A new entropy based summary function for topological data analysis

N. Atienza, R. Gonzalez-Diaz, M. Soriano-Trigueros [1]

Departamento de Matemticas Aplicadas I
Universidad de Sevilla
Seville, Spain

Abstract

Topological data analysis (TDA) aims to obtain useful information from data sets using topological concepts. In particular, it may help to infer from finite sample when a configuration space is a manifold. So far, there is no automatic process to decide the main topological features of a given sampled manifold. In this article, we present an entropy-based summary function which may help to decide the most relevant Betti numbers from finite samples of a given manifold.

Keywords: persistent homology, entropy, topological data analysis.

1 Introduction

In order to obtain global features from a discrete data set, for example a point cloud in $\mathbb{R}^n$, we need to create a structure able to summarize the relation between them. A classical strategy is to create a proximity graph connecting points with edges and divide it in communities. In order to enrich the analysis

[1] Partially supported by MINECO, FEDER/UE under grant MTM2015-67072-P
Email: {natienza,rogodi,msoriano4}@us.es

of data sets and take advantage of the topological techniques we may use simplicial complexes. Note that a graph is formed by a set of vertices V and a set of edges. Each edge may be seen as a subset formed by two elements of V and each vertex as a subset formed by one element of V. Then, the graph may be seen as a family of subsets K of V each of them with one or two elements and satisfying that: If $a \in K$, then $\{v\} \subset a$ implies $\{v\} \in K$. A natural generalization of graphs is [2, p. 53]:

Definition 1.1 [Abstract simplicial complex] Let V be a finite set. A family K of subsets of V is an abstract $s$implicial complex if for every subsets $\sigma \in K$ and $\mu \subset V$, we have $\mu \subset \sigma$ implies $\mu \in K$.

When we want to visualize a graph, this can be represented in $\mathbb{R}^3$ drawing vertices (called 0-simplices) as points and edges (called 1-simplices) as lines in such a way that no edges cross each other. In the same way, simplicial complexes can be visualized in $\mathbb{R}^n$ using in addition triangles (called 2-simplices) for relating 3 points, tetrahedron (3-simplices) for 4 points and so on without self intersection of the simplices for a sufficiently large n. In order to increase even more the information carried by the simplicial complex, a nested sequence of subcomplexes can be defined.

Definition 1.2 [Filtration] Consider a simplicial complex K. A filtration on K is a sequence of simplicial complexes such that

$$K_0 \subset \ldots \subset K_{m-1} \subset K_m = K. \tag{1}$$

For example, consider a point cloud in a Euclidean space and consider its proximity graph G_d where the edges connect vertices whose distance is smaller than d. Therefore, if the number of points is finite, we have a set of distances $d_1 \leq \ldots \leq d_m$ where the proximity graphs change. We are interested in the *Vietoris-Rips complexes*, R_d, which can be constructed assigning to each $(n+1)$-vertex clique in G_d the corresponding n-simplex. Note that $d_1 \leq d_2$ implies $R_{d_1} \subset R_{d_2}$, then $R_{d_1} \subset \ldots \subset R_{d_{m-1}} \subset R_{d_m}$ is a filtration called the Vietoris-Rips filtration. See figure 1.

A topological invariant is an attribute assigned to an object that keeps unchanged under continuous deformation. We can use topological invariants to compare objects and establish differences between them. The aim of TDA is to use these notions to deduce properties from finite samples of the object. The main tool used is homology, which can be seen as a way of describing the "holes" of the given object, and is computed using algebraic techniques.

Consider the following simplicial complex which consists in a hollow triangle sharing an edge with a filled triangle, see figure 2. How can we detect

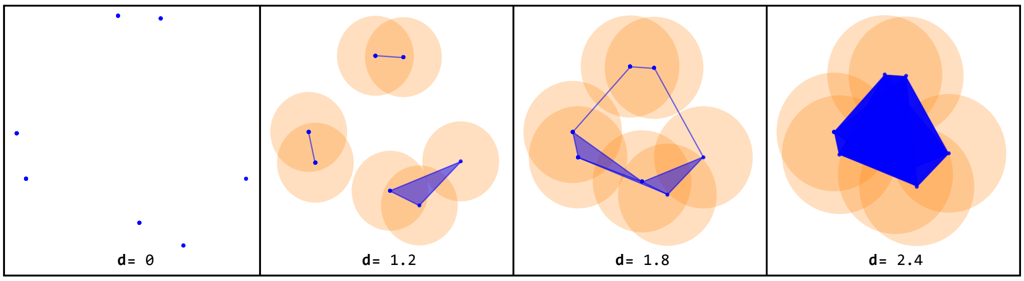

Fig. 1. Simplicial complexes K_d of a Vietoris-Rips filtration for $d = 0, 1, 2, 1, 8, 2, 4$.

the hole of the first triangle? Consider the vector space generated by vertexes (C_0) and edges (C_1) with coefficients in $\mathbb{Z}/\mathbb{Z}2$ (i.e. $1+1=0$). Then, the cycle surrounding the hole can be express as $ab + bd + da$. Define the boundary linear operator ∂_1 sending each edge to its vertexes , the path $ab+bd+da$ will be a *cycle* if every vertex in the path appears an even number of times when the boundary operator is applied: $\partial_1(ab+bd+da) = \partial_1(ab)+\partial_1(bd)+\partial_1(da) = a+b+b+d+d+a = 0$. In general, if a path is a cycle, its boundary is zero. Nevertheless, not all cycles are holes, for example, $bc + cd + db$. This is due to the boundary of the 2-simplex bcd being $bc+cd+db$, i.e. $\partial_2(bcd) = bc+cd+db$.

