On the stability of persistent entropy and new summary functions for Topological Data Analysis

N. Atienza, R. Gonzalez Diaz, M. Soriano Trigueros

June 7, 2018

Abstract

Persistent entropy of persistence barcodes, which is based on the Shannon entropy, has been recently defined and successfully applied to different scenarios: characterization of the idiotypic immune network, detection of the transition between the preictal and ictal states in EEG signals, or the classification problem of real long-length noisy signals of DC electrical motors, to name a few. In this paper, we study properties of persistent entropy and prove its stability under small perturbations in the given input data. From this concept, we define three summary functions and show how to use them to detect patterns and topological features.

1 Introduction

In the last decade, correct and efficient interpretation of data has become a key problem in science and industry. In this context, topological data analysis (TDA) attempts to create reliable methods based on topological features of spaces in order to obtain useful information from data sets. Intuitively, topological features can be seen as qualitative geometric properties relating the notions of proximity and continuity.

The standard workflow is the following: Start with a data set, for example a point cloud, endowed with some notion of proximity (usually a metric) depending on the kind of information we want to obtain. Then, create a simplicial complex and a filter function on it to encapsulate this information. A nested sequence of increasing subcomplexes is then computed using the filter function. Calculate the homology groups of each simplicial complex (intuitively, each element of the homology groups represents a "hole" in the simplicial complex). Finally, treat all these homology groups together as the subject of study, leading to the key concept of persistent homology.

Persistent homology summarizes hidden structural features of the given data set and can be compactly represented using persistence barcodes [5], diagrams [12] and, more recently, landscapes [3]. There exist stability results showing that these representations are robust under small perturbations of the given data (see, for example, [11]). During the last decade, this approach has been successfully applied in many areas (see, for example, [14]). Nowadays, there exist numerous softwares to compute persistent homology and its representations. A nice study of the performance of different available softwares is made in [21].

Persistence barcodes, diagrams or landscapes (endowed with a metric) are normally used to compare different given data sets. Nevertheless, using a scalar function instead of a metric space could be easier to interpret for people not being familiar with this topic. Actually, this function may be used not only for comparing different spaces but also for obtaining topological properties from them. Persistent entropy [8, 23] based on Shannon entropy [25] is a perfect candidate for this approach. Some applications of persistent entropy are given, for example, in [23], [20], [24] and [2]. Persistent entropy is used in [1] to distinguish topological features from noise. Nevertheless, it seems that there is a lack of stability results guaranteeing the reliability of persistent entropy (although first steps in this direction have already been done, for example, in [24, 1]). The main objective of this paper is proving such a stability result of persistent entropy.

The simplicity of persistent entropy is at the same time its main virtue and its main weakness. For this reason, in this paper, we define a new stable summary function which may be used to describe persistence barcodes. Finally, two modified versions of this summary function are created in order to detect topological features and patterns of point clouds embedded in a manifold. Examples illustrating the usefulness of these new functions are also given.

The paper is organized as follows: After reminding the existing theory regarding persistent homology and persistent entropy in Section 2, we provide stability results of persistent entropy in Section 3. In Section 4, we introduce three summary functions derived from the concept of persistent entropy and study their stability. Examples showing the applicability of these functions are also given. The paper ends with a section devoted to conclusions and future work.

2 Background

In this section, we give a quick overview about how algebraic topology is applied to data analysis and recall the definition of persistent entropy. An instructive book showing the main algebraic topology tools for data analysis is [11]. A general introduction to algebraic topology can be found in [15].

As explained in the introduction, to apply topological tools to data analysis, we first need to "encode" the information provided by the data into a simplicial complex.

Definition 2.1 (Abstract simplicial complex)**.** Let S be a finite set. A family of subsets K of S is an abstract $s$implicial complex if for every subsets $\sigma \in K$ and $\mu \subset S$, we have that $\mu \subset \sigma$ implies $\mu \in K$. A subset in K of $m+1$ elements of S is called an m-simplex.

In other words, if two subsets of a simplicial complex K have elements in common, then their intersection is a simplex in K formed by these common elements.

This combinatorial object have a geometrical interpretation. Consider S as a set of points of $\mathbb{R}^n$. Fix $m \leq n$. An m-simplex σ is a subset of $m+1$ affinely independent points of S, denoted by $\sigma = \langle x_0, \dots, x_m \rangle$. A 0-simplex is a point of S, a 1-simplex is a segment joining two points of S, a 2-simplex is a filled triangle, a 3-simplex is a filled tetrahedron and so on. When the finite set S represents some data, its nature may establish some relation depending on the context. We can use this to enrich the information carried by the complex, using the concept of filtration.

Definition 2.2 (Filtration)**.** A $f$ilter on a simplicial complex K is a monotonic function $f : K \to \mathbb{R}$ satisfying that $\mu \subset \sigma$ implies $f(\mu) \leq f(\sigma)$. A filtration on K, obtained from f, is the sequence of simplicial complexes $\left(K_t\right)_{t \in \mathbb{R}}$ where $K_t = f^{-1}(-\infty, t]$.

Note that, because of the monotonicity of f, the set K_t is a simplicial complex for all t, and $t_1 \geq t_2$ implies that $K_{t_1} \supseteq K_{t_2}$. The parameter t will be called time. The next definition is an example of filtration and require S to be a metric space.

Definition 2.3 (Vietoris-Rips filtration)**.** Let S be a finite set of points endowed with a distance d_s. The Vietoris-Rips filtration of S is the sequence $\left(Rips(S,t)\right)_{t \in \mathbb{R}}$ obtained from the filter function $f([x_0, \dots, x_m]) = \max_{i,j} d_S(x_i, x_j)$ where, for each $t \in \mathbb{R}$, the simplices of the Vietoris-Rips simplicial complex $Rips(S,t)$ are defined as:

$$\sigma = \langle x_0, \dots, x_m \rangle \in Rips(S,t) \Leftrightarrow d_S(x_i, x_j) \leq t \text{ for all } i \; j.$$

2.1 Persistent homology and persistence barcodes

Homology groups of a simplicial complex provides a formal interpretation of its geometric "holes". Persistent homology captures the variation of the homology groups of the simplicial complexes in a filtration. This information can be represented using persistence barcodes.

Given a simplicial complex K, an m-chain c is a formal sum of m-simplices in K. That is, $c = \sum_{i=1}^{k} a_i \sigma_i$ where, for $1 \leq i \leq k$, σ_i is an m-simplex in K and a_i is a coefficient in an unital ring R. Usually, $R = \mathbb{Z}/2\mathbb{Z}$ and then $a_i \in \{0,1\}$ satisfies that $a_i + a_j = 0$ iff $a_i = a_j = 0$ or $a_i = a_j = 1$, for $1 \leq i, j \leq k$. In order to relate the m-chains of a given simplicial complex K with its m-dimensional "holes" we need the boundary operator ∂_m. If $\langle x_0, \dots, x_m \rangle$ is an m-simplex in K then,

$$\partial_m \langle x_0, \dots, x_m \rangle = \sum_{i=0}^{m} \langle x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_m \rangle.$$

We can extent this definition to any m-chain by linearity. Note that $\partial_{m-1} \circ \partial_m = 0$ or, in other words, the boundary of a boundary is null. The "holes" of K are detected from chains whose

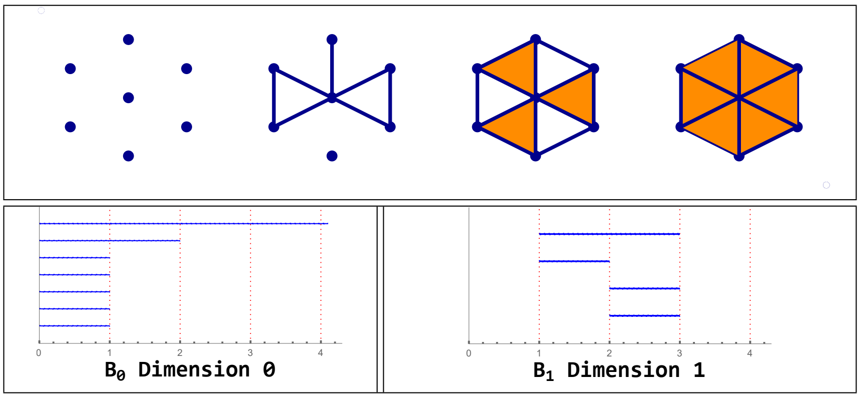


Figure 1: Top: example of a filtration $\mathcal{F}$. Bottom: 0-th and 1-st persistence barcodes of $\mathcal{F}$.

boundary is zero without being a boundary themselves. The m-dimensional homology group of K is then defined as the quotient group

$$H_m(K) = \frac{Ker\ \partial_m}{Img\ \partial_{m+1}},$$

and its m-dimensional Betti numbers as $\beta_m = rank\ H_m(K)$. Intuitively β_0 is the number of connected components of K, β_1 the number of 2-dimensional holes, β_2 the number of cavities and so on.

In order to study the variation of homology groups of the simplicial complexes in a given filtration, we need the concept of persistent homology.

Definition 2.4 (Persistent homology)**.** Let $\mathcal{F} = \big(K_t\big)_{t\in\mathbb{R}}$ be a filtration. For each $t \in \mathbb{R}$, let $H_m(K_t)$ be the m-dimensional homology group of K_t. For every $a \leq b$ and m, consider the function $v_m^{a,b} : H_m(K_a) \to H_m(K_b)$ induced by the inclusion $K_a \hookrightarrow K_b$. The family of homology groups $\big(H_m(K_t)\big)_{t\in\mathbb{R}}$ together with the functions $\big(v_m^{a,b}\big)_{t\in\mathbb{R}}$ is called the m-th *persistent homology* of the filtration $\mathcal{F}$.

Remark 2.5. Let $[\sigma]$ be a class of the quotient group $H_m(K_t)$. Let $t_1 = \sup\{\, a : (v_m^{a,t})^{-1}[\sigma] = \emptyset\}$ and $t_2 = \inf\{\, a : v_m^{t,a}[\sigma] \neq 0\}$. Then, $t_1 \leq t \leq t_2$. In other words, a generator σ of the class $[\sigma]$ appears in t_1 and keeps "alive" (as an image of the functions v) until the moment t_2 where its image becomes 0. Then, $[\sigma]$ is a *persistent homology class* and t_1 and t_2 are its *birth and death times.*

In this paper, we assume that $rank\ H_m(K_t) < \infty$ for all t, m and that the total number of persistent homology classes is finite. The information obtained by persistent homology can be represented, for example, via persistence barcodes or diagrams.

