AIMS Mathematics, 10(9): 21080–21103.
DOI:10.3934/math.2025942
Received: 03 June 2025
Revised: 31 July 2025
Accepted: 13 August 2025
Published: 15 September 2025

https://www.aimspress.com/journal/Math

Research article

Cosine similarity and orthogonality of persistence diagrams

Azmeer Nordin*, Mohd Salmi Md Noorani, Nurulkamal Masseran*, Mohd Sabri Ismail and Nur Firyal Roslan

Department of Mathematical Sciences, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

* **Correspondence:** Email: nfazmeer.nordin@ukm.edu.my, kamalmsn@ukm.edu.my.

Abstract: Topological data analysis is an approach to study the shape of a data set by means of topology. Its main object of study is the persistence diagram, which represents the topological features of the data set at different spatial resolutions. Multiple data sets can be compared by the similarity of their diagrams to understand their behaviors relative to each other. The bottleneck and Wasserstein distances are often used as a tool to indicate the similarity. In this paper, we introduce the cosine similarity as a new indicator for the similarity between persistence diagrams and investigate its properties. Furthermore, it leads to the new notion of orthogonality between persistence diagrams. It turns out that the orthogonality refers to perfect dissimilarity between persistence diagrams under the cosine similarity. Through data demonstration, the cosine similarity is shown to be more accurate than the standard distances to measure the similarity between persistence diagrams.

Keywords: persistence diagram; persistence landscape; cosine similarity; orthogonality; persistent homology; topological data analysis
Mathematics Subject Classification: 46C05, 55N31, 68T09

1. Introduction

Topological data analysis is a fast-growing field for analyzing complex data via theories and techniques from topology. It extracts information from the topological structure or shape of the data which may be obscured by traditional methods in data analysis. It is adaptable to various types of data and also robust to noise [1]. Due to its advantages, it has seen wide applications in diverse fields [2, 3], such as enhancing classical machine learning models [4], identifying cancer tumors in oncology [5], detecting warning signs of crashes in financial markets [6, 7], and recognizing patterns in multimedia data [8]. These exemplify only a fraction of the possibilities, as the list remains inexhaustible.

The central tool in topological data analysis is the persistent homology [9], which is the application

of homology theory on a data set. In particular, it constructs the shape of the data set via a certain mathematical structure called a simplicial complex, and then computes its topological features such as connected components, holes, and cavities at different spatial resolutions. The existences of those features along the resolutions are recorded in a scatter plot, which is referred to as a persistence diagram. Further analysis is conducted on the diagram to obtain meaningful information depending on purposes of the study.

Many studies pertaining to applications of persistent homology will involve working with multiple data sets. Their persistence diagrams can be compared with each other in a certain way to examine whether the data sets have similar structures or behaviors. For instance, this is useful to distinguish an abnormal data set under an extreme event in a particular situation. This is demonstrated in some studies of air pollution based on daily data of Air Quality Index, in which any month with a haze episode is characterized by having persistence diagram with abundant features associated to holes [10, 11]. The comparison between persistence diagrams in these works are done simply by observing their numbers of features without the need of other machinery.

Although there are more, the above example is enough to emphasize the importance of comparing persistence diagrams to understand the behaviors of data sets in question. To be more precise, we compare diagrams in terms of their similarity, i.e., how close the points of one diagram to those of another one on the same axes. A pair of diagrams is considered similar if both collections of points are sufficiently close, while dissimilar if both are too far apart. This depends on the tool or indicator used to quantify similarity and ascertain whether two diagrams are similar or dissimilar.

The common way to measure the similarity is to use the bottleneck and Wasserstein distances for persistence diagrams [9, 12]. Specifically, two diagrams are deemed similar if either distance is small. This idea of similarity via distances has been utilized in [13] to detect causality between two variables via bottleneck distance, and in [14] to classify clusters between several data sets via Wasserstein distance for topological clustering.

However, in certain cases, the distances may not be accurate to signify the similarity. It is possible for two persistence diagrams which are closely similar to have a noticeably large distance, while those that are clearly dissimilar may record a very small distance (see Examples 3.1 and 3.2). For this reason, a few alternative tools or indicators have been designed in the literature to measure the similarity between persistence diagrams more accurately, such as the RST parametric model [15] and the vectorization of persistence diagrams into complex polynomials [16].

In this paper, we propose a new indicator called cosine similarity as an alternative to measure the similarity between persistence diagrams. This is motivated by the same notion in Euclidean space, in which the cosine similarity for non-zero vectors $\vec{u}, \vec{v} \in \mathbb{R}^d$ is defined as

$$S_C(\vec{u}, \vec{v}) = \cos\alpha = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}||||\vec{v}||},$$

where α denotes the angle between them. It indicates how close the vectors are aligned in the same direction regardless of their magnitudes. It is widely applied in text mining to measure the similarity between two documents in terms of subject matter [17], and in data mining to measure the cohesion within data clusters [18].

In our approach, we employ persistence landscapes [19,20], which are an alternative presentation of persistence diagrams, to use the above formula for our cosine similarity. Each landscape is a sequence

of piecewise linear functions, and hence can be associated to a certain norm and inner product, unlike a persistence diagram. The norm for these landscapes is commonly utilized as a moving summary in persistent homology analysis of time series in various situations such as river water levels [21, 22] and financial data [23, 24], but the applications of its inner product are seldom explored in the literature.

As a related result here, the inner product motivates a new notion of orthogonality between persistence diagrams. In particular, we say that two diagrams are orthogonal if their landscapes admit a zero inner product. It turns out that the orthogonality refers to perfect dissimilarity between two persistence diagrams under the cosine similarity, i.e., both diagrams are considered entirely different from each other.

To sum up, the aim of this paper is to introduce the cosine similarity for persistence diagrams and investigate its properties, which are discussed in Section 3. Moreover, we explore the notion of orthogonality for the diagrams and provide some justifications on why it means perfect dissimilarity under the cosine similarity in Section 4. Finally, we compare the accuracy of cosine similarity against the bottleneck, Wasserstein, and other relevant distances to measure similarity of persistence diagrams through a data demonstration in Section 5.

2. Persistent homology

This section provides a brief introduction to the persistent homology which covers some common notions such as simplicial complexes, homology groups, filtrations, persistence diagrams, and persistence landscapes. A detailed account of this theory can be referred to [9, 12].

Let P be a non-empty finite set. It shall be called a *point set*. An *abstract simplicial complex* $\mathcal{K}$ is a collection of non-empty subsets of P such that for any set $\sigma \in \mathcal{K}$, its non-empty subset $\tau \subseteq \sigma$ is also in $\mathcal{K}$. Each set $\sigma \in \mathcal{K}$ of size $k+1$ is called a *k-simplex*. Its non-empty subset $\tau \subseteq \sigma$ of size $\ell+1$ is called an *ℓ-face* of σ. A $(k-1)$-face of σ is called a *facet*.

A complex $\mathcal{K}$ can be geometrically visualized in $\mathbb{R}^{|P|-1}$ by associating an element of P to the zero vector while others to standard unit vectors. A simplex $\sigma \in \mathcal{K}$ corresponds to the convex hull of the relevant vectors in $\mathbb{R}^{|P|-1}$. Hence, a simplex can be thought as a geometrical shape under this visualization. For example, a 0-simplex, a 1-simplex, a 2-simplex, and a 3-simplex are a point, an edge, a solid triangle and a solid tetrahedron, respectively. This visualization is called a geometric realization of $\mathcal{K}$. This is useful if we want to relate certain topological features to a simplicial complex, such as connected components and holes. We shall mention that there is another notion of a complex for a point set in an Euclidean space, which is called a geometric simplicial complex, but it is not required in this paper.

A function $f : \mathcal{K} \to \mathbb{R}$ is said to be *simplex-wise monotone* if $f(\tau) \leq f(\sigma)$ whenever τ is a face of σ. In this case, the subcollection $\mathcal{K}_\theta = \{\sigma \in \mathcal{K} \mid f(\sigma) \leq \theta\}$ is a subcomplex of $\mathcal{K}$ for any $\theta \in \mathbb{R}$. It is clear that $\mathcal{K}_\theta \subseteq \mathcal{K}_{\theta'}$ for $\theta \leq \theta'$. So, the collection of complexes $\{\mathcal{K}_\theta \mid \theta \geq \theta_0\}$, where $\theta_0 = \min f$, forms a nested inclusion as θ increases. It is called the simplicial filtration of $\mathcal{K}$ by f.

For example, Vietoris-Rips filtration is often considered in topological data analysis due to its simple construction. In this case, the complex $\mathcal{K}$ is taken as the power set of P. Suppose that P is equipped with a distance d. We define a simplex-wise monotone function f on $\mathcal{K}$ as

$$f(\sigma) = \max_{p,q \in \sigma} \mathsf{d}(p, q).$$

The subcomplex $\mathcal{K}_\theta$ is called the *Vietoris-Rips complex* of parameter θ. In particular, a subset $\sigma \subseteq P$ is in $\mathcal{K}_\theta$ if and only if $\mathrm{d}(p,q) \leq \theta$ for any pair $p, q \in \sigma$. The *Vietoris-Rips filtration* is the collection $\{\mathcal{K}_\theta \mid \theta \geq 0\}$ in this case.