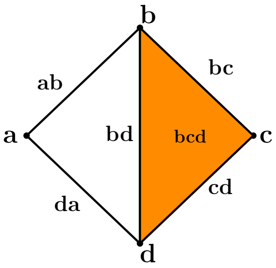

Fig. 2. Simplicial complex with one hole.

In general, the boundary linear operator of a simplicial complex K is defined as:

$$\partial_i[v_0, \dots, v_i] = \sum_{s=0}^{i} [v_0, \dots, \hat{v}_s, \dots, v_i] \tag{2}$$

Where $[v_0, \dots, v_i]$ is an i-simplex of K with vertexes $v_0, \dots, v_i$ and $\hat{v}_s$ means v_s has been removed. It is verified that $\partial_{i+1} \circ \partial_i = 0$ so the i-th homology can be defined as

$$H_i(K) = \frac{\ker \partial_i}{\operatorname{img} \partial_{i+1}} \tag{3}$$

which, as we have mentioned before, is a topological invariant. The i-th homology represents the i-th dimensional holes of the simplicial complex. For example, the 0-th homology represents connected components, the 1-th homology represents cycles and the 2-th homology represents voids of K.

Observe in our example, $ab + bd + da$ and $ab + bc + cd + da$ are considered the same hole due to the equivalence relation. The number of independent holes at each dimension is called the i-th Betti number.

This concept can be extended to filtrations. The inclusion $K_j \hookrightarrow K_{j+1}$ induce a linear map between vector spaces $H(K_j) \to H(K_{j+1})$. Intuitively when a hole disappears (i.e., it is in K_j but not in K_{j+1} for some j), this map sends it to zero. When this happens we say it dies at time j. When a hole appears by the first time (i.e., it is in K_j but not in K_{j-1} for some j) we say it has born at time j. We represent the moment of birth and death time of the generators of homology (the independent holes) using barcodes. The bottleneck distance, d, makes barcodes a metric space, [2, p. 180].

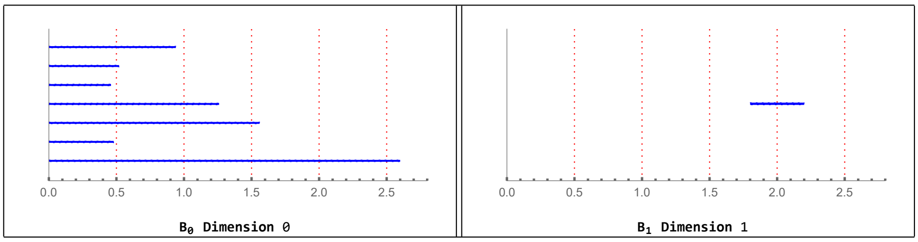

Fig. 3. Barcodes representing birth and death times of the homology of dimension 0 (connected components) and 1 (cycles) of the filtration in Figure 1.

2 Entropy-based functional summary

Let B be a barcode. Enumerate the length of its bars as $\{\ell_i\}_{i=1}^n$ and let $L = \sum_{i=1}^n \ell_i$. Then, the real number:

$$E(B) = \sum_{i=1}^{n} -\frac{\ell_i}{L} \log \frac{\ell_i}{L} \tag{4}$$

is called its persistent entropy. The greater the number of bars is and the more homogeneous they are, the greater entropy is. Some applications may be found in [5,6]. The maximum value of $E(B)$ is $\log(n)$ and is reached when $\ell_1 = \ldots = \ell_n$. In [1] we have proved this stability result for pesistent entropy:

Theorem 2.1 *For any two finite metric spaces (X, d_X) and (Y, d_Y), let A, B be the persistence barcodes coming from $Rips(X,t)|_{t\in\mathbb{R}}$ and $Rips(Y,t)|_{t\in\mathbb{R}}$ and*

d_{GH} the Gromov-Hausdorff distance. Consider n_{max} the maximum number of bars of A and B and L_a, L_b the total sum of their respective bars. If $\ell_a = L_a/n_{\max}$ and $\ell_b = L_b/n_{\max}$, $\ell_{\max} = \max\{\ell_a, \ell_b\}$. If $d_\infty(A,B) \leq \frac{1}{8}\ell_{\max}$ then

$$d_{GH}(X,Y) \leq \delta \Rightarrow |E(A) - E(B)| \leq \frac{4\delta}{\ell_{\max}} \left[\log(n_{\max}) - \log\left(\frac{4\delta}{\ell_{\max}}\right)\right]. \tag{5}$$