Definition 2.6 (Persistence barcodes)**.** Let $\mathcal{H}$ be the m-th persistent homology of a filtration $\mathcal{F}$. For each m-th persistent homology class α in $\mathcal{H}$, let $t_1(\alpha)$ and $t_2(\alpha)$ be its birth and death times. Then, $\mathcal{H}$ can be encoded as a multiset of intervals $\big\{[t_1(\alpha), t_2(\alpha)]\big\}_{\alpha\in\mathcal{H}}$. This multiset is the m-th persistence barcode of $\mathcal{F}$. Let $\mathcal{B}$ denote the space of all possible persistence barcodes.

An example of persistence barcode is showed in Figure 1.

If the birth and death times of persistent homology classes of $\mathcal{H}$ are encoded as points in $\mathbb{R}^2$ (i.e., $\big\{(t_1(\alpha), t_2(\alpha))\big\}_{\alpha\in\mathcal{H}} \subset \mathbb{R}^2$) then we obtain a multiset of points which is called the *persistence diagram of* $\mathcal{F}$.

Persistence barcodes (or diagrams) can be used to classify spaces that change along time (encoded as a filtration). In order to compare two different persistence barcodes for such classification task, we need to define a metric on $\mathcal{B}$.

Definition 2.7 (Wasserstein distance)**.** Consider $A, B \in \mathcal{B}$ and $1 \leq p \leq \infty$. Define the *p-th Wasserstein distance* as

$$d_p(A,B) = \left(\min_{\gamma} \sum_{i=1} \max\left\{|x_i^a - x_{\gamma(i)}^b|^p, |y_i^a - y_{\gamma(i)}^b|^p\right\}\right)^{\frac{1}{p}},$$

where γ is any bijection between the multisets $A = \{[x_i^a, y_i^a]\}_{i=1}^{n_a}$ and $B = \{[x_i^b, y_i^b]\}_{i=1}^{n_b}$. In case $n_a \neq n_b$, we can add intervals of zero length ($[t, t]$) until both multisets A and B have cardinal $n_{max} = \max\{n_a, n_b\}$. The limit case $p = \infty$ is called the bottleneck distance and is given by

$$d_\infty(A, B) = \min_\gamma \max_i \left(\max\left\{ |x_i^a - x_{\gamma(i)}^b|, |y_i^a - y_{\gamma(i)}^b| \right\} \right).$$

Observe that we have replaced the inf and sup terms in the original definition of Wassertein and bottleneck distance [11, p. 180-183] by min and max terms because, in this paper, persistence barcodes have always a finite number of intervals.

Remark 2.8. When $y_i^a = y_{\gamma(i)}^b = \infty$ we will assume $|y_i^a - y_{\gamma(i)}^b| = 0$ so the max function takes the maximum of the finite values only.

We finish this section with some well-known stability results, supporting the idea that similar inputs produce similar outputs. The second theorem is a consequence of the first one.

Theorem 2.9 ([9]). *Let $f, g : X \to \mathbb{R}$ be two tame Lipschitz functions on a metric space X whose triangulations grow polynomially with constant exponent $j \geq 1$. Then, there are constant $C \geq 1$ and $k \geq j$ such that the p-th Wasserstein distance between their corresponding m-th persistence barcodes, A and B, satisfies that for every $p \geq k$.*

$$d_p(A, B) \leq C||f - g||_\infty^{1-\frac{k}{p}}.$$

When $p = \infty$, the constant C is no longer necessary, obtaining the most commonly used simplified version which appears in [11, p. 183].

Proposition 2.10 ([11, p. 183]). *Let K be a simplicial complex and let $f, g : K \to \mathbb{R}$ be two monotonic functions. If $A, B \in \mathcal{B}$ are their corresponding m-th persistence barcodes, then*

$$d_\infty(A, B) \leq ||f - g||_\infty.$$

Theorem 2.11 ([6]). *Consider two finite metric spaces (X, d_X), (Y, d_Y). Let A, B be the two m-th persistence barcodes obtained, respectively, from $Rips(X, t)|_{t\in\mathbb{R}}$ and $Rips(Y, t)|_{t\in\mathbb{R}}$. Then $d_\infty(A, B) \leq d_{GH}(X, Y)$, where d_{GH} denotes the Gromov-Hausdorff (GH) distance.*

We could conclude that bottleneck distance gives simple expressions for the stability results and seems the best distance to work with.

2.2 Persistent Fntropy

Persistent entropy was first introduced in [8] and formally defined in [23]. The idea of persistent entropy is to somehow apply Shannon entropy to persistence barcodes.

Since classical Shannon entropy is defined for finite probability distribution, $\{(p_1, \ldots, p_n)$: $p_1 + \cdots + p_n = 1, 0 \leq p_i \leq 1\}$ then, we first need to normalize persistence barcodes.

Definition 2.12 (Persistent entropy). Consider a persistence barcode $A = \{[x_i^a, y_i^a]\}_{i=1}^{n_a}$ where $\max_i\{y_i\} < \infty$. Persistent entropy of A is:

$$E(A) = -\sum_i^{n_a} \frac{\ell_i^a}{L_a} \log\left(\frac{\ell_i^a}{L_a}\right),$$

where $\ell_i^a = y_i^a - x_i^a$ and $L_a = \ell_1^a + \cdots + \ell_{n_a}^a$.

The maximum possible value of persistent entropy $E(A)$ is $\log(n_a)$ and is reached when all intervals of A have the same length (the uniform distribution in probability terms). The minimum value is 0 and coincides with the case when there is only one interval (i.e. $n_a = 1$). In general, the greater the number of intervals is and the more homogeneous they are, the greater persistent entropy is.

Persistent entropy is only defined here for persistence barcodes with intervals of finite length (also called *finite intervals*). Later, we will discuss what can be done when intervals of infinite length (also called *infinite intervals*) are present in the barcode.

3 Stability of persistent entropy

In this section we provide some stability results regarding persistent entropy guaranteeing that given two persistence barcodes with small Wasserstein or bottleneck distance, their corresponding persistent entropy will be similar. Before proceeding with this task, we clarify through definitions and lemmas some concepts related with persistent entropy, which have not been treated rigorously so far. This allows us to link well-known Shannon entropy stability results of finite probability distribution with the new concept of persistent entropy.

3.1 Preliminary lemmas

In this subsection, we will define the subspaces and norms we are going to use in the sequel, and provide some relations between them.

From now on, given two persistence barcodes $A = \{[x_i^a, y_i^a]\}_{i=1}^{n_a}$ and $B = \{[x_i^b, y_i^b]\}_{i=1}^{n_a}$, denote $n_{\max} = \max\{n_a, n_b\}$, $\ell_i^a = y_i^a - x_i^a$, $L_a = \sum_i^{n_a} \ell_i^a$, $\ell_i^b = y_i^b - x_i^b$, $L_b = \sum_i^{n_b} \ell_i^a$, $L_{\max} = \max\{L_a, L_b\}$ and $L_{\min} = \min\{L_a, L_b\}$.

Remember that the value of $d_p(A, B)$ is reached for a concrete bijection γ between A and B (see Definition 2.7).

Remark 3.1. For simplicity of notation, in several proofs, we will sort the intervals of the persistence barcodes A and B in such a way that the bijection γ, for which the equality

$$d_p(A,B) = \left(\sum_{i=1}^{n} \max\left\{ |x_i^a - x_{\gamma(i)}^b|^p, |y_i^a - y_{\gamma(i)}^b|^p \right\} \right)^{\frac{1}{p}}$$

will be reached when γ is the identity: $\gamma_{Id}(i) = i$.

We first recall a well-known lemma regarding p-norms.

Lemma 3.2. *Let $x \in \mathbb{R}^n$ and $p, q \in \mathbb{R}$. Let $||x||_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ and $||x||_\infty = \max\{|x_i|\}$. If $1 \le q < p \le \infty$ then $||x||_p \le ||x||_q \le n^{\frac{1}{q}-\frac{1}{p}} ||x||_p$.*

Now we introduce some useful subspaces of the space $\mathcal{B}$ of persistence bacodes.

Definition 3.3 (Sets $\mathcal{B}_F, \mathcal{B}_0, \mathcal{B}_N$)**.** Define:

- the set of persistence barcodes with finite length intervals as $\mathcal{B}_F = \{A \in \mathcal{B} : \forall [x_i^a, y_i^a] \in A, \, y_i^a < \infty\}$;
- the set of persistence barcodes whose intervals were born at the origin as $\mathcal{B}_0 = \{A \in \mathcal{B} : \forall [x_i^a, y_i^a] \in A, \, x_i^a = 0\}$;
- the set of persistence barcodes with "normalized" intervals as $\mathcal{B}_N = \{A \in \mathcal{B} : \forall [x_i^a, y_i^a] \in A, \, \sum_i y_i^a - x_i^a = 1\}$.

Now let us extend Lemma 3.2 to the Wasserstein distance.

Corollary 3.4. *Let d_p be the p-th Wasserstein distance for persistence barcodes. If $A, B \in \mathcal{B}_F$ and $1 \le q < p \le \infty$ then $d_p(A,B) \le d_q(A,B) \le n_{\max}^{\frac{1}{q}-\frac{1}{p}} d_p(A,B)$.*

Proof. Sort the intervals of A and B such that $\gamma_{Id}(i) = i$ as in Remark 3.1. Then, the second inequality can be proven as follows:

$$\begin{aligned} d_q(A,B) &= \left(\min_\gamma \sum_i \max\left\{ |x_i^a - x_{\gamma(i)}^b|^q, |y_i^a - y_{\gamma(i)}^b|^q \right\} \right)^{\frac{1}{q}} \\ &\le \left(\sum_i \max\left\{ |x_i^a - x_i^b|^q, |y_i^a - y_i^b|^q \right\} \right)^{\frac{1}{q}} \\ &\le \left(n_{\max}^{\frac{1}{q}-\frac{1}{p}} \sum_i \max\left\{ |x_i^a - x_i^b|^p, |y_i^a - y_i^b|^p \right\} \right)^{\frac{1}{p}} = n_{\max}^{\frac{1}{q}-\frac{1}{p}} d_p(A,B). \end{aligned}$$

The first inequality $d_p(A,B) \le d_q(A,B)$ can be proven in an analogous way. □

Each persistence barcode in the space $\mathcal{B}_0 \cap \mathcal{B}_N \subset \mathcal{B}_F$ can be identified with a finite probability distribution $\{p_i\}_i$, which is the original domain of Shannon entropy. When we compute persistent entropy on barcodes in $\mathcal{B}_F$, we are indirectly using the following projections as lemma 3.6 shows.