Given a filtration, we are interested to observe the changes in the topological structure of $\mathcal{K}_\theta$ as θ increases. For this purpose, we shall introduce the homology groups induced by a general complex $\mathcal{K}$. Let $\mathcal{S}_k = \{\sigma_1, \sigma_2, \ldots, \sigma_{n_k}\}$ be a collection of k-simplices in $\mathcal{K}$. We define a formal series of $\mathcal{S}_k$ over the field $\mathbb{Z}_2$ as $\sum_{i=1}^{n_k} a_i\sigma_i$ where $a_i \in \mathbb{Z}_2$. The series is called a *k-chain*. The collection of all k-chains of $\mathcal{K}$ forms an abelian group, which is called the k^{th} *chain group* $\mathsf{C}_k(\mathcal{K})$. We set $\mathsf{C}_k(\mathcal{K})$ as the trivial group if $\mathcal{S}_k$ is empty.

For a k-simplex σ, consider the collection of its facets $\{\tau_1, \tau_2, \ldots, \tau_{k+1}\}$. We denote $\partial_k(\sigma) = \sum_{i=1}^{k+1} \tau_i$, which is a $(k-1)$-chain of $\mathcal{K}$. We extend this map as a homomorphism $\partial_k : \mathsf{C}_k(\mathcal{K}) \to \mathsf{C}_{k-1}(\mathcal{K})$ by defining

$$\partial_k\left(\sum_{i=1}^{n_k} a_i\sigma_i\right) = \sum_{i=1}^{n_k} a_i\partial_k(\sigma_i).$$

The kernel of ∂_k is called the k^{th} cycle group $\mathsf{Z}_k(\mathcal{K})$, while the image of ∂_{k+1} is called the k^{th} boundary group $\mathsf{B}_k(\mathcal{K})$. The k^{th} *homology group* is then defined as the quotient group $\mathsf{H}_k(\mathcal{K}) = \mathsf{Z}_k(\mathcal{K})/\mathsf{B}_k(\mathcal{K})$.

Each homology group relates to a certain type of topological features in a complex $\mathcal{K}$. For example, the zeroth and first homology groups reflect the connected components and holes respectively in $\mathcal{K}$, or to be precise, in its geometric realization [25].

Now, consider a simplicial filtration $\{\mathcal{K}_\theta \mid \theta \geq \theta_0\}$. Since $\mathcal{K}_\theta \subseteq \mathcal{K}_{\theta'}$ for $\theta \leq \theta'$, there is a natural homomorphism $\phi_k^{\theta,\theta'} : \mathsf{H}_k(\mathcal{K}_\theta) \to \mathsf{H}_k(\mathcal{K}_{\theta'})$ given by

$$\phi_k^{\theta,\theta'} : c + \mathsf{B}_k(\mathcal{K}_\theta) \mapsto c + \mathsf{B}_k(\mathcal{K}_{\theta'})$$

for any chain $c \in \mathsf{Z}_k(\mathcal{K}_\theta)$. A non-trivial class in $\mathsf{H}_k(\mathcal{K}_\theta)$ may be reduced to be trivial in $\mathsf{H}_k(\mathcal{K}_{\theta'})$ under this map. This motivates the idea of *birth* and *death* of a non-trivial class in the filtration. It describes a topological feature in $\mathcal{K}_\theta$ as mentioned previously. For this reason, we shall refer to a non-trivial class in the filtration simply as a *feature*.

A feature ξ is born at $b \in \mathbb{R}$ if $\xi \in \mathsf{H}_k(\mathcal{K}_b)$ and $(\phi_k^{\theta,b})^{-1}(\xi) = \emptyset$ for any $\theta < b$. Next, it dies at $d \in \mathbb{R}$ if $\phi_k^{b,d}(\xi) = \mathbf{0}$, but $\phi_k^{b,\theta}(\xi) \neq \mathbf{0}$ for any $\theta < d$. It is possible that ξ never dies, i.e., $\phi_k^{b,\theta}(\xi) \neq \mathbf{0}$ for any $\theta > b$. We denote $d = \infty$ in this case. The lifespan of ξ is defined as $\ell = d - b$.

From now on, we focus on the features with respect to the k^{th} homology group. Let ξ be a feature that is born at b and dies at d. It can be represented as a point (b, d). The persistence diagram D is the set of points (allowing multiplicities) for those features. It can be visualized as a scatter plot of deaths against births of the features. Note that the points in that plot must lie above the diagonal line $\Delta = \{(\theta, \theta) \mid \theta \geq \theta_0\}$.

A feature whose point is relatively far from the diagonal line is said to be *persistent*. Such a feature has a long lifespan. If the point set is sampled from a manifold, then this feature, such as a connected component or a hole, is expected to exist in that manifold. On the other hand, an impersistent feature in the diagram is simply considered a noise due to the filtration, and the abundance of such features may indicate that the sample itself is affected by noise [26].

We shall introduce two types of distances for persistence diagrams D_1 and D_2. For this purpose, we shall construct a bijection between the diagrams in a certain way. We allow them to add a finite number

of points from Δ (allowing multiplicities) to create new sets D_1^Δ and D_2^Δ, and construct a bijection $\gamma : D_1^\Delta \to D_2^\Delta$. For example, we define γ as follows:

$$\gamma : (b_1, d_1) \mapsto \left(\frac{b_1 + d_1}{2}, \frac{b_1 + d_1}{2}\right) \quad \text{and} \quad \gamma : \left(\frac{b_2 + d_2}{2}, \frac{b_2 + d_2}{2}\right) \mapsto (b_2, d_2)$$

for all points $(b_1, d_1) \in D_1$ and $(b_2, d_2) \in D_2$. We call this bijection a *trivial matching*. It pairs each point in D_1 and D_2 to the closest point in Δ.

Denote the set of those bijections as Γ. We allow the expression $\infty - \infty = 0$ here. The bottleneck distance is defined as

$$W_\infty(D_1, D_2) = \inf_{\gamma \in \Gamma} \left(\sup_{x \in D_1^\Delta} \|x - \gamma(x)\|_\infty \right),$$

where $\|\cdot\|_\infty$ denotes the supremum norm in $\mathbb{R}^2$. Another alternative is the *p-Wasserstein distance* for $p \geq 1$, which is given by

$$W_p(D_1, D_2) = \inf_{\gamma \in \Gamma} \left(\sum_{x \in D_1^\Delta} \|x - \gamma(x)\|_\infty{}^p \right)^{1/p}.$$

The bijection γ that attains the infimum for either distance is called a *perfect matching*.

The set of possible persistence diagrams do not form a vector space as there is no natural operation that can be defined on the set. Several attempts have been made in the literature to transform those diagrams into suitable vectors in a Banach or Hilbert space. This is called a *vectorization* of persistence diagrams. One such transformation is the persistence landscape [19, 20].

For this purpose, we assume that each feature has a finite death. For example, this assumption holds in Vietoris-Rips filtration, except for the zeroth homology group. In this case, there is exactly one feature that never dies [27]. It is often omitted in further analysis beyond the persistence diagram.

Suppose that a persistence diagram D has n features. For a feature ξ_i with birth b_i and death d_i, we construct a function $g_i : \mathbb{R} \to \mathbb{R}$ as

$$g_i(t) = \begin{cases} t - b_i & \text{if } t \in [b_i, (b_i + d_i)/2], \\ d_i - t & \text{if } t \in [(b_i + d_i)/2, d_i], \\ 0 & \text{otherwise.} \end{cases}$$

Next, we define the j^{th} layer function $\lambda_j : \mathbb{R} \to \mathbb{R}$ as

$$\lambda_j(t) = \max{}^{(j)} \{g_1(t), g_2(t), \ldots, g_n(t)\},$$

where $\max^{(j)}$ denotes the j^{th} maximum of the above set. For clarity, we set $\lambda_j(t) = 0$ for $j > n$. The vector $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots)$ is called the persistence landscape of D. Denote the transformation as $\varphi : D \mapsto \vec{\lambda}$. It is proved in [28] that φ is an injection from the set $\mathcal{D}$ of possible persistence diagrams to its image $\varphi(\mathcal{D})$. So, every persistence landscape corresponds to a unique persistence diagram.