We will use this function to automatically detect the underlying shape of the data. A manifold is a space which locally looks like a euclidean space (e.g. a curve, a torus or a hypersphere). Manifolds appears in nature as the possible configuration space of a physical system or experiment output. The work of Latschev [4] implies that for dense enough point clouds contained in a manifold the Vietoris-Rips filtration is homotopically equivalent to the manifold during a period of time and consequently in this period both of them have the same homology. Unfortunately, this period of time is difficult to compute and the homology of the manifold is inferred using subjective criteria. In particular, the set of intervals of the barcodes which are considered topological features are expected to satisfy the following properties in that period of time:

- The lengths of the alive bars during that period are big and similar between them. (This means the contribution of these bars to the persistent entropy is big in comparison with the others).
- Few bars are alive in that period.
- The period these bars are the only ones alive, is long.

Our aim is to associate to each barcode a function with higher values in the period which satisfies these properties. Considering partial sums of persistent entropy can help with the first requisite and must play a role in the function, which we define as follows.

$$F_B(t) = -\frac{T(t)}{W(t)} \sum_{i=1}^{n} w_i(t) \frac{\ell_i}{L} \log\left(\frac{\ell_i}{L}\right) \tag{6}$$

Where $w_i(t)$ is 1 if the i-th bar is alive at t and 0 otherwise, $W(t) = \sum_{i=1}^{n} w_i(t)$ is the number of bars which are alive in that moment and $T(t)$ is the length of time bars in t are the only ones alive.

Once we have the function associated to the barcode, we select its higher values and see which bars are alive at that time. See figure 4, where using the summary function, we recover the homology of a circle from a 10 points sample. This summary function could be used as a statistic and help to infer the underlying shape of data.

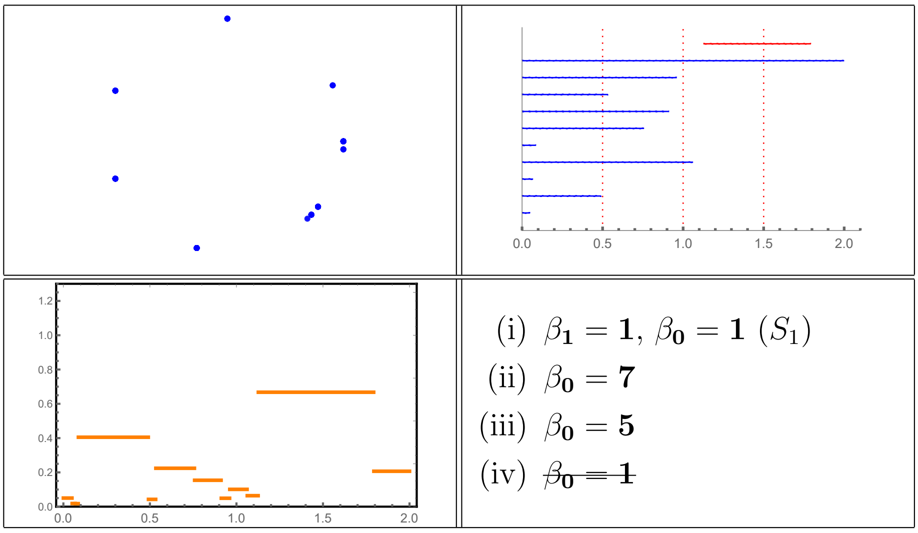

Fig. 4. Top-left: ten points contained in a circle. Top-right: barcode of its Vietoris-Rips filtration. Note that at time $t = 1.5$ the expected Betti numbers are reached: one cycle and one connected component. Bottom-left: the summary function associated to the barcode. Bottom-right: the associated betti numbers obtained from the summary functions.

References

[1] N. Atienza, R. Gonzalez-Diaz, M. Soriano-Trigueros. *On the stability of persistent entropy and new summary function for TDA.* http://arxiv/abs/1803.08304

[2] H. Edelsbrunner, J.L. Harer. *Computational Topology: An Introduction.* American Mathematical Society, 2010.

[3] J.C. Hausmann. *On the VietorisRips complexes and a cohomology theory for metric spaces.* Annals of Mathematics Studies 138, pages 175-188, 1995.

[4] J. Latschev. *Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold.* Archiv der Mathematik, Volume 77, Issue 6, pages 522-528, 2001.

[5] M. Rucco, F. Castiglione, E. Merelli, M. Pettini. *Characterisation of the Idiotypic Immune Network Through Persistent Entropy.* Proceedings of ECCS 2014, pages 117-128. Springer Proceedings in Complexity.

[6] M. Rucco, R. Gonzalez-Diaz, M. J. Jimenez, N. Atienza, C. Cristalli, E. Concettoni, A. Ferrante, E. Merelli. *A new topological entropy-based approach for measuring similarities among piecewise linear functions*, Signal Processing 134, 2017, pages 130-138.