Definition 3.5 (Projections π, ψ' and ψ)**.** Define:

$$\pi : \mathcal{B}_F \to \mathcal{B}_0 \cap \mathcal{B}_F \;\; \text{where } A = \{[x_i^a, y_i^a]\} \mapsto \pi(A) = \{[0, y_i^a - x_i^a]\};$$

$$\psi' : \mathcal{B}_0 \cap \mathcal{B}_F \to \mathcal{B}_0 \cap \mathcal{B}_N \;\; \text{where } A = \{[0, \ell_i^a]\} \mapsto \psi'(A) = \left\{ \left[0, \frac{\ell_i^a}{L_a} \right] \right\};$$

$$\psi : \mathcal{B}_F \to \mathcal{B}_0 \cap \mathcal{B}_N \;\; \text{where } \psi = \psi' \circ \pi.$$

Examples of these projections are showed in Figure 2.

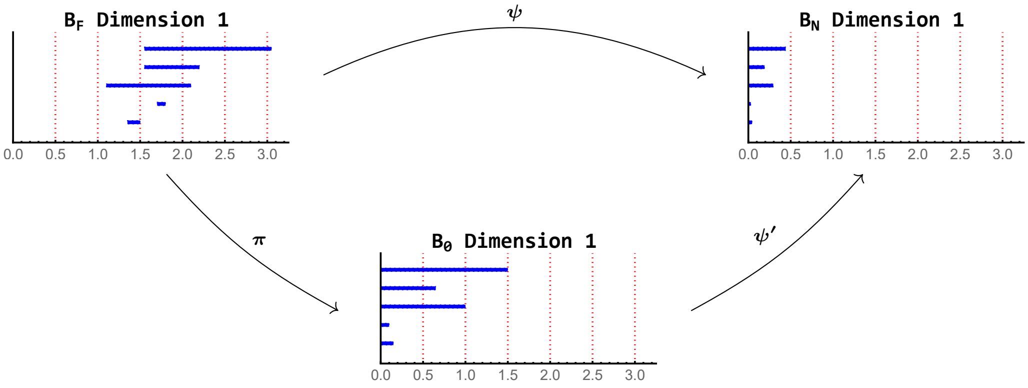


Figure 2: Examples of projections π, ψ' and ψ.

Lemma 3.6. *If $A \in \mathcal{B}_F$ then $E(\psi(A)) = E(A)$.*

The result below states that when we translate the intervals of given persistence barcodes A and B to the origin, the distance between them may double.

Lemma 3.7. *If $A, B \in \mathcal{B}_F$ and $1 \leq p \leq \infty$ then $d_p(\pi(A), \pi(B)) \leq 2d_p(A, B)$.*

Proof. Sort the intervals of A and B such that $\gamma_{Id}(i) = i$ as in Remark 3.1. Then, as $\pi(A) = \{0, y_i^a - x_i^a\}$ and $\pi(B) = \{0, y_i^b - x_i^b\}$, we have:

$$\begin{aligned}
d_p(\pi(A), \pi(B))^p &= \min_{\gamma} \sum_{i=1}^{n_{\max}} \max_i \left\{ 0, |(y_i^a - x_i^a) - (y_{\gamma(i)}^b - x_{\gamma(i)}^b)|^p \right\} \\
&= \min_{\gamma} \sum_{i=1}^{n_{\max}} |(y_i^a - x_i^a) - (y_{\gamma(i)}^b - x_{\gamma(i)}^b)|^p \leq \sum_{i=1}^{n_{\max}} |(y_i^a - x_i^a) - (y_i^b - x_i^b)|^p \\
&\leq \sum_{i=1}^{n_{\max}} \left(|(y_i^a - y_i^b)| + |(x_i^b - x_i^a)| \right)^p \leq \sum_{i=1}^{n_{\max}} \left(2 \max \left\{ |x_i^a - x_i^b|^p, |y_i^a - y_i^b|^p \right\} \right) \\
&= 2^p d_p(A, B)^p.
\end{aligned}$$

□

In order to fix what we consider "big" or "small" error, we give a general definition of the relative variation with respect to the average length.

Definition 3.8 (Relative error)**.** The relative variation for the p-th Wasserstein distance with coefficient $1 \leq p \leq \infty$ of $A, B \in \mathcal{B}_F$ is given by

$$r_p(A, B) = \frac{2n_{\max}^{1-\frac{1}{p}} d_p(A, B)}{L_{\max}} = \frac{2d_p(A, B)}{\ell_p}$$

where $\ell_p = L_{\max}/n_{\max}^{1-\frac{1}{p}}$ can be seen as a weighted average, depending on p, of the length of the intervals. For example, for $p = \infty$, ℓ_p is the standard average of the length of the intervals, and for $p = 1$, it is just the sum.

Let us see how projection ψ affects to the distance d_p.

Lemma 3.9. *If $A, B \in \mathcal{B}_F$ then $d_p\big(\psi(A), \psi(B)\big) \leq \frac{4n_{\max}^{1-\frac{1}{p}} d_p(A,B)}{L_{\max}} = 2r_p(A,B)$.*

Proof. Consider $\pi(A) = \{(0, \ell_i^a)\}_{i=1}^{n_a}$ and $\pi(B) = \{(0, \ell_i^b)\}_{i=1}^{n_b}$. Then,

$$d_p\big(\psi(A), \psi(B)\big)^p = \min_\gamma \sum_{i=1}^{n_{\max}} \left| \frac{\ell_i^a}{L_a} - \frac{\ell_{\gamma(i)}^b}{L_b} \right|^p = \min_\gamma \sum_{i=1}^{n_{\max}} \left| \frac{\ell_i^a L_b - \ell_{\gamma(i)}^b L_a}{L_a L_b} \right|^p .$$

Note that ℓ_i^a or ℓ_i^b might be 0 if intervals $[t,t]$ were needed for creating each bijection γ. If we sort the intervals of A and B such that $\gamma_{Id}(i) = i$ as in Remark 3.1, we obtain

$$d_p\big(\psi(A), \psi(B)\big)^p \leq \sum_{i=1}^{n_{\max}} \left| \frac{\ell_i^a L_b - \ell_i^b L_a}{L_a L_b} \right|^p .$$

We can suppose without loss of generality that $L_{\max} = L_a \geq L_b$. We have two cases: $\ell_i^a L_b \geq \ell_i^b L_a$ and $\ell_i^a L_b \leq \ell_i^b L_a$. In the first case:

$$\left| \frac{\ell_i^a L_b - \ell_i^b L_a}{L_a L_b} \right|^p = \left(\frac{\ell_i^a L_b - \ell_i^b L_a}{L_a L_b} \right)^p \leq \left(\frac{\ell_i^a L_b - \ell_i^b L_b}{L_a L_b} \right)^p = \left(\frac{\ell_i^a - \ell_i^b}{L_a} \right)^p .$$

The other case (i.e., when $\ell_i^a L_b \leq \ell_i^b L_a$) is slightly more difficult. First, using Lemma 3.2 we have:

$$0 \leq L_a - L_b = \sum_{i=1}^{n_{\max}} \ell_i^a - \ell_i^b \leq \sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b| \leq n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}} .$$

Therefore, $L_a \leq L_b + n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}}$. Using this expression, we obtain:

$$\begin{aligned}
\left| \frac{\ell_i^b L_b - \ell_i^a L_a}{L_a L_b} \right|^p &\leq \left(\frac{\ell_i^b \left(L_b + n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}} \right) - \ell_i^a L_b}{L_a L_b} \right)^p \\
&= \left(\frac{\ell_i^b - \ell_i^a}{L_a} + \frac{\ell_i^b n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}}}{L_a L_b} \right)^p .
\end{aligned}$$

This last value gives us a greater bound than the one before. Using it as the worst possible scenario we obtain:

$$\begin{aligned}
d_p(\psi(A), \psi(B))^p &= \min_\gamma \sum_{i=1}^{n_{\max}} \left| \frac{\ell_{\gamma(i)}^b}{L_b} - \frac{\ell_i^a}{L_a} \right|^p \\
&\leq \sum_{i=1}^{n_{\max}} \left(\frac{|\ell_i^b - \ell_i^a|}{L_a} + \frac{\ell_i^b n_{max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}}}{L_a L_b} \right)^p \\
&\leq \left(\sum_{i=1}^{n_{\max}} \frac{|\ell_i^b - \ell_i^a|}{L_a} + \sum_{i=1}^{n_{\max}} \frac{\ell_i^b n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}}}{L_a L_b} \right)^p \\
&= \left(\sum_{i=1}^{n_{\max}} \frac{|\ell_i^b - \ell_i^a|}{L_a} + \frac{n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}}}{L_a} \right)^p \\
&\leq \left(2\, \frac{n_{\max}^{1-\frac{1}{p}} \left(\sum_{i=1}^{n_{\max}} |\ell_i^a - \ell_i^b|^p \right)^{\frac{1}{p}}}{L_a} \right)^p = \frac{2^p n_{\max}^{p-1} d_p(\pi(A), \pi(B))^p}{L_a^p} .
\end{aligned}$$

In the third line we have used $|x|^p + |y|^p \leq (|x| + |y|)^p$ for $p \geq 1$ and in the fourth, Lemma 3.2. Finally, eliminating the exponent p in both sides of the inequality, writing $L_{\max}$ instead of L_a and applying Lemma 3.7 we obtain:

$$d_p(\psi(A), \psi(B)) \leq \frac{2n_{\max}^{1-\frac{1}{p}} d_p(\pi(A), \pi(B))}{L_{\max}} \leq \frac{4n_{\max}^{1-\frac{1}{p}} d_p(A, B)}{L_{\max}} = 2r_p(A, B).$$

□

3.2 Persistence barcodes with infinite intervals

As mentioned in Definition 2.12, persistent entropy is defined only for persistence barcodes with intervals of finite length. Nevertheless, it is quite common to find persistence barcodes with infinite intervals in practice and, depending on the application, they might be important or not. If they are, it will be interesting to define projections $\mathcal{B} \to \mathcal{B}_F$ that preserve the information carried by them. As we are seeking stability results, these projections must keep a control on the distance. A common approach is to change the infinite intervals by finite ones. For example, in [24] the endpoints of infinite intervals in each persistence barcode is sent to the maximum endpoint of the finite intervals of that barcode plus a constant. In this subsection, we formally define this projection and prove that, despite the distance between persistence barcodes might be modified by it, the variation can be controlled.