We shall note that the set $\varphi(\mathcal{D})$ itself is not a vector space, but it can be embedded into a relevant normed space. For a persistence landscape $\vec{\lambda}$, its *p-norm* for $p \geq 1$ and supremum norm are given respectively as

$$\|\vec{\lambda}\|_p = \left(\sum_{j=1}^{\infty} \left(\int \left|\lambda_j(t)\right|^p dt \right) \right)^{1/p},$$

and

$$\|\vec{\lambda}\|_\infty = \sup_{j,t} \left|\lambda_j(t)\right|.$$

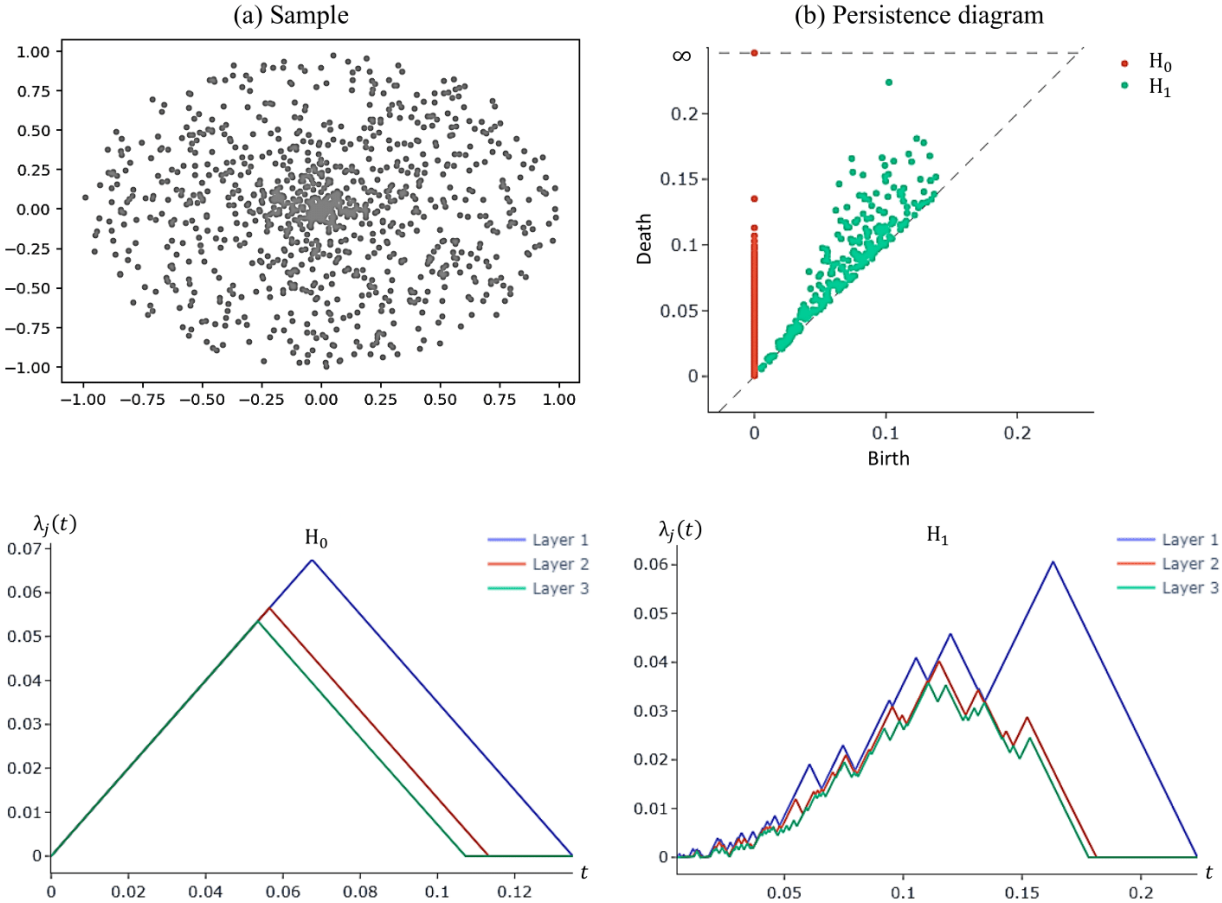
For $p = 2$, the norm induces an inner product as follows:

$$\langle \vec{\lambda}, \vec{\mu} \rangle = \sum_{j=1}^{\infty} \left(\int \lambda_j(t) \cdot \mu_j(t)\, dt \right),$$

where $\vec{\lambda}$ and $\vec{\mu} = (\mu_1, \mu_2, \ldots)$ are persistence landscapes.

The persistence diagram and landscape are stable; that is, they are robust against a small perturbation onto the point set P [1,20]. In such a situation, they differ slightly in terms of their distances or norms.

Figure 1 illustrates the persistence diagram and landscape for a sample from the unit disc in $\mathbb{R}^2$. The notations H_0 and H_1 below refer to the respective collection of features with respect to the zeroth and first homology groups in the Vietoris-Rips filtration of the sample.



(c) Persistence landscape (first three layer functions)

Figure 1. Persistence diagram and landscape for a sample from the unit disc in $\mathbb{R}^2$.

3. Similarity between persistence diagrams

We concentrate on the idea of similarity between persistence diagrams in this section. As mentioned previously, the similarity means how close the points in one diagram are to those of another one. We will introduce and explore the properties of cosine similarity as an indicator for the similarity. First, we shall explain why distances may not be an accurate approach for this purpose.

3.1. Issues with distances

Recall that the bottleneck and Wasserstein distances are commonly utilized to indicate the similarity between two persistence diagrams. In particular, these distances measure the degree of difference between two diagrams. The distances defined by the norms of their persistence landscapes can be considered for this purpose as well.

It is generally thought that two diagrams are closely similar whenever their distance is close to zero. However, this assertion is actually false in certain cases. It is possible for two diagrams to be almost equal, but their distance is significantly larger than zero. Conversely, two diagrams that are entirely different may record a very small distance. These are illustrated in the following examples.

Example 3.1. *Figure 2 shows two persistence diagrams*

$$D_1 = \{(\theta, \theta + 1) \mid \theta = 0, 1, \ldots, n-1\} \quad \textit{and} \quad D_2 = D_1 \cup \{(n, n+1)\}$$

for some $n \in \mathbb{N}$.

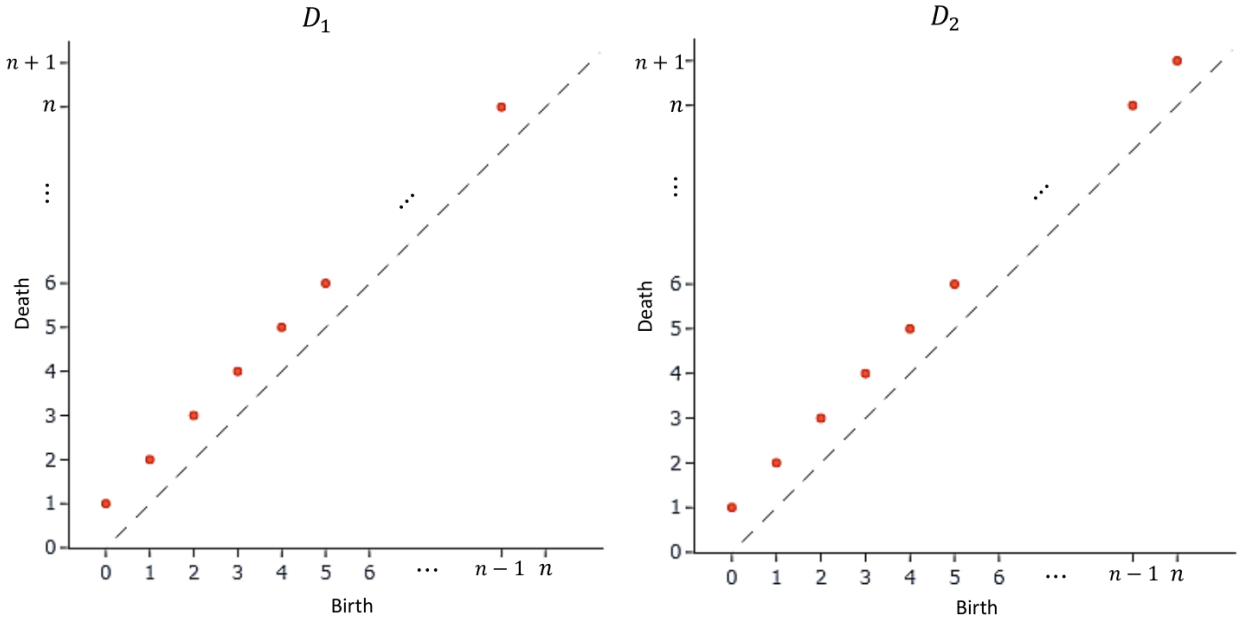


Figure 2. Persistence diagrams D_1 and D_2.

It is easy to verify the following:

$$W_\infty(D_1, D_2) = W_p(D_1, D_2) = \frac{1}{2},$$

$$\|\varphi(D_1) - \varphi(D_2)\|_\infty = \frac{1}{2} \quad \text{and} \quad \|\varphi(D_1) - \varphi(D_2)\|_p = \frac{1}{2}\left(\frac{1}{p+1}\right)^{\frac{1}{p}}.$$

The above distances are non-zero and constant for any n. We shall expect a significant difference between both diagrams based on the distances. However, they have n common points and differ only by a point. As n increases, the diagrams have a higher proportion of common points and thus are considered almost equal. However, this is not reflected by the distances. This example demonstrates that two diagrams can be very similar, but their distances are notably larger than zero.

Example 3.2. *Figure 3 shows two persistence diagrams*

$$D_3 = \left\{ \left(\theta + \frac{1}{2} - \frac{1}{2m}, \theta + \frac{1}{2} + \frac{1}{2m} \right) \middle| \ \theta = 0, 1, \ldots, n-1 \right\},$$

and

$$D_4 = \left\{ \left(\theta + \frac{1}{2} - \frac{1}{2m}, \theta + \frac{1}{2} + \frac{1}{2m} \right) \middle| \ \theta = 2n, 2n+1, \ldots, 2n+(n-1) \right\}$$

for some $m, n \in \mathbb{N}$ on the same plot.

