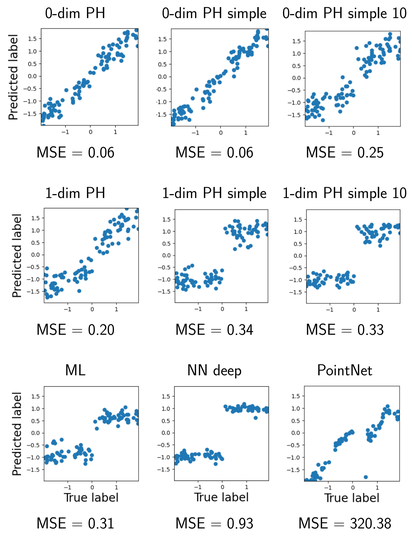


Curvature

A balanced dataset of point clouds with 500 points sampled from unit disks D_K on manifolds of constant negative, zero or positive curvature K is constructed as in [1]. The label of a point cloud is the curvature K of the underlying disk D_K.

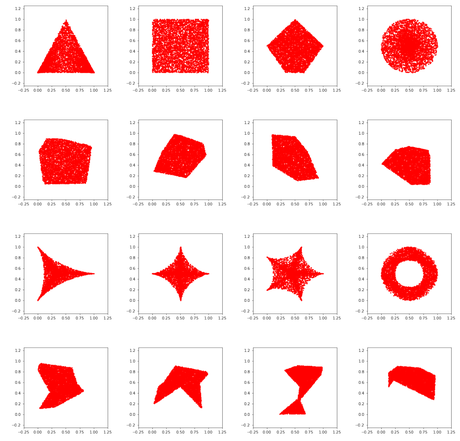


The mean squared errors (MSE) and the regression lines show that PH indeed detects curvature, outperforming other methods. We also see that the performance drops if we only focus on the longest 10 intervals (PH simple 10), so that the many short intervals together capture the geometry of interest for this problem.

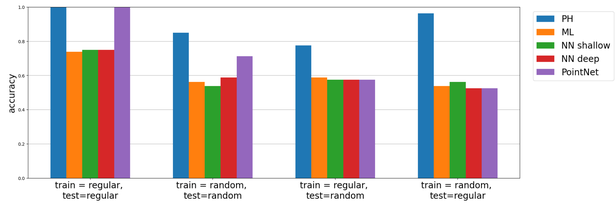


Convexity

We construct a balanced dataset by sampling $5\,000$ points from convex and concave shapes in $\mathbb{R}^2$. A point cloud has label 1 if it is sampled from a convex shape, and 0 otherwise.



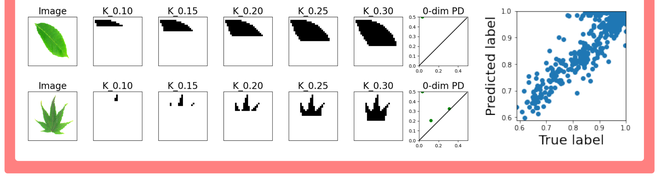
The classification accuracies under different conditions (with different train and test data) show that PH is able to detect convexity, outperforming other baseline methods.



Theorem

Let $X \subset \mathbb{R}^d$ be triangulizable. We have that X is convex if and only if for every line α in $\mathbb{R}^d$ the 0-dimensional PH with respect to the tubular filtration function $\tau_\alpha \colon \mathbb{R}^d \to \mathbb{R}$, $\tau_\alpha(x) = \mathrm{dist}(x, \alpha)$, contains exactly one persistence interval.

We also demonstrate that PH can detect a measure of convexity in the FLAVIA [2] image dataset of plant leaves.



Conclusions

A number of longest persistence intervals reflects the topology of data (Betti numbers β_k, the numbers of k-dimensional cycles), and the additional PH information on the birth and death values of the topological features can reveal different geometric properties such as curvature or convexity, depending on the choice of filtration and signature.

Summary of guidelines

SIGNAL	(Betti β_k) number of k-dimensional cycles	curvature	convexity
FILTRATION	Rips simplicial complex, filtered by euclidean distance	Rips simplicial complex, filtered by geodesic distance	cubical complex, filtered by tubular function
SIGNATURE	longest k-dimensional cycles	many short 0- and 1-dimensional cycles	second longest 0-dimensional cycle

References

[1] Bubenik, Peter, et al., *Persistent homology detects curvature*, Inverse Problems 36.2 (2020): 025008.

[2] Wu, Stephen Gang, et al., *A leaf recognition algorithm for plant classification using probabilistic neural network*, 2007 IEEE International Symposium on Signal Processing and Information Technology.