Definition 3.10 (Projection τ)**.** Let $A, B \in \mathcal{B}$ and let u_a and u_b be, respectively, the maximum endpoints of their finite intervals. Fix $C \geq 0$. Define projection

$$\tau : \mathcal{B} \to \mathcal{B}_F \;\text{ where } A = \{(x_i^a, y_i^a)\} \mapsto \tau(A) = \{(x_i^a, z_i^a)\}$$

being $z_i^a = u_a + C$ if $y_i^a = \infty$ and $z_i^a = y_i^a$ otherwise.

Recall that if the number of infinite intervals in A and B is different, then $d_p(A, B) = \infty$. In other case, we have the following result.

Lemma 3.11. *If two persistence barcodes $A, B \in \mathcal{B}$ have the same number m_∞ of infinite intervals then, for any p being $1 \leq p \leq \infty$, projection τ satisfies that*

$$d_p(\tau(A), \tau(B)) \leq \left(d_p(A, B)^p + m_\infty\, d_\infty(A, B)^p\right)^{\frac{1}{p}}.$$

Proof. Sort the intervals of A and B in such a way that their first m_∞ intervals are the infinite ones and for the rest, $\gamma_{Id}(i) = i$ as in Remark 3.1. We have:

$$\begin{aligned}
d_p(\tau(A), \tau(B))^p &= \min_{\gamma} \sum_{i=1}^{n_{\max}} \max\left\{|x_i^a - x_{\gamma(i)}^b|^p, |z_i^a - z_{\gamma(i)}^b|^p\right\} \\
&\leq \sum_{i=1}^{n_{\max}} \max\left\{|x_i^a - x_i^b|^p, |z_i^a - z_i^b|^p\right\} \\
&= \sum_{i=1}^{m_\infty} \max\left\{|x_i^a - x_i^b|^p, |u^a - u^b|^p\right\} + \sum_{i=1}^{n_{\max}-m_\infty} \max\left\{|x_i^a - x_i^b|^p, |y_i^a - y_i^b|^p\right\} \\
&= \sum_{i=1}^{m_\infty} \max\left\{|x_i^a - x_i^b|^p, |u^a - u^b|^p\right\} + d_p(A, B)^p - \sum_{i=1}^{m_\infty} |x_i^a - x_i^b|^p \\
&= \sum_{i=1}^{m_\infty} \max\left\{0, |u^a - u^b|^p - |x_i^a - x_i^b|^p\right\} + d_p(A, B)^p \leq m_\infty |u^a - u^b|^p + d_p(A, B)^p.
\end{aligned}$$

In the third equality, we have used that $\max\left\{|x_i^a - x_i^b|^p, |\infty - \infty|\right\} = |x_i^a - x_i^b|^p$ by definition (see Remark 2.8). Now, we only have to prove that $|u^a - u^b| \leq d_\infty(A, B)$. Assuming, without loss of generality, that $u^a \geq u^b$, suppose $u^a - u^b > d_\infty(A, B)$, as there was at least one interval with endpoint u^a. Then, there exists another endpoint x in B matched with u_a by the bottelneck distance with $|u^a - x| \leq d_\infty(A, B)$. Then,

$$u^a - x \leq d_\infty(A, B) < u^a - u^b \Rightarrow u^b < x.$$

By definition, $x \leq u^b$ leading to a contradiction. □

Another possible approach consists on sending the infinite intervals to a fixed value common for all persistence barcodes we are dealing with at that moment. The more importance we want to give to these intervals, the greater this value should be.

Definition 3.12 (Projection ϕ)**.** Let $\{A_j\}_j \subset \mathcal{B}$ be a finite family of persistence barcodes and let C be a constant greater than the maximum length of all finite intervals in the persistence barcodes of the given family. Fix p being $1 \leq p \leq \infty$. Define projection

$$\phi : \mathcal{B} \to \mathcal{B}_F \text{ where } A_j = \{(x_i^j, y_i^j)\} \mapsto \phi(A_j) = \{(x_i^j, z_i^j)\},$$

being $z_i^j = C$ if $y_i^j = \infty$ and $z_i^j = y_i^j$ otherwise.

The following lemma guarantees that projection ϕ does not increase (Wassertein or bottleneck) distance between persistence barcodes.

Lemma 3.13. *Consider a finite family of persistence barcodes $\{A_j\}_j \subset \mathcal{B}$. Then projection ϕ satisfies that $d_p(\phi(A_j), \phi(A_h)) \leq d_p(A_j, A_h), \ \forall j, h$.*

Proof. Fixing the bijection γ which gives the exact value of $d_p(A_j, A_h)$ we can deduce that $d_p(\phi(A_j), \phi(A_h))$ will be at most $d_p(A_j, A_h)$, since there could exist a different bijection giving a lower value of the p-th Wasserstein distance $d_p(\phi(A_j), \phi(A_h))$. □

3.3 Stability result

Two important results about the stability of persistent homology were recalled in Theorem 2.9 and Theorem 2.11. They guarantee that if two filter functions or two metric spaces are similar, then their corresponding persistence barcodes will be similar as well. There exist stability results for Shannon entropy when defined in a probability distribution. In order to combine these results to persistent entropy, we just need to adapt them to the metric space of persistence barcodes.

First of all, in [1] the continuity of persistent entropy with respect to the bottleneck distance is proven. The following proposition generalize this result to the Wasserstein distance.

Proposition 3.14. *Let $A, B \in \mathcal{B}_F$ and let d_p be the p-th Wasserstein distance with $1 \leq p \leq \infty$. If we fix the maximum numbers of intervals $n_{\max}$ and the minimum total length $L_{\min}$, then the persistent entropy E is continuous on $(\mathcal{B}_F, d_p)$:*

$$\forall \varepsilon \ \exists \delta \text{ such that } d_p(A, B) \leq \delta \Rightarrow |E(A) - E(B)| \leq \varepsilon.$$

Proof. The proof is straightforward using $d_\infty(A, B) \leq d_p(A, B)$ (Corollary 3.4). □

The stability problem of Shannon entropy has been previously studied by Lesche in [18] for the 1-norm due to its importance in physics. That bound can be slightly improved as shown in [10, p. 664].

Theorem 3.15 ([10, p. 664])**.** *Let P and Q be two finite probability distributions (seen as vectors in $\mathbb{R}^n$), and let $E_S(P)$ and $E_S(Q)$, respectively, their Shannon entropy. If $||P - Q||_1 \leq 1/2$ then*

$$|E_S(P) - E_S(Q)| \leq ||P - Q||_1 \log(n) - ||P - Q||_1 \log ||P - Q||_1$$

Note that the restriction $||P - Q||_1 \leq 1/2$ is reasonable because $||P - Q||_1$ is at most 2. Besides, since the space $\mathcal{B}_0 \cap \mathcal{B}_N$ can be interpreted as finite probability distributions, we can first project the persistence barcodes of $\mathcal{B}$ onto $\mathcal{B}_0 \cap \mathcal{B}_N$ and then apply the previous theorem to obtain the desired stability result.

Theorem 3.16 (Stability of Persistent Entropy)**.** *Consider $A, B \in \mathcal{B}_F$. If the relative error $r_p(A, B)$ is less than $1/4$ then $|E(A) - E(B)| \leq 2r_p(A, B) \left[\log(n_{\max}) - \log\left(2r_p(A, B)\right)\right]$.*

Proof. We first use Lemma 3.4 to transform the p-norm into the 1-norm. Then, we normalize the given persistence barcodes and apply Lemma 3.9 and Theorem 3.15. □

Table 1 shows some numerical examples. Despite the bound of $|E(A) - E(B)|$ may tend to infinity for arbitrary large $n_{\max}$, the relative value $\frac{|E(A)-E(B)|}{\log(n_{\max})}$ is bounded when $n_{\max}$ tends to infinity. In other words,

$$\lim_{n_{\max} \to \infty} \frac{|E(A) - E(B)|}{\log(n_{\max})} \leq 2r_p(A, B).$$

We can deduce the following two stability results using Theorem 2.9, Theorem 2.11 and Theorem 3.16.

	Relative error			
$\mathbf{n}_{\max}$	**0.1**	**0.05**	**0.025**	**0.01**
10	0.339794	0.2	0.115051	0.0539794
510	0.251631	0.136933	0.0740258	0.0325498
1010	0.246531	0.133285	0.0716526	0.0313102
1510	0.243975	0.131457	0.070463	0.0306888
2010	0.242321	0.130274	0.0696935	0.0302868
2510	0.24112	0.129415	0.0691346	0.0299949
3010	0.240187	0.128747	0.0687007	0.0297682
3510	0.239431	0.128206	0.0683486	0.0295843
4010	0.238798	0.127754	0.0680541	0.0294305
4510	0.238256	0.127366	0.067802	0.0292988
5010	0.237784	0.127028	0.0675823	0.029184

Table 1: Bounds of relative values $\frac{|E(A)-E(B)|}{\log(n_{\max})}$ for different number of intervals (columns) and relative errors r_∞ (rows).

Theorem 3.17. *Let K be a simplicial complex and let $f, g : K \to \mathbb{R}$ be two monotonic functions. Let $A, B \in \mathcal{B}$ be their corresponding persistence barcodes. The average length of the intervals of A and B are, respectively, $\ell^a_\infty = L_a/n_{\max}$ and $\ell^b_\infty = L_b/n_{\max}$. Let $\ell_{\max} = \max\{\ell^a_\infty, \ell^b_\infty\}$. If $d_\infty(A, B) \leq \frac{1}{8}\ell_{\max}$ then*

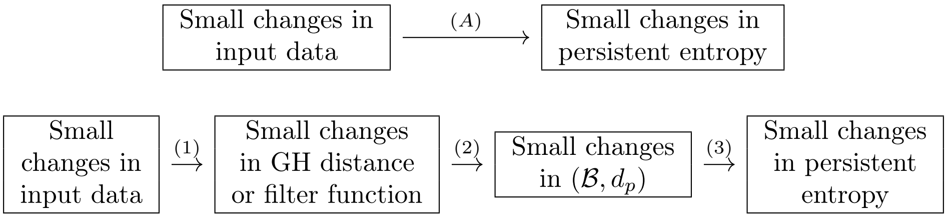
$$||f - g||_\infty \leq \delta \Rightarrow |E(A) - E(B)| \leq \frac{4\delta}{\ell_{\max}} \left[\log(n_{\max}) - \log\left(\frac{4\delta}{\ell_{\max}}\right)\right].$$

Theorem 3.18. *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 $Rips(Y, t)|_{t\in\mathbb{R}}$ respectively. The average length of the intervals of A and B are, respectively, $\ell^a_\infty = L_a/n_{\max}$ and $\ell^b_\infty = L_b/n_{\max}$. Let $\ell_{\max} = \max\{\ell^a_\infty, \ell^b_\infty\}$. If $d_\infty(A, B) \leq \frac{1}{8}\ell_{\max}$ then*

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

The condition $d_\infty(A, B) \leq \frac{1}{8}\ell_{\max}$ comes from imposing $r_p(A, B) \leq 1/2$. In order to remove the infinite intervals, we have to use projection ϕ (see Lemma 3.13).