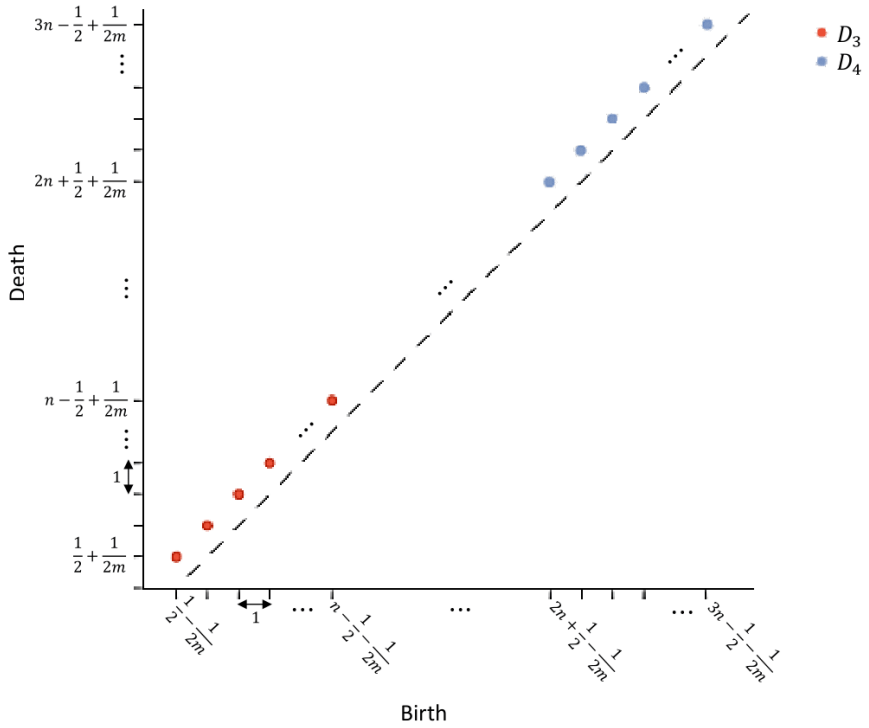


Figure 3. Persistence diagrams D_3 and D_4.

Simple calculations show that

$$W_\infty(D_3, D_4) = \frac{1}{2m}, \qquad W_p(D_3, D_4) = \frac{(2n)^{\frac{1}{p}}}{2m},$$

$$\|\varphi(D_3) - \varphi(D_4)\|_\infty = \frac{1}{2m} \quad \text{and} \quad \|\varphi(D_3) - \varphi(D_4)\|_p = \frac{1}{2m}\left(\frac{2n}{m(p+1)}\right)^{\frac{1}{p}}.$$

By fixing n, the distances approach zero as m increases. We shall expect both diagrams to be closely similar based on the distances. However, for sufficiently large n, the points of D_3 are clearly very far apart from those of D_4 for all m. Hence, both diagrams can be considered entirely different. This example shows that two diagrams can be dissimilar, but their distances are very small.

3.2. Cosine similarity

The previous argument shows that distances may not be accurate to indicate similarity between persistence diagrams in some cases. Due to this reason, there is a need for a better indicator for the similarity. We introduce the cosine similarity for this purpose. In this section, the notation $\|\cdot\|$ refers to the 2-norm for persistence landscapes.

Definition 3.3. *For non-empty persistence diagrams D_1 and D_2, the* (landscape) cosine similarity ς *is defined as*

$$\varsigma(D_1, D_2) = \frac{\langle \varphi(D_1), \varphi(D_2) \rangle}{\|\varphi(D_1)\| \cdot \|\varphi(D_2)\|}.$$

The assumptions on D_1 and D_2 are required in the above definition to avoid any zero norm in the denominator. Only the empty persistence diagram is mapped to the zero landscape under φ whose norm is zero.

We investigate the range of ς. It is clear that ς is non-negative. The Cauchy-Schwarz inequality [29] implies that

$$\langle \varphi(D_1), \varphi(D_2) \rangle \leq \|\varphi(D_1)\| \cdot \|\varphi(D_2)\|. \tag{3.1}$$

This proves that $\varsigma \in [0, 1]$. Intuitively, from the above inequality, the cosine similarity calculates the proportion of inner product over its possible maximum value. Here, the values 0 and 1 correspond to the extreme cases for ς, which shall be called the *perfect dissimilarity* and *perfect similarity* between diagrams, respectively. We shall interpret the meaning of these cases in detail.

The case $\varsigma(D_1, D_2) = 1$ corresponds to equality in (3.1). This holds if and only if $\varphi(D_2) = \alpha \cdot \varphi(D_1)$ for a real $\alpha > 0$. Based on the definition of a persistence landscape, each layer function is piecewise linear with slope 0, 1, or -1. This implies $\alpha = 1$. Since φ is injective, the equation $\varphi(D_2) = \varphi(D_1)$ is equivalent to $D_2 = D_1$. Thus, the perfect similarity under ς simply means the equality of persistence diagrams.

The case $\varsigma(D_1, D_2) = 0$ indicates a zero inner product between persistence landscapes of D_1 and D_2. In this instance, the diagrams shall be referred as orthogonal. We will discuss the notion of orthogonality in Section 4 and justify why it signifies perfect dissimilarity between persistence diagrams.

The cosine similarity is stable against a small perturbation in persistence diagrams.

Proposition 3.4. *Let D_1 and D_2 be non-empty persistence diagrams. Fix a real $\eta > 0$ such that*

$$\eta^2 = \min\left\{1, \frac{\|\varphi(D_1)\|^2}{8}\left(W_\infty(D_1, \emptyset) + \frac{1}{3}\right)^{-1}, \frac{\|\varphi(D_2)\|^2}{8}\left(W_\infty(D_2, \emptyset) + \frac{1}{3}\right)^{-1}\right\}.$$

Then, there are positive constants c_1 and c_2 which depend on D_1 and D_2 such that

$$\left|\varsigma(\tilde{D}_1, \tilde{D}_2) - \varsigma(D_1, D_2)\right| \leq c_1 \cdot W_2(D_1, \tilde{D}_1) + c_2 \cdot W_2(D_2, \tilde{D}_2)$$

for any pair of non-empty persistence diagrams $\tilde{D}_1$ and $\tilde{D}_2$ with

$$W_2(D_1, \tilde{D}_1) \leq \eta \quad \text{and} \quad W_2(D_2, \tilde{D}_2) \leq \eta.$$

Proof. Recall the equation for an inner product as follows:

$$||\varphi(D_1) - \varphi(D_2)||^2 = ||\varphi(D_1)||^2 + ||\varphi(D_2)||^2 - 2\langle \varphi(D_1), \varphi(D_2) \rangle . \tag{3.2}$$

We rearrange this equation to obtain

$$2 \cdot \varsigma(D_1, D_2) = \frac{||\varphi(D_1)||}{||\varphi(D_2)||} + \frac{||\varphi(D_2)||}{||\varphi(D_1)||} - \frac{||\varphi(D_1) - \varphi(D_2)||^2}{||\varphi(D_1)|| \cdot ||\varphi(D_2)||}.$$

The same holds for $\varsigma(\tilde{D}_1, \tilde{D}_2)$. Hence,

$$\begin{aligned} 2\left|\varsigma(\tilde{D}_1, \tilde{D}_2) - \varsigma(D_1, D_2)\right| \leq & \left|\frac{||\varphi(\tilde{D}_1)||}{||\varphi(\tilde{D}_2)||} - \frac{||\varphi(D_1)||}{||\varphi(D_2)||}\right| + \left|\frac{||\varphi(\tilde{D}_2)||}{||\varphi(\tilde{D}_1)||} - \frac{||\varphi(D_2)||}{||\varphi(D_1)||}\right| \\ & + \left|\frac{||\varphi(\tilde{D}_1) - \varphi(\tilde{D}_2)||^2}{||\varphi(\tilde{D}_1)|| \cdot ||\varphi(\tilde{D}_2)||} - \frac{||\varphi(D_1) - \varphi(D_2)||^2}{||\varphi(D_1)|| \cdot ||\varphi(D_2)||}\right|. \end{aligned}$$

We shall find a suitable bound for each modulus in the right-hand side (RHS) of the above inequality.

For each $i \in \{1, 2\}$, it is proved in [19] that

$$||\varphi(\tilde{D}_i) - \varphi(D_i)||^2 \leq 2\left(W_\infty(D_i, \emptyset) + \frac{1}{3}\right) W_2(D_i, \tilde{D}_i)^2, \tag{3.3}$$

if $W_2(D_i, \tilde{D}_i) \leq 1$. We apply the triangle inequality and the assumption on η to the above result to obtain

$$\left|||\varphi(\tilde{D}_i)|| - ||\varphi(D_i)||\right| \leq \frac{||\varphi(D_i)||}{2},$$

or, equivalently,

$$\frac{||\varphi(D_i)||}{2} \leq ||\varphi(\tilde{D}_i)|| \leq \frac{3\,||\varphi(D_i)||}{2}. \tag{3.4}$$

With these findings, we use the triangle inequality, (3.3), and (3.4) to determine the bound for the first modulus as follows:

$$\begin{aligned} & \left|\frac{||\varphi(\tilde{D}_2)||}{||\varphi(\tilde{D}_1)||} - \frac{||\varphi(D_2)||}{||\varphi(D_1)||}\right| \\ & \quad \leq \frac{||\varphi(D_2)|| \cdot \left|||\varphi(\tilde{D}_1)|| - ||\varphi(D_1)||\right| + ||\varphi(D_1)|| \cdot \left|||\varphi(\tilde{D}_2)|| - ||\varphi(D_2)||\right|}{||\varphi(D_2)|| \cdot ||\varphi(\tilde{D}_2)||} \\ & \quad \leq \frac{2\,||\varphi(D_2)|| \cdot \left|||\varphi(\tilde{D}_1)|| - ||\varphi(D_1)||\right| + 2\,||\varphi(D_1)|| \cdot \left|||\varphi(\tilde{D}_2)|| - ||\varphi(D_2)||\right|}{||\varphi(D_2)||^2} \\ & \quad \leq \frac{2\sqrt{2}}{||\varphi(D_2)||}\left(W_\infty(D_1, \emptyset) + \frac{1}{3}\right)^{\frac{1}{2}} W_2(D_1, \tilde{D}_1) \end{aligned}$$

$$+\frac{2\sqrt{2}\,\|\varphi(D_1)\|}{\|\varphi(D_2)\|^2}\left(W_\infty(D_2,\emptyset)+\frac{1}{3}\right)^{\frac{1}{2}}W_2(D_2,\tilde{D}_2).$$

A similar result is achieved for the second modulus by symmetry.

We employ the same idea for the third modulus, but it is more cumbersome. Again, we utilize the triangle inequality and (3.4), leading to

$$\begin{aligned}
&\left|\frac{\|\varphi(\tilde{D}_1)-\varphi(\tilde{D}_2)\|^2}{\|\varphi(\tilde{D}_1)\|\cdot\|\varphi(\tilde{D}_2)\|}-\frac{\|\varphi(D_1)-\varphi(D_2)\|^2}{\|\varphi(D_1)\|\cdot\|\varphi(D_2)\|}\right| \\
&\quad\le\frac{4}{\|\varphi(D_1)\|\cdot\|\varphi(D_2)\|}\cdot\left|\|\varphi(\tilde{D}_1)-\varphi(\tilde{D}_2)\|^2-\|\varphi(D_1)-\varphi(D_2)\|^2\right| \\
&\qquad+\frac{4\,\|\varphi(D_1)-\varphi(D_2)\|^2}{\|\varphi(D_1)\|\cdot\|\varphi(D_2)\|^2}\cdot\left|\|\varphi(\tilde{D}_2)\|-\|\varphi(D_2)\|\right| \\
&\qquad+\frac{4\,\|\varphi(D_1)-\varphi(D_2)\|^2}{\|\varphi(D_1)\|^2\cdot\|\varphi(D_2)\|^2}\cdot\|\varphi(\tilde{D}_1)\|\cdot\left|\|\varphi(\tilde{D}_1)\|-\|\varphi(D_1)\|\right|.
\end{aligned}$$

It remains to find a bound for each term in the RHS of the above inequality. The aim here is to eliminate the dependency on $\tilde{D}_1$ and $\tilde{D}_2$. So, it is sufficient to consider only the expressions containing these diagrams. In the second and third terms, the bounds for the relevant expressions are obtained easily from (3.3) and (3.4), which are

$$\left|\|\varphi(\tilde{D}_2)\|-\|\varphi(D_2)\|\right|\le\sqrt{2}\left(W_\infty(D_2,\emptyset)+\frac{1}{3}\right)^{\frac{1}{2}}W_2(D_2,\tilde{D}_2),$$

and

$$\|\varphi(\tilde{D}_1)\|\cdot\left|\|\varphi(\tilde{D}_1)\|-\|\varphi(D_1)\|\right|\le\frac{3\sqrt{2}\,\|\varphi(D_1)\|}{2}\left(W_\infty(D_1,\emptyset)+\frac{1}{3}\right)^{\frac{1}{2}}W_2(D_1,\tilde{D}_1).$$

For the first term, we consider two separate expressions below, and then apply triangle inequality, (3.3), and (3.4) as follows:

$$\begin{aligned}
&\|\varphi(\tilde{D}_1)-\varphi(\tilde{D}_2)\|+\|\varphi(D_1)-\varphi(D_2)\| \\
&\quad\le\|\varphi(\tilde{D}_1)\|+\|\varphi(\tilde{D}_2)\|+\|\varphi(D_1)-\varphi(D_2)\| \\
&\quad\le\frac{3\,\|\varphi(D_1)\|}{2}+\frac{3\,\|\varphi(D_2)\|}{2}+\|\varphi(D_1)-\varphi(D_2)\|,
\end{aligned}$$

and

$$\begin{aligned}
&\left|\|\varphi(\tilde{D}_1)-\varphi(\tilde{D}_2)\|-\|\varphi(D_1)-\varphi(D_2)\|\right| \\
&\quad\le\|\varphi(\tilde{D}_1)-\varphi(D_1)\|+\|\varphi(\tilde{D}_2)-\varphi(D_2)\| \\
&\quad\le\sqrt{2}\left(W_\infty(D_1,\emptyset)+\frac{1}{3}\right)^{\frac{1}{2}}W_2(D_1,\tilde{D}_1)+\sqrt{2}\left(W_\infty(D_2,\emptyset)+\frac{1}{3}\right)^{\frac{1}{2}}W_2(D_2,\tilde{D}_2).
\end{aligned}$$

Overall, we combine all relevant inequalities to calculate the constants c_1 and c_2. □

Although not required here, the previous proposition implies that ς is a continuous map. For this purpose, we equip the set $\mathcal{D}$ of persistence diagrams with 2-Wasserstein distance. We consider the product metric on $\mathcal{D}^2$. The continuity of ς is proved below.

Corollary 3.5. *The map $\varsigma : (\mathcal{D} \setminus \{\emptyset\})^2 \to \mathbb{R}$ is continuous.*

Proof. Let D_1 and D_2 be non-empty persistence diagrams. Consider the positive real η, c_1, and c_2 from Proposition 1. Fix a real $\epsilon > 0$. Set

$$\delta = \min\left\{\eta, \frac{\epsilon}{2c_1}, \frac{\epsilon}{2c_2}\right\}.$$

The inequality

$$\sqrt{W_2(D_1, \tilde{D}_1)^2 + W_2(D_2, \tilde{D}_2)^2} < \delta$$

implies that $W_2(D_1, \tilde{D}_1) \leq \eta$ and $W_2(D_2, \tilde{D}_2) \leq \eta$. So, the conclusion of (3.4) holds. Together with the assumption on δ, we obtain

$$|\varsigma(\tilde{D}_1, \tilde{D}_2) - \varsigma(D_1, D_2)| \leq c_1 \cdot W_2(D_1, \tilde{D}_1) + c_2 \cdot W_2(D_2, \tilde{D}_2) < \epsilon.$$

This shows the continuity of ς. □

For completeness, we introduce the cosine distance as the counterpart to cosine similarity. We will use this notion for demonstration in Section 5 as a direct comparison to other distances between persistence diagrams. We shall emphasize that the term itself is a misnomer since it is not a metric, but has been named this way in data science [18].

Definition 3.6. *For non-empty persistence diagrams D_1 and D_2, the* (landscape) cosine distance ς^* *is defined as*

$$\varsigma^*(D_1, D_2) = 1 - \varsigma(D_1, D_2).$$

Example 3.7. *(Examples 3.1 and 3.2 revisited) We can verify that*

$$\varsigma(D_1, D_2) = \frac{n}{\sqrt{n(n+1)}}.$$

This value approaches 1 as n increases, which indicates that D_1 and D_2 become almost equal. This agrees with our direct observation on both diagrams.

On the other hand, we calculate that $\varsigma(D_3, D_4) = 0$ for any m and n, which implies that D_3 and D_4 are perfectly dissimilar. This fits our previous explanation that their collections of points are far apart from one another, and thus both diagrams deserve to be considered entirely different.

4. Orthogonality between persistence diagrams

According to cosine similarity, two persistence diagrams are perfectly dissimilar if they are orthogonal. We shall justify why the notion of orthogonality is suitable to describe the perfect dissimilarity between persistence diagrams.

Definition 4.1. *Persistence diagrams D_1 and D_2 are said to be* orthogonal *if*

$$\langle \varphi(D_1), \varphi(D_2) \rangle = 0.$$

The above definition of orthogonality relies on the idea of persistence landscapes. We shall provide an equivalent definition which concentrates on persistence diagrams themselves. To avoid confusion below, the notation (b, d) refers to a point in a diagram, while $(b, d)_{\mathbb{R}}$ denotes an open interval of $\mathbb{R}$.

Proposition 4.2. *The persistence diagrams D_1 and D_2 are orthogonal if and only if for any pair of points $(b_1, d_1) \in D_1$ and $(b_2, d_2) \in D_2$, the intervals $(b_1, d_1)_{\mathbb{R}}$ and $(b_2, d_2)_{\mathbb{R}}$ have an empty intersection.*

Proof. For $i \in \{1, 2\}$, define

$$L_i = \bigcup_{(b,d)\in D_i} (b, d)_{\mathbb{R}}\,.$$

For $j \in \mathbb{N}$, denote $\varphi(D_i)_j$ as the j^{th} layer function of persistence landscape $\varphi(D_i)$, and also $\operatorname{supp}\left(\varphi(D_i)_j\right)$ as its support, i.e. the set of values $t \in \mathbb{R}$ where $\varphi(D_i)_j(t) > 0$. It is clear that

$$\operatorname{supp}\left(\varphi(D_i)_1\right) = L_i \quad \text{and} \quad \operatorname{supp}\left(\varphi(D_i)_j\right) \subseteq L_i$$

for $j > 1$.