It seems appropriate to recapitulate now the results of this section before continuing. As shown in the following diagram, at the beginning of the section we wanted to prove implication (A). In order to do it, we have separated the problem in three parts ((1), (2) and (3)):


Small changes in input data
(A)
Small changes in persistent entropy
Small changes in input data
(1)
Small changes in GH distance or filter function
(2)
Small changes in $(\mathcal{B}, d_p)$
(3)
Small changes in persistent entropy

Implication (1) is given by the formalization of the problem and implication (2) is given by Theorem 2.9 and Theorem 2.11 mentioned in the background section. The proof of implication (3) is the main aim of this section (Theorem 3.16). Putting all together we obtain Theorem 3.17 and Theorem 3.18.

4 Entropy-based summary functions

The simplicity of persistent entropy limits its application to distinguish persistence barcodes. Nevertheless, it is able to measure two interesting features simultaneously: the number of intervals and their heterogeneity. In order to obtain deeper statistical information from persistence barcodes we need to take a step forwards. Following this idea, different kinds of summary functions have been used in TDA to obtain statistical information from persistence barcodes such as silhouettes [7] and intensity maps [22]. In this section, we will define three summary piecewise constant functions using persistent entropy. The first associate to each moment of time a partial sum of the persistent entropy of the barcode. The second is the normalization of the first one. Both are stable respect to the bottleneck distance. The last one is designed to measure different features to the bottleneck distance, and then is not possible to apply the step (3) in the previous diagram. A more detailed study of this last function would have to be carried in another paper.

4.1 Entropy summary function (ES-function)

We define now a new function which pairs each barcode $A \in \mathcal{B}_F$ with a piecewise constant function (also known as step functions) in $\mathbb{R}$. This new function resumes information about the number of intervals and their homogeneity and, as we will prove at the end of this subsection, is stable with respect to the bottleneck distance. In Remark 2.5 we defined when a class was alive. Since the birth and death time of each class is encoded by intervals in a given persistence barcode A, we say that an interval $[x_i^a, y_i^a] \in A$ *is alive at* t if $x_i^a < t < y_i^a$.

Definition 4.1 (Entropy summary function (ES-function))**.** Consider a persistence barcode $A = \{[x_i^a,\ y_i^a]\}_{i=1}^{n_a}$ in $\mathcal{B}_F$. Define its entropy summary function (ES-function) as the piecewise linear function:

$$S(A)[t] = -\sum_{i=1}^{n_a} w_i^a(t)\frac{\ell_i^a}{L_a}\log\left(\frac{\ell_i^a}{L_a}\right)$$

where $w_i^a(t) = 1$ if $x_i^a \leq t \leq y_i^a$ and $w_i^a(t) = 0$ otherwise.

Note that $S : \mathcal{B}_F \to \mathcal{C}$ and $S(A) : \mathbb{R} \to \mathbb{R}$, being $\mathcal{C}$ the space of piecewise constant functions.

Remark 4.2. ES-function pairs the instant t and the persistence barcode A with the partial sum of $E(A)$ corresponding to the intervals of A that are alive at that moment t. See Figure 3.

The following result states that ES-function is stable with respect to the bottleneck distance.

Theorem 4.3 (Stability of the ES-function)**.** *Let S be the ES- function, d_∞ the bottleneck distance and A, B two persistence barcodes in $\mathcal{B}_F$. If the relative error $r_\infty(A,B)$ is less or equal than* $1/4$*, then we have:*

$$\begin{aligned} ||S(A)[t] - S(B)[t]||_1 &\leq 2L_{\min} r_\infty(A,B)\log\left[2r_\infty(A,B)\right] + 2d_\infty(A,B)\log n_{\max} \\ &\leq 2r_\infty(A,B)\left(L_{\min}\log\left[2r_\infty(A,B)\right] + L_{\max}\frac{\log n_{\max}}{n_{\max}}\right). \end{aligned}$$

Remark 4.4. Recall that $||f||_1 = \int_{\mathbb{R}} |f(t)|dt$ for a given function $f : \mathbb{R} \to \mathbb{R}$. Notice that all functions appearing in this subsection are bounded and have compact support in $\mathbb{R}$ therefore their 1-norm is always finite.

Proof. Let us prove the first inequality. Sort the intervals of A and B such that $\gamma_{Id}(i) = i$ as in Remark 3.1. Note that $w_i^a(t) = w_i^a(t)w_i^b(t) + w_i^a(t)(1 - w_i^b(t))$. Denote the expression $\frac{\ell_i^a}{L_a}\log\left(\frac{\ell_i^a}{L_a}\right)$ by s_i^a. Then:

$$\begin{aligned} &||S(A) - S(B)||_1 = \\ &\quad = \left|\left|\sum_{i=1}^{n_{\max}} (w_i^a(t)w_i^b(t) + w_i^a(t)(1 - w_i^b(t)))s_i^a - (w_i^b(t)w_i^a(t) + w_i^b(t)(1 - w_i^a(t)))s_i^b\right|\right|_1 \\ &\quad = \left|\left|\sum_{i=1}^{n_{\max}} w_i^a(t)w_i^b(t)(s_i^a - s_i^b) + w_i^a(t)(1 - w_i^b(t))s_i^a - w_i^b(t)(1 - w_i^a(t))s_i^b\right|\right|_1 \\ &\quad \leq \sum_{i=1}^{n_{\max}} \left|\left|w_i^a(t)w_i^b(t)\right|\right|_1 |s_i^a - s_i^b| + \left|\left|w_i^a(t)(1 - w_i^b(t))s_i^a\right|\right|_1 + \left|\left|w_i^b(t)(1 - w_i^a(t))s_i^b\right|\right|_1. \end{aligned}$$

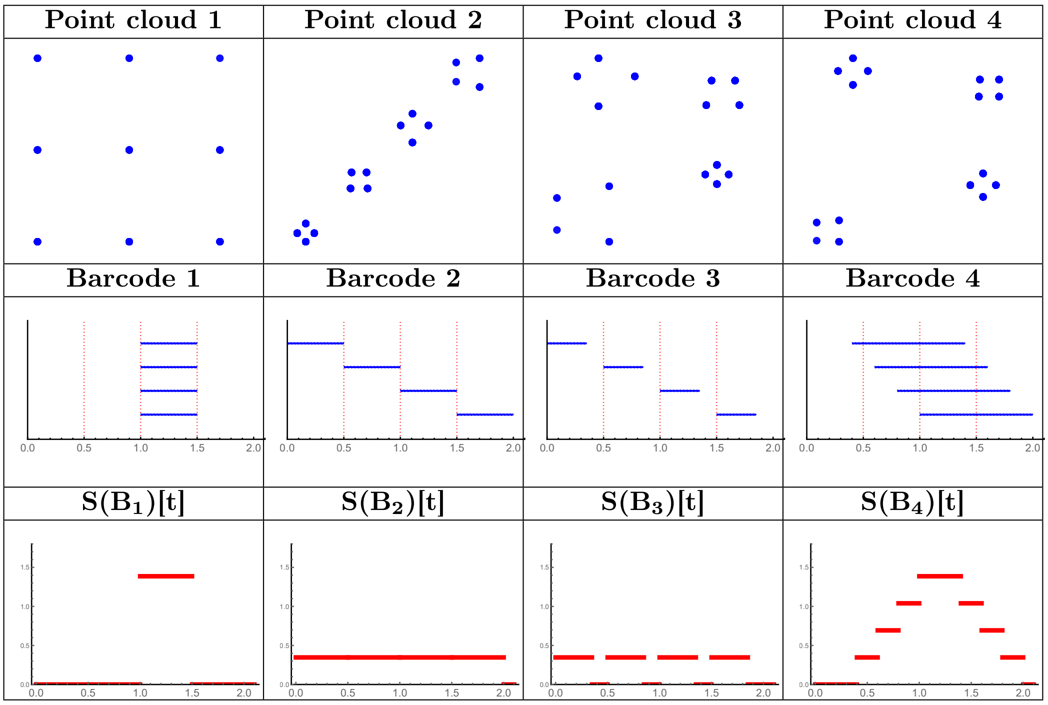


Figure 3: In the first row we show four different point clouds. The 1-st persistence barcodes of their associated Vietoris-Rips filtration appear in the second row. Observe that all of them have the same persistent entropy $E = 1.38629$. In the third row we can see their ES-function for each instant t. Note that all of them are different in spite of the fact that the four persistent barcodes have the same persistent entropy.