Suppose that D_1 and D_2 are orthogonal. We have

$$\int \varphi(D_1)_1(t) \cdot \varphi(D_2)_1(t)\, dt \leq \langle \varphi\left(D_1\right), \varphi\left(D_2\right) \rangle = 0.$$

Each layer function above is continuous according to [19], and so is their product. Since the integral is zero, the function $\varphi(D_1)_1 \cdot \varphi(D_2)_2$ is zero everywhere. This shows that

$$L_1 \cap L_2 = \operatorname{supp}\left(\varphi(D_1)_1\right) \cdot \operatorname{supp}\left(\varphi(D_2)_1\right) = \operatorname{supp}\left(\varphi(D_1)_1 \cdot \varphi(D_2)_1\right) = \emptyset$$

as desired.

Conversely, suppose that $L_1 \cap L_2 = \emptyset$. Note that

$$\operatorname{supp}\left(\varphi(D_1)_j \cdot \varphi(D_2)_j\right) = \operatorname{supp}\left(\varphi(D_1)_j\right) \cdot \operatorname{supp}\left(\varphi(D_2)_j\right) \subseteq L_1 \cap L_2 = \emptyset$$

for any j. This implies that $\varphi(D_1)_j \cdot \varphi(D_2)_j$ is a zero function, and consequently $\langle \varphi(D_1), \varphi(D_2) \rangle = 0$. □

Corollary 4.3. *If the persistence diagrams D_1 and D_2 are orthogonal, then*

$$\|\varphi(D_1) - \varphi(D_2)\|_\infty = \sup\{\|\varphi(D_1)\|_\infty, \|\varphi(D_2)\|_\infty\},$$

and

$$\|\varphi(D_1) - \varphi(D_2)\|_p{}^p = \|\varphi(D_1)\|_p{}^p + \|\varphi(D_2)\|_p{}^p.$$

Proof. This is straightforward by definitions of the norms and also the implication of Proposition 4.2. □

The previous proposition gives an intuitive meaning of orthogonality for persistence diagrams from the view of their simplicial filtrations. Given two orthogonal diagrams, whenever a feature exists during a certain range of filtration parameter for a diagram, there is none for another diagram in the same range. Due to this behavior, their ground simplicial complexes can be seen as a clear contrast to each other, and thus the same can be said of the persistence diagrams as well. This is indeed a sound reason on why the orthogonality symbolizes a perfect dissimilarity between diagrams, though we have a second justification below.

Recall the trivial matching based on the definitions of bottleneck and Wasserstein distances in Section 2. In general, the bijection is considered as a bad pairing of points between two persistence diagrams. It disregards the pairs of points in these diagrams who are closer to each other compared to points on the diagonal line. It is commonly expected that trivial matching is not perfect. However, it is indeed a perfect matching if the diagrams are orthogonal.

Proposition 4.4. *If persistence diagrams D_1 and D_2 are orthogonal, then the trivial matching between them is perfect under the bottleneck and Wasserstein distances.*

Proof. This clearly holds if either D_1 and D_2 is empty. We assume that both diagrams are non-empty. For ease of notation, define

$$\omega_\infty(\gamma) = \sup_{x \in D_1^\Delta} \|x - \gamma(x)\|_\infty \quad \text{and} \quad \omega_p(\gamma) = \sum_{x \in D_1^\Delta} \|x - \gamma(x)\|_\infty{}^p$$

for a bijection $\gamma : D_1^\Delta \to D_2^\Delta$ in the definitions of bottleneck and Wasserstein distances. Denote κ as the trivial matching. For points $x = (b_1, d_1) \in D_1$ and $y = (b_2, d_2) \in D_2$, their paired points on the diagonal line Δ are

$$\kappa(x) = \left(\frac{b_1 + d_1}{2}, \frac{b_1 + d_1}{2}\right) \quad \text{and} \quad \kappa^{-1}(y) = \left(\frac{b_2 + d_2}{2}, \frac{b_2 + d_2}{2}\right).$$

Note that the points above are the closest to x and y respectively among points on Δ with respect to the supremum norm in $\mathbb{R}^2$.

Suppose that D_1 and D_2 are orthogonal. We claim that

$$\sup\{\|x - \kappa(x)\|_\infty, \|y - \kappa^{-1}(y)\|_\infty\} \le \|x - y\|_\infty, \tag{4.1}$$

and

$$\|x - \kappa(x)\|_\infty{}^p + \|y - \kappa^{-1}(y)\|_\infty{}^p \le \|x - y\|_\infty{}^p. \tag{4.2}$$

Proposition 4.2 states that $(b_1, d_1)_\mathbb{R} \cap (b_2, d_2)_\mathbb{R} = \emptyset$. This implies that either $d_1 \le b_2$ or $d_2 \le b_1$. Without loss of generality, we assume the former. The above inequalities are proved, respectively, as follows:

$$\begin{aligned} \sup\left\{\frac{d_1 - b_1}{2}, \frac{d_2 - b_2}{2}\right\} &< \sup\{d_1 - b_1, d_2 - b_2\} \\ &\le \sup\{b_2 - b_1, d_2 - d_1\}, \end{aligned}$$

and

$$2\left(\frac{d_1 - b_1}{2}\right)^p + 2\left(\frac{d_2 - b_2}{2}\right)^p \le (d_1 - b_1)^p + (d_2 - b_2)^p$$

$$\leq (b_2 - b_1)^p + (d_2 - d_1)^p$$
$$\leq 2 \cdot \sup\{(b_2 - b_1)^p, (d_2 - d_1)^p\}.$$

Consider a bijection $\gamma : D_1^\Delta \to D_2^\Delta$. For a point $x \in D_1$, note that if $\gamma(x) \in \Delta$, then

$$\|x - \kappa(x)\|_\infty \leq \|x - \gamma(x)\|_\infty,$$

since $\kappa(x)$ is the closest to x among points on Δ. If $\gamma(x) = y \in D_2$, then inequalities (4.1) and (4.2) hold. These findings conclude that

$$\omega_\infty(\kappa) \leq \omega_\infty(\gamma) \quad \text{and} \quad \omega_p(\kappa) \leq \omega_p(\gamma)$$

for any bijection γ, and thus κ is a perfect matching. □

Corollary 4.5. *If the persistence diagrams D_1 and D_2 are orthogonal, then*

$$W_\infty(D_1, D_2) = \sup\{W_\infty(D_1, \emptyset), W_\infty(D_2, \emptyset)\}$$

and

$$W_p(D_1, D_2)^p = W_p(D_1, \emptyset)^p + W_p(D_2, \emptyset)^p.$$

Proof. This is straightforward by definitions of the distances and also the implication of Proposition 4.4. □

The last proposition provides an intuitive meaning for orthogonality of persistence diagrams with respect to the distributions of their points on the same plot. The points in a diagram are sufficiently far apart from those in another one to avoid better pairing than the trivial matching. It can be thought that the diagrams occupy separate regions on the plot, and they deserve to be considered entirely different. This provides another justification on why orthogonality signifies a perfect dissimilarity between persistence diagrams.

As a remark, although two orthogonal diagrams are deemed entirely different based on both justifications above, it disagrees with the interpretation of distances. Recall that a distance indicates degree of difference. So, it is expected that two orthogonal diagrams have a significantly large distance. However, they can record a very small distance, as illustrated in Example 3.2.

5. Data demonstration

We shall demonstrate the accuracy of cosine similarity against the standard distances to indicate similarity between persistence diagrams. Here, the cosine distance is used instead of cosine similarity to match the narrative with other distances. We will show that the cosine distance is able to distinguish clearly whether two diagrams are closely similar or dissimilar. This demonstration is confined to simple shapes in $\mathbb{R}^2$, but it is satisfactory to provide insight to our claim.

5.1. Data preparation

To prepare persistence diagrams in this demonstration, several data sets are sampled from certain regions or shapes in $\mathbb{R}^2$. The sets are listed as follows:

1) Q and Q' from the unit disc $\mathcal{Q} = \left\{x \in \mathbb{R}^2 \mid \|x\|_2 \leq 1\right\}$,
2) R and R' from an annulus $\mathcal{R} = \left\{x \in \mathbb{R}^2 \mid 0.5 \leq \|x\|_2 \leq 1\right\}$,
3) S and S' from the unit circle $\mathcal{S} = \left\{x \in \mathbb{R}^2 \mid \|x\|_2 = 1\right\}$.

Each data set contains 3000 points which are chosen randomly and uniformly from the designated shape. Persistence diagrams for the sets are obtained via Vietoris-Rips filtration. We observe features in the zeroth and first homology groups for these diagrams, which are denoted as H_0 and H_1, respectively. The diagrams are then compared against each other via the bottleneck distance W_∞, Wasserstein distance W_2, distances defined by the norms $\|\cdot\|_\infty$ and $\|\cdot\|_2$ for persistence landscapes, and, lastly, cosine distance ς^*. The persistent homology method is carried out via a Python package called giotto-tda [30].