We first compute a bound for $\sum_{i=1}^{n_{\max}} \big|\big|w_i^a(t)w_i^b(t)\big|\big|_1 |s_i^a - s_i^b|$. Note that

$$\sum_{i=1}^{n_{\max}} ||w_i^a(t)w_i^b(t)||_1 \leq \sum_{i=1}^{n_{\max}} \min\{\ell_i^a, \ell_i^b\} \leq L_{\min}.$$

And since function $-x \log x$ is concave then $|x_1 - x_2| \leq \epsilon \Rightarrow |-x_1 \log x_1 + x_2 \log x_2| \leq -\epsilon \log \epsilon$. In this case,

$$\epsilon = \max\left\{\frac{\ell_i^a}{L_a} - \frac{\ell_i^b}{L_b}\right\} \leq \frac{4n_{\max} d_\infty(A,B)}{L_{\max}} = 2r_\infty(A,B)$$

by Lemma 3.9, and then $|s_i^a - s_i^b| \leq 2r_\infty(A,B) \log\left(2r_\infty(A,B)\right)$. Therefore,

$$\sum_{i=1}^{n_{\max}} \big|\big|w_i^a(t)w_i^b(t)\big|\big|_1 |s_i^a - s_i^b| \leq 2L_{\min} r_\infty(A,B) \log\left(2r_\infty(A,B)\right). \tag{1}$$

Now, we calculate the bound for $\sum_{i=1}^{n_{\max}} \big|\big|w_i^a(t)(1-w_i^b(t))s_i^a\big|\big|_1 + \big|\big|w_i^b(t)(1-w_i^a(t))s_i^b\big|\big|_1$. Consider the function $w_i^b(t)(1-w_i^a(t))$. Its integral gives the period of time in which the i-th interval of B is alive and the i-th interval of A is not. This might happen in both: the initial and the end of the period of time. Therefore, if $\epsilon_i = \max\{|x_i^a - x_i^b|, |y_i^a - y_i^b|\}$ then:

$$\int_{\mathbb{R}} w_i^b(t)(1-w_i^a(t))dt \leq 2\epsilon_i \leq 2d_\infty(A,B).$$

Note that both intervals cannot be the only ones alive in both extreme of the period of time simultaneously, therefore we also have

$$\epsilon_i \leq \int_{\mathbb{R}} w_i^b(t)(1-w_i^a(t))dt \leq 2\epsilon_i \Rightarrow \int_{\mathbb{R}} w_i^a(t)(1-w_i^b(t))dt = 0.$$

and vice versa. Using $\sum_{i=1}^{n_{\max}} s_i^a = E(A)$ we can deduce:

$$\begin{aligned}
&\sum_{i=1}^{n_{\max}} \left|\left|w_i^a(t)(1-w_i^b(t))\right|\right|_1 s_i^a + \left|\left|w_i^b(t)(1-w_i^a(t))\right|\right|_1 s_i^b\\
&\leq \sum_{i=1}^{n_{\max}} s_i^a \int_{\mathbb{R}} w_i^a(t)(1-w_i^b(t)) + s_i^b \int_{\mathbb{R}} w_i^b(t)(1-w_i^a(t))\\
&\leq \max\left\{ \sum_{i=1}^{n_{\max}} \epsilon_i(s_i^a+s_i^b), \sum_{i=1}^{n_{\max}} 2\epsilon_i s_i^a, \sum_{i=1}^{n_{\max}} 2\epsilon_i s_i^b \right\}\\
&\leq \max\left\{ d_\infty(A,B)[E(A)+E(B)], 2d_\infty(A,B)E(A), 2d_\infty(A,B)E(B) \right\}\\
&\leq d_\infty(A,B)\max\left\{ [E(A)+E(B)], 2E(A), 2E(B) \right\} = d_\infty(A,B) 2\log n_{\max}. \qquad (2)
\end{aligned}$$

Putting together (1) and (2) we obtain the desired bound. Using the definition of r_∞ we can deduce the second inequality presented in the theorem. □

When n tends to infinity, we can deduce from the theorem above that:

$$\lim_{n_{max}\to\infty} ||S(A)-S(B)||_1 \leq 2L_{\min} r_\infty(A,B)\log\left[2r_\infty(A,B)\right].$$

4.2 Normalized entropy summary function (NES-function)

One of the main aims of persistent homology is to represent the shape of the input data set. In some applications, like image analysis or material science (see [4] for a review), it may be important to detect some repetitive pattern independently of the size of the input data set. In this particular case, a comparison in the metric space $(\mathcal{B}, d_p)$ is not useful due to its strongly dependence on the number of long intervals. Our aim now is to create a function to distinguish patterns independently of the number of intervals.

Definition 4.5 (Normalized entropy summary function (NES-function)). Consider a persistence barcode $A = \{[x_i^a, y_i^a]\}_{i=1}^{n_a}$ in $\mathcal{B}_F$. Normalized entropy summary function (NES-function) of A is defined as:

$$NES(A)[t] = \frac{S(A)[t]}{||S(A)||_1}.$$

We show examples of repetitive patterns in Figure 4. In the first row, the first two images indicate different patterns both given by quadrilaterals. The second and the third images have the same pattern but different number of points. In the second row, in each image, we take the vertices of the quadrilaterals and use the Vietoris-Rips filtration to obtain the corresponding persistence barcodes. In the third row, 30% of points are displaced or removed, with respect to the second row. The result of computing persistent homology and NES-function on both examples is shown in Figure 5 and Figure 6, we can observe that NSE-function seems to be robust to noise and to the number of points in the pattern. This observation is supported by the 1-norm distance matrix showed in Figure 7.

Theorem 4.6 (Stability of the NES-function). *Under the same conditions appearing in theorem 4.3, we have*

$$\begin{aligned}
||NES(A)-NES(B)||_1 &\leq \frac{||S(A)-S(B)||_1}{\min\{||S(A)||_1, ||S(B)||_1\}}\\
&\leq \frac{2r_\infty(A,B)\left(L_{\min}\log\left[2r_\infty(A,B)\right] + L_{\max}\frac{\log n_{\max}}{n_{\max}}\right)}{min\left\{||S(A)||_1, ||S(B)||_1\right\}}.
\end{aligned}$$

Proof. It is straight forward.

$$\begin{aligned}
\left|\left|\frac{S(A)}{||S(A)||_1} - \frac{S(B)}{||S(B)||_1}\right|\right|_1 &= \frac{\left|\left|||S(B)||_1 S(A) - ||S(A)||_1 S(B)\right|\right|_1}{||S(A)||_1||S(B)||_1} \leq \frac{\max\{||S(A)||_1||S(B)||_1\}}{||S(A)||_1||S(B)||_1}\\
&= \frac{||S(A)-S(B)||_1}{\min\{||S(A)||_1, ||S(B)||_1\}}.
\end{aligned}$$

□

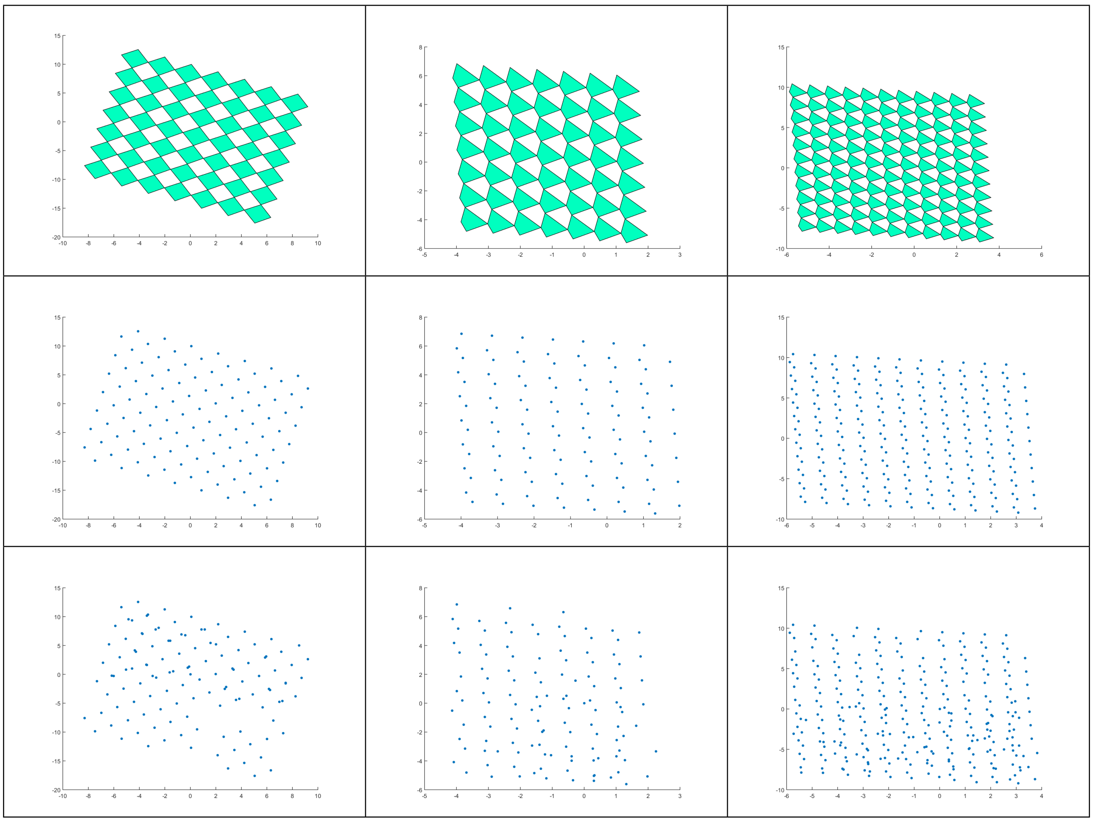


Figure 4: The first row shows quadrilaterals forming a pattern; in the second, the points of the quadrilateral are pictured; and, in the last row, noise is added to the point cloud data. The first two columns have similar number of points but different pattern while the last two have the same pattern but different number of points.

4.3 Time-based entropy summary function (TES-function) for detecting topological features

A direct consequence of the work carried out by Hausmann [16] and Latschev [17] is that if a point cloud in a manifold is dense enough then its Vietoris-Rips filtration will be homotopic to the manifold during a period of time I. Nevertheless, it is not possible to compute that period of time in practice. This problem is classically sorted out using persistent homology and considering long intervals, in the corresponding persistence barcode, as topological features. In this subsection, we will define the time-based entropy summary function F which is a modified version of ES-function, to automatically distinguish topological features from noise, and locate the periods of time where the corresponding Vietoris-Rips complexes may be homotopic to a manifold.

In order to achieve this goal, our function will pair each moment t with a higher value when the classes represented by the alive intervals in that time are topological features. Usually, this happens when:

- The length of the alive intervals at the moment t are big and similar.
- Few intervals are alive at that moment.
- The period of time these intervals are the only ones alive is long.

ES-function S satisfies the first condition but not the others. In order to get the second condition, we can obtain the "average contribution" to the persistent entropy, of each interval in the persistence barcode, dividing S by the number of intervals alive at the moment t, $W_a(t)$. Besides, for the third condition we can weight these contributions multiplying S by the period of time for which the set of alive intervals keep unchanged with respect to the moment t, $T_a(t)$.