For each data set, its points are selected in the above manner so that the filtration is able to capture the topological structure of the shape accurately. Given a parameter value, its Vietoris-Rips complex can be regarded as a reconstruction of the said shape from the data set. Specifically, each 1-simplex in this complex creates an edge between two points, while each 2-simplex forms a triangular region enclosed by three points in the data set. The union of these edges and triangles approximates the original shape. It consists of connected components and holes which are specified by persistent diagrams with respect to H_0 and H_1. Any geometrical feature in each shape, such as holes of $\mathcal{R}$ and $\mathcal{S}$, will be presented in the diagrams of corresponding data sets.

5.2. Objectives

In this demonstration, we shall observe the following:

1) Distances between persistence diagrams with respect to H_0 and H_1 for three pairs Q and Q', R and R', and S and S'.
2) Distances between persistence diagrams with respect to H_1 among Q, R, and S.

Specifically, we examine whether each distance can convey the similarity between those diagrams clearly. The objectives are decided this way because, in theory, we have prior knowledge on whether those diagrams are closely similar or very dissimilar. So, we can compare this theoretical information with the results obtained in this demonstration with confidence.

For the first objective, two data sets are set up for each shape to demonstrate close similarity between their persistence diagrams. Since both are sampled from the same shape randomly and uniformly, their diagrams must have a slight difference only. So, their distances shall be small, including the cosine distance.

The second objective is done to verify extent of similarity and dissimilarity between persistence diagrams with respect to H_1 for the relevant data sets. Recall that features in H_1 relate to holes formed during a simplicial filtration. Each of the shapes $\mathcal{R}$ and $\mathcal{S}$ has a hole, unlike $\mathcal{Q}$. This hole will be recorded as a persistent feature in the diagrams with respect to H_1 for data sets R and S. Such feature will be absent in the diagram for Q. Thus, the diagram with respect to H_1 for R shall be similar to that for S, but dissimilar to that for Q. The cosine distance is expected to address this similarity or dissimilarity clearly in comparison to other distances.

5.3. *Results*

Figures 4–6 show the persistence diagrams for the data sets. The distances between relevant diagrams are summarized in Tables 1–6.

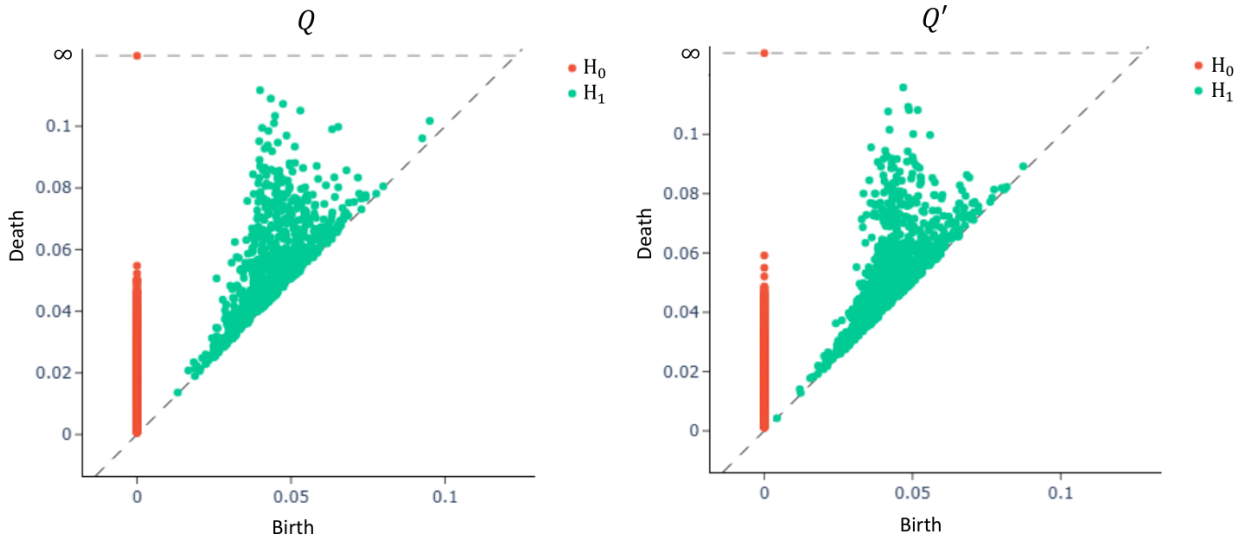


Figure 4. Persistence diagrams for data sets Q and Q'.

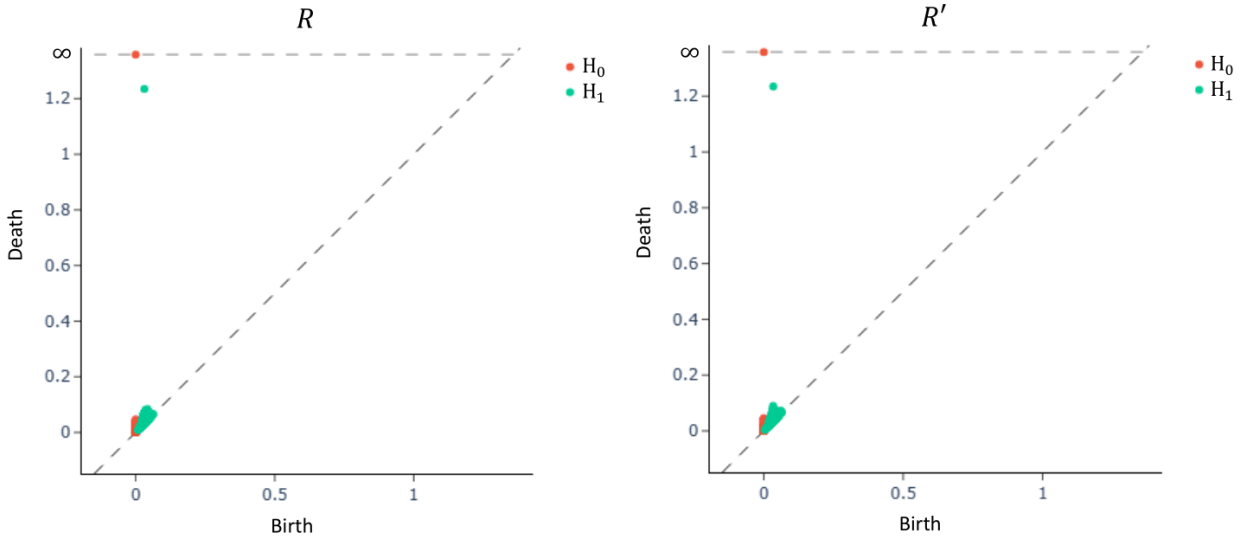


Figure 5. Persistence diagrams for data sets R and R'.

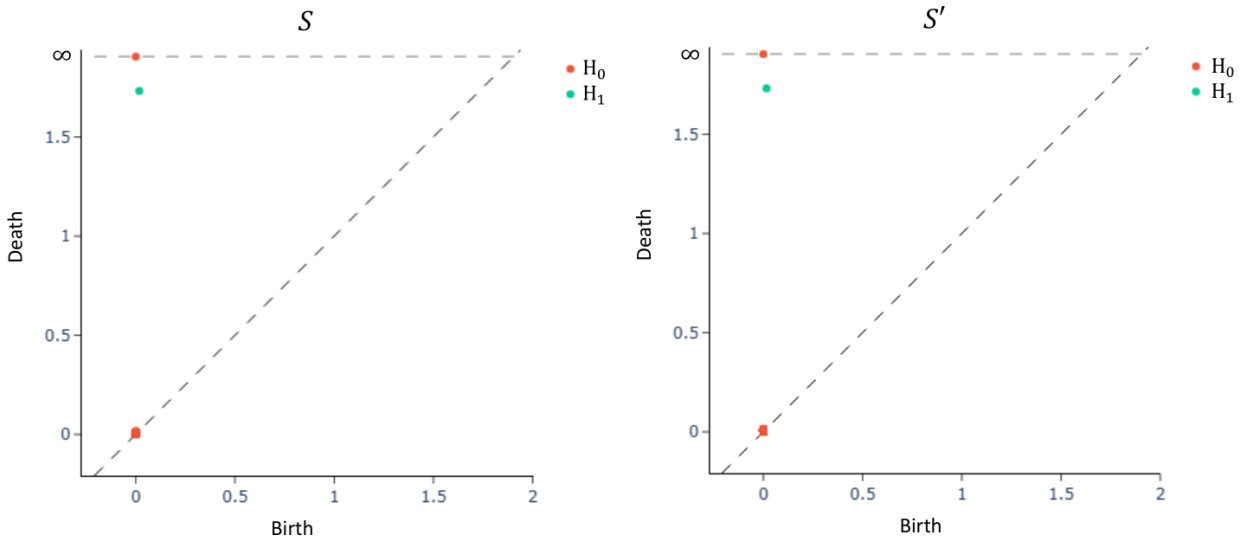


Figure 6. Persistence diagrams for data sets S and S'.

Table 1. Distances between persistence diagrams for two data sets from the same shape.

H_0

Data sets	W_∞	W_2	$\|\cdot\|_\infty$	$\|\cdot\|_2$	ς^*
Q and Q'	0.004360	0.015458	0.004360	0.000722	0.010417
R and R'	0.001629	0.013998	0.001629	0.000245	0.002320
S and S'	0.001921	0.006393	0.001921	0.000165	0.020274

H_1

Data sets	W_∞	W_2	$\|\cdot\|_\infty$	$\|\cdot\|_2$	ς^*
Q and Q'	0.010991	0.041436	0.005256	0.001085	0.014863
R and R'	0.008980	0.034152	0.008365	0.002123	0.000009
S and S'	0.002066	0.002066	0.002066	0.001911	0.000003

Table 2. Bottleneck distance between persistence diagrams for the data sets.