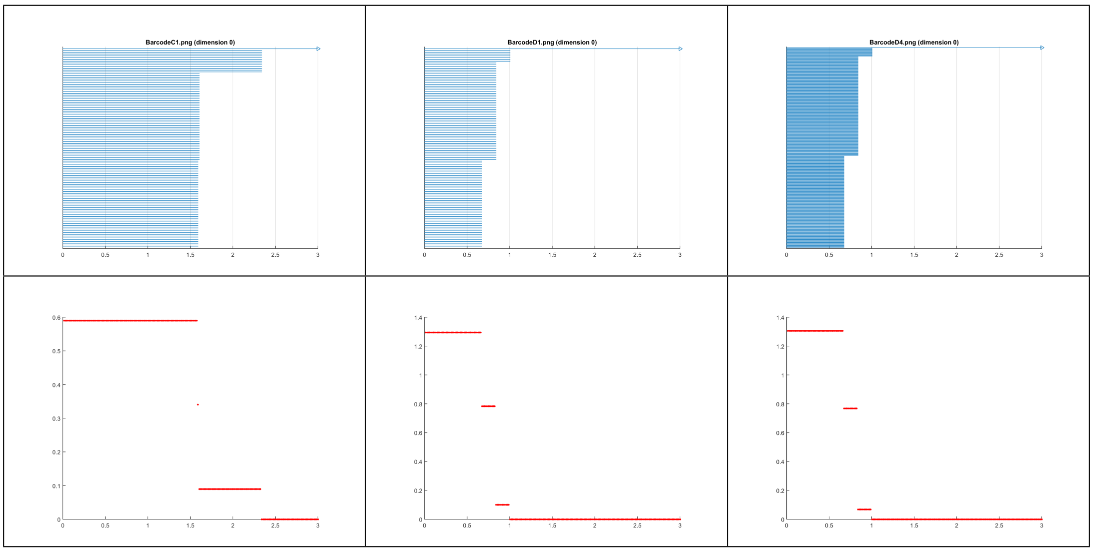


Figure 5: Persistence barcodes of the Vietoris-Rips filtration (fist row) and NES-function on the persistence barcodes (second row), all associated to the point clouds (without noise) showed in the second row of Figure 4.

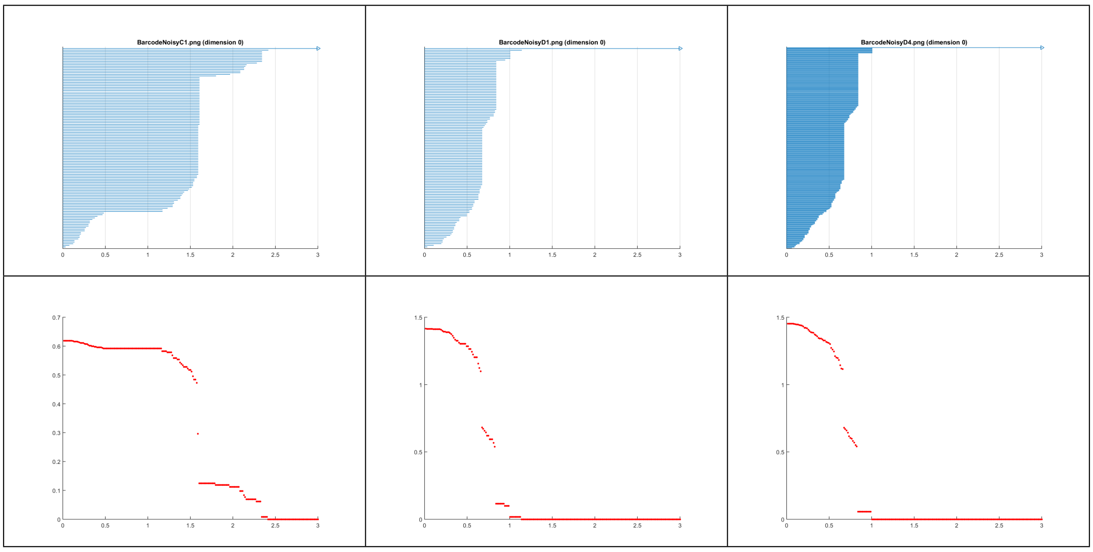


Figure 6: Persistence barcodes of the Vietoris-Rips filtration (first row) and NES-function on the persistence barcodes (second row), all associated with the point clouds (with noise) showed in the third row of Figure 4.

Definition 4.7 (Time-based entropy summary function (TES-function))**.** Consider a persistence barcode $A = \{[x_i^a, y_i^a]\}_{i=1}^{n_a}$ in $\mathcal{B}_F$. Time-based entropy summary function (TES-function) of A is defined as:

$$F(A)[t] = \frac{T_a(t)}{W_a(t)} S(A)[t],$$

where $W_a(t) = \sum_{i=1}^{n_a} w_i^a(t)$, and $T_a(t) = \max\big\{s_1 > 0 : |w_i^a(t) - w_i^a(t+\lambda s_1)| = 0,\ \forall \lambda \in [0,1];\ \forall i = 1 \dots n_a\big\} + \max\big\{s_2 > 0 : |w_i^a(t) - w_i^a(t-\lambda s_2)| = 0,\ \forall \lambda \in [0,1];\ \forall i = 1 \dots n_a\big\}$ is the period of time during which the alive intervals in t persist.

Function $A \mapsto T_a$ is not continuous with respect to the bottleneck distance then $F(A)$ is not continuous neither. The main reason for this fact is that bottleneck distance ignores noise while F is sensitive to it since it is designed with the purpose of detecting topological features.

We use now the circle S^1 as a toy example to study the potential use of TES-function. In fact,

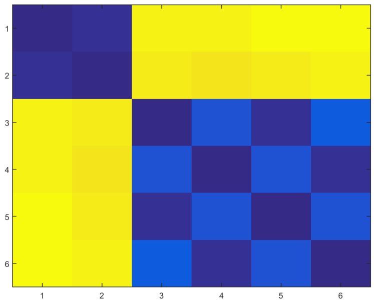


Figure 7: 1-norm distance matrix of NES-function on the persistence barcodes associated to the six point clouds showed in Figure 4. Observe that NES-function distinguishes the patterns but not the sizes of the point clouds.

the circle S^1 is a manifold whose Vietoris-Rips complexes exhibit a rich behavior. In [19], it is proved that, if X is a dense subset of S^1 and $m \in \mathbb{Z}^*$ then

$$Rips(X,t) \simeq S^{2m+1} \quad \text{when} \quad \frac{m}{2m+1} < t \leq \frac{m+1}{2m+3}.$$

Finite samples are expected to approximate this result. We use TES-function associated to the Vietoris-Rips complex of finite samples of the circle to verify it.

The methodology is as follows: Compute persistence barcodes, up to dimension 5, associated to finite samples of points in the circle and then apply TES-function. For each sample of the circle, take the intervals corresponding to the highest values of its TES-function as topological features, ignoring the contractible case which is: $\beta_0 = 1, \beta_i = 0 \; \forall i > 0$.

In our experiments, we have tested TES-function on the persistence barcodes associated to nine point clouds with 40 points. Two of these point clouds are shown in Figure 8. We have observed that we always obtain the Betti numbers of the circle S^1 as the main topological feature. Betti numbers of S^3 appears three times as the second most important, two times as the third, the forth and the sixth; indicating they are topological features of the filtration independent of the distribution of the points. The rest of the topological features depend on the distribution of the points in the point cloud and consist of $\beta_0 > 1$ or, occasionally, $\beta_2 = 1$ and $\beta_0 = 1$. Betti numbers of S^5 do not appear as important topological features since the point clouds are not dense enough to generate it or it appears with a very short length.

Notice this method is an automatized process. It would be interesting to see how it responds to extremal cases, for example, finding out the minimum number of points needed to recognize the homology of the circle S^1. In Figure 9 we show two examples of a new test with 9 point clouds of 10 points each. In this case, five of them detect the homology of S^1 as the main feature. In two of the remaining ones, no cycle is created due to the point distribution. When the experiment is done with 8 points, its persistent homology usually does not find any cycle and therefore the process does not detect the homology of S^1.

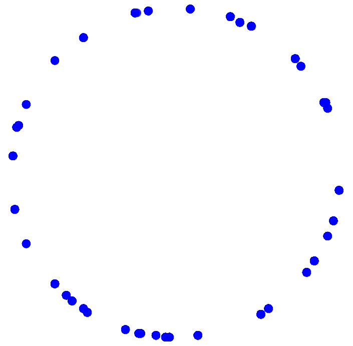
5 Conclusions and future work

We have proved the stability of persistent entropy justifying its application in topological data analysis. What is more, we have used persistent entropy to define an stable summary function called ES-function. We have constructed, from it, two new summary functions called NES-function (to distinguish different patterns) and TES-function (to detect topological features).

The computations carried in the paper has been done using the package "TDA" for R (see [13]), and Javaplex for Matlab (see [26]). Besides, the graphics have been generated using both, Wolfram Mathematica and Matlab. The code used for generating the examples can be found in http://grupo.us.es/cimagroup/.

As future work, a hypothetical stability result for NES- and TES-functions would have to be developed. New experiments should be carried in order to get a deep insight of the properties of the functions and their possible applications.

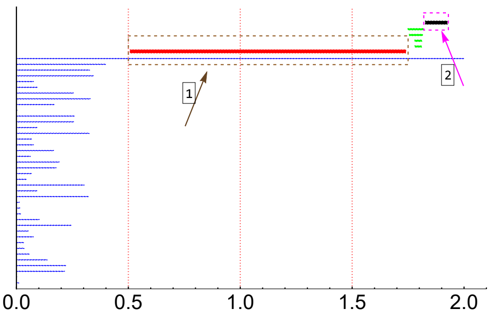
Point Cloud 1



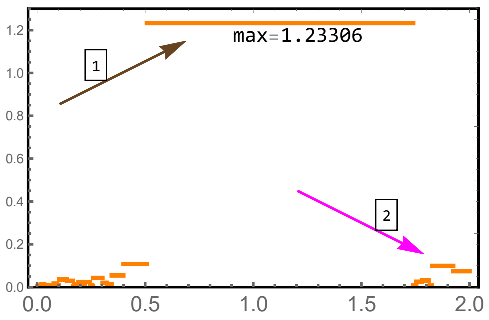
Topological features 1

1. $\beta_1 = \mathbf{1}$, $\beta_0 = \mathbf{1}$ (S_1)
2. $\beta_3 = \mathbf{1}$, $\beta_0 = \mathbf{1}$ (S_3)
3. ~~$\beta_0 = \mathbf{1}$~~
4. ~~$\beta_0 = \mathbf{1}$~~
5. $\beta_0 = \mathbf{2}$
6. $\beta_0 = \mathbf{16}$

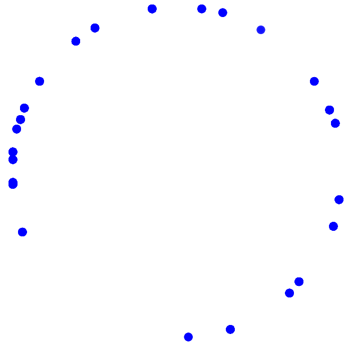
Persistence barcode 1



TES-function 1



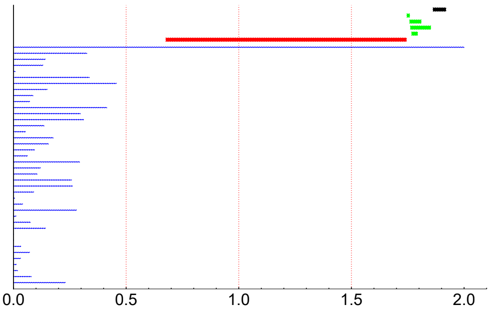
Point Cloud 2



Topological features 2

1. $\beta_1 = \mathbf{1}$, $\beta_0 = \mathbf{1}$ (S_1)
2. ~~$\beta_0 = \mathbf{1}$~~
3. ~~$\beta_0 = \mathbf{1}$~~
4. $\beta_0 = \mathbf{3}$
5. $\beta_0 = \mathbf{2}$
6. $\beta_3 = \mathbf{1}$, $\beta_0 = \mathbf{1}$ (S_3)

Persistence barcode 2



TES-function 2

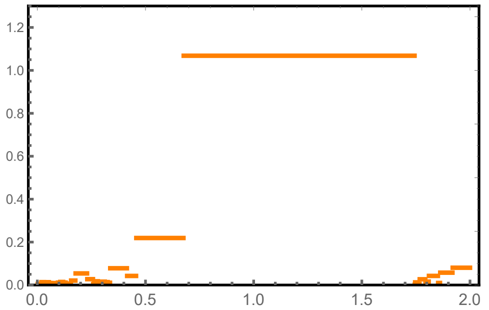


Figure 8: On the left (row 1 and row 3), point clouds with 40 points each are displayed. On the right (same rows), the topological features of their associated Vietoris-Rips filtrations are shown in order of importance in accordance with the values of TES-function. In the first example, the maximum value of **TES-function 1**, marked with the arrow one, $max = 1.23306$, is reached in the period of time $(0.51, 1.7)$ which corresponds to the time when there are two intervals of 0-th (blue) and 1-th (red) dimensions in **Persistence barcode 1**. These two intervals are compatibles with the homology of the circle S^1 which appears in **Topological features 1** in the first place. The next bigger value, pointed by the arrow two, corresponds with the Betti numbers of S_3. We always discard the contractible case $\beta_0 = 1$, $\beta_i = 0$ for $i > 0$ as a topological feature.

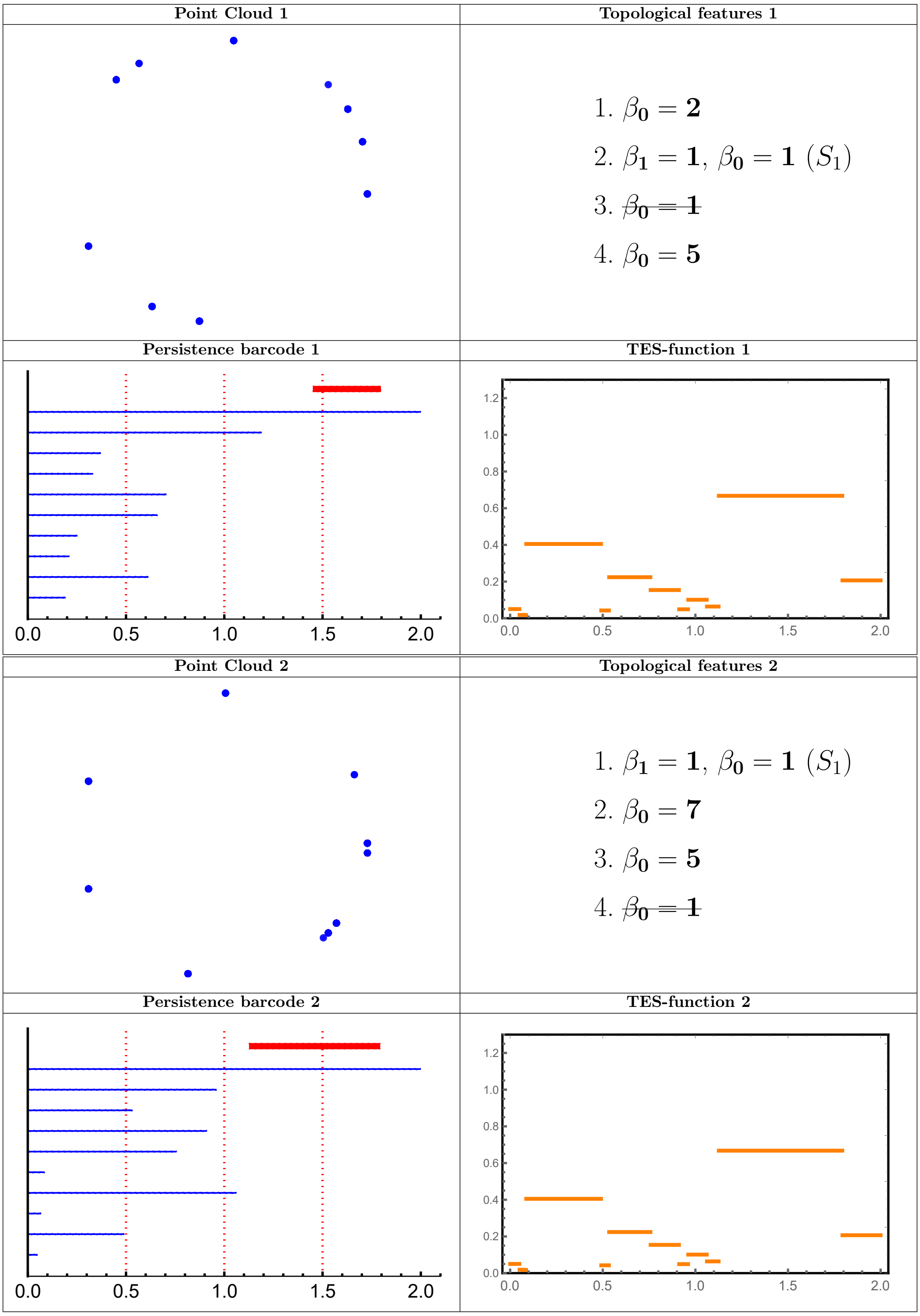


Figure 9: In this case, contrary to Figure 8, the point clouds only have 10 points. If the points appear forming clusters, the TES function may consider them more important than the fact of being contained in a circle.

References

[1] N. Atienza, R. Gonzalez-Diaz, and M. Rucco. Persistent entropy for separating topological features from noise in vietoris-rips complexes. *Journal of Intelligent Information Systems*, (7):1–19, 2017.

[2] J. Binchi, E. Merelli, M. Rucco, G. Petri, and F. Vaccarino. jholes: A tool for understanding biological complex networks via clique weight rank persistent homology. *Electronic Notes in Theoretical Computer Science*, 306:5–18, 2014.

[3] P. Bubenik. Statistical topology using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015.

[4] M. Buchet, Y. Hiraoka, and I. Obayashi. Persistent homology and materials informatics. In I.Tanaka, editor, *Nanoinformatics*, pages 75–95. Springer, Singapore, 2018.

[5] G. Carlson, A. Zomorodian, A. Collins, and L.J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(02):149–187, 2005.

[6] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Mémoli, and S.Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28(5):1393–1403, 2009.

[7] F. Chazal, B.T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *Journal of Computational Geometry*, 6(2):140–161, 2015.

[8] H. Chintakunta, T. Gentimis, R. Gonzalez-Diaz, M. J. Jimenez, and H. Krim. An entropy based persistence barcode. *Pattern Recognition*, 48(2):391–401, February 2015.

[9] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, , and Y. Mileyko. Lipschitz functions have l_p-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, April 2010.

[10] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, 2nd edition, 2006.

[11] H. Edelsbrunner and J.L. Harer. *Computational Topology: An Introduction.* American Mathematical Society, 1st edition, 2010.

[12] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28(4):511–533, November 2002.

[13] B.T. Fasy, J. Kim, F. Lecci, C. Maria, V. Rouvreau . The included GUDHI is authored by Clement Maria, Dionysus by Dmitriy Morozov, PHAT by Ulrich Bauer, Michael Kerber, and Jan Reininghaus. *TDA: Statistical Tools for Topological Data Analysis*, 2017. R package version 1.6.

[14] M. Ferri. Persistent topology for natural data analysis — a survey. *Towards Integrative Machine Learning and Knowledge Extraction. Lecture Notes in Computer Science. Springer*, 10344(2):127–139, 2017.

[15] A. Hatcher. *Algebraic Topology.* Cambridge University Press, 1st edition, 2002.

[16] J.C. Hausmann. On the vietoris–rips complexes and a cohomology theory for metric spaces. In F. Quinn, editor, *Prospects in topology : proceedings of a conference in honor of William Browder*, volume 138, page 175–188. Princeton, N.J. Princeton, N.J. : Princeton University Press, 1995.

[17] J. Latschev. Vietoris-rips complexes of metric spaces near a closed riemannian manifold. *Archiv der Mathematik*, 77(6):522–528, 2001.

[18] B. Lesche. Instabilities of rényi entropies. *Journal of Statistical Physics*, 27(2):419–422, 1982.

[19] H. Adams M. Adamaszek. The vietoris-rips complexes of a circle. *Pacific Journal of Mathematics*, 290(1):1–40, 2017.

[20] E. Merelli, M. Piangerelli, M. Rucco, and D. Toller. A topological approach for multivariate time series characterization: the epileptic brain. *EAI Endorsed Transactions on Self-Adaptive Systems*, 16(7), 5 2016.

[21] N. Otter, M.A. Porter, U. Tillmann, P. Grindrod, and H.A. Harrington. A roadmap for the computation of persistent homology. *Entropy*, 17(6), 2017.

[22] P. Pranav, H. Edelsbrunner, R. van de Weygaert, G. Vegter, M. Kerber, B.J.T. Jones, and M. Wintraecken. The topology of the cosmic web in terms of persistent betti numbers. *Monthly Notices of the Royal Astronomical Society*, 465(4):4281–4310, March 2017.

[23] M. Rucco, F. Castiglione, E. Merelli, and M. Pettini. Characterisation of the idiotypic immune network through persistent entropy. In S. Battiston, F. De Pellegrini, G. Caldarelli, and E. Merelli, editors, *Proceedings of ECCS 2014. Proceedings of ECCS 2014*, pages 117–128. Springer Proceedings in Complexity, 2014.

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

[25] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[26] Tausz, Andrew, Vejdemo-Johansson, Mikael, Adams, and Henry. JavaPlex: A research software package for persistent (co)homology. In H. Hong and C. Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at