H_0

	Q	R	S
Q	0	0.013372	0.027399
R	0.013372	0	0.023379
S	0.027399	0.023379	0

H_1

	Q	R	S
Q	0	0.601816	0.856912
R	0.601816	0	0.497006
S	0.856912	0.497006	0

Table 3. 2-Wasserstein distance between persistence diagrams for the data sets.

	H_0				H_1		
	Q	R	S		Q	R	S
Q	0	0.352694	0.632645	Q	0	0.635436	0.887606
R	0.352694	0	0.452928	R	0.635436	0	0.518727
S	0.632645	0.452928	0	S	0.887606	0.518727	0

Table 4. Distance defined by supremum norm between persistence landscapes for the data sets.

	H_0				H_1		
	Q	R	S		Q	R	S
Q	0	0.013372	0.027394	Q	0	0.601763	0.856864
R	0.013372	0	0.023376	R	0.601763	0	0.497006
S	0.027394	0.023376	0	S	0.856864	0.497006	0

Table 5. Distance defined by 2-norm between persistence landscapes for the data sets.

	H_0				H_1		
	Q	R	S		Q	R	S
Q	0	0.001230	0.003578	Q	0	0.381141	0.647616
R	0.001230	0	0.002756	R	0.381141	0	0.387817
S	0.003578	0.002756	0	S	0.647616	0.387817	0

Table 6. Cosine distance between persistence diagrams for the data sets.

	H_0				H_1		
	Q	R	S		Q	R	S
Q	0	0.041416	0.725444	Q	0	0.973207	0.979322
R	0.041416	0	0.651488	R	0.973207	0	0.160770
S	0.725444	0.651488	0	S	0.979322	0.160770	0

5.4. Discussion

In Figures 4–6, each persistence diagram with respect to H_0 has exactly one feature that never dies, as explained in 2. It is omitted in the calculations for distances. Notice that there is a persistent feature in the diagram for R with respect to H_1, which reflects the hole of annulus $\mathcal{R}$. Such feature is also observed for R', S, and S' due to a similar reason.

The persistence diagrams for Q and Q' in Figure 4 seem to be closely similar. This is indeed confirmed by all distances in Table 1, where they are noticeably small. This is expected since the data sets originate from the same shape $\mathcal{Q}$. This is true for other pairs of data sets as well. From this finding, the cosine distance is able to indicate close similarity between two persistence diagrams by recording

a small value. However, this is nothing special since other distances conclude the same too. It is worth noting that the performance of cosine distance against other distances varies throughout the pairs of data sets. For example, in the pair S and S', the cosine distance records a higher value than other distances for H_0, but it is the opposite for H_1. Nonetheless, the values given by the cosine distance are small enough to ascertain the close similarity between diagrams for each pair of data sets.

Next, we examine the persistence diagrams with respect to H_1 for the data sets Q, R, and S. Note that the persistent feature for R and S is located way above the diagonal line in Figures 5 and 6. Meanwhile, the features in the diagram of Q accumulate around the diagonal line in Figure 4. This contributes a large difference between the diagram of each R and S and that of Q. Thus, we establish that the diagram of R is similar to that of S, while dissimilar to that of Q. This fact is confirmed by the cosine distance in Table 6. It is small for R and S to indicate their similarity, while it is close to 1 for Q and R to ascertain their dissimilarity.

However, the same fact is not conveyed clearly by the other distances. Each of their values for R and S is noticeably close to that for Q and R. This is especially apparent for the distance defined by $\|\cdot\|_2$ in Table 5. Hence, in this case, these distances cannot distinguish similarity or dissimilarity between persistence diagrams.

It is interesting to note that each distance records a relatively high value for Q and S, which implies a large difference between their diagrams with respect to H_1. This is expected since their diagrams are dissimilar. However, this fact is clearly indicated by the cosine distance whose value is close to 1, and even around the value for Q and R. Unlike other distances, it is more confident to distinguish the similarity or dissimilarity between the diagrams for those three data sets by using the cosine distance.

As a conclusion, our demonstration shows that the cosine distance, and thus cosine similarity, are more accurate to indicate similarity between the diagrams in comparison to other distances. We shall clarify that the demonstration has been replicated multiple times with different samples from those three shapes, and our conclusion on the cosine similarity still holds in each run.

6. Conclusions

In this paper, we have introduced the cosine similarity as a new indicator for similarity between persistence diagrams. It measures the extent of similarity within two extremes, which are perfect similarity and dissimilarity. While the former simply means equality of two diagrams, the perfect dissimilarity refers to the notion of orthogonality between the diagrams. The cosine similarity is found to be more accurate in this purpose, unlike the common distances for persistence diagrams, such as the bottleneck and Wasserstein distances.

For future work, we shall provide some ideas regarding the research problem on similarity between persistence diagrams. The demonstration in Section 5 is limited to simple geometric shapes in $\mathbb{R}^2$. We are yet to test the cosine similarity in potentially critical cases, such as data sets of higher dimension, from non-manifold structures such as fractals, or in the presence of noise. Further works are needed to generalize the findings here on the accuracy of cosine similarity to cover such cases.

As mentioned in Section 1, there are other approaches in the literature to measure the similarity. We are yet to compare them with our cosine similarity in terms of accuracy and efficiency. For example, the RST parametric model [15] is able to account for geometrical and topological differences on the shapes of data sets. This cannot be achieved by the cosine similarity since data sets from two shapes of

different sizes, but topologically equal are perceived to be different under the cosine similarity. More research is needed to investigate the advantages and disadvantages of each method, including the cosine similarity itself.

It is possible to propose a variation of the cosine similarity. For example, we define another variation as

$$\varrho(D_1, D_2) = \frac{2\langle\varphi(D_1), \varphi(D_2)\rangle}{\|\varphi(D_1)\|^2 + \|\varphi(D_2)\|^2}$$

for persistence diagrams D_1 and D_2, where one of them is non-empty. This is motivated by (3.2). In fact, this equation enables the corresponding cosine distance to be expressed nicely as

$$\varrho^*(D_1, D_2) = \frac{\|\varphi(D_1) - \varphi(D_2)\|^2}{\|\varphi(D_1)\|^2 + \|\varphi(D_2)\|^2}.$$

All results in this paper still hold for this variation. It is worth noting that ϱ is strictly smaller than ς except in case of perfect similarity or dissimilarity. This is proved easily by applying the AM-GM inequality [29] to Cauchy-Schwarz inequality in (3.1).

Instead of persistence landscapes, the cosine similarity can be defined by employing a different vectorization of persistence diagrams, provided that its images are in an inner product space. There are some vectorizations in the literature that have potential for this purpose, such as the Betti curve [9], persistence silhouette [31], life entropy curve [32], heat kernel [33], persistence image [34], persistence block [35], PD thresholding curve [36], and other persistence curves [37]. However, the perfect similarity and dissimilarity may have different meanings depending on the vectorization.

Author contributions

Azmeer Nordin: conceptualization, data curation, formal analysis, software, validation, visualization, writing-original draft; Mohd Salmi Md Noorani: conceptualization, supervision, validation, writing-review and editing; Nurulkamal Masseran: funding acquisition, project administration, supervision, writing-review and editing; Mohd Sabri Ismail: conceptualization, software, validation, writing-review and editing; Nur Firyal Roslan: writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the research grant DIP-2024-004 provided by Universiti Kebangsaan Malaysia. The authors are thankful to the referees for constructive comments and suggestions which help to improve the quality of this paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington, A roadmap for the computation of persistent homology, *EPJ Data Sci.*, **6** (2017), 17. https://doi.org/10.1140/epjds/s13688-017-0109-5
2. L. Wasserman, Topological data analysis, *Annu. Rev. Stat. Appl.*, **5** (2018), 501–532. https://doi.org/10.1146/annurev-statistics-031017-100045
3. F. Chazal, B. Michel, An introduction to topological data analysis: fundamental and practical aspects for data scientists, *Front. Artif. Intell.*, **4** (2021), 667963. https://doi.org/10.3389/frai.2021.667963
4. F. Hensel, M. Moor, B. Rieck, A survey of topological machine learning methods, *Front. Artif. Intell.*, **4** (2021), 681108. https://doi.org/10.3389/frai.2021.681108
5. Y. Skaf, R. Laubenbacher, Topological data analysis in biomedicine: A review, *J. Biomed. Inform.*, **130** (2022), 104082. https://doi.org/10.1016/j.jbi.2022.104082
6. M. Gidea, Y. Katz, Topological data analysis of financial time series: Landscapes of crashes, *Physica A*, **491** (2018), 820–834. https://doi.org/10.1016/j.physa.2017.09.028
7. S. Arvanitis, M. Detsis, Mild explocivity, persistent homology and cryptocurrencies' bubbles: An empirical exercise, *AIMS Mathematics*, **9** (2024), 896–917. https://doi.org/10.3934/math.2024045
8. C. Y. Li, M. Ovsjanikov, F. Chazal, Persistence-based structural recognition, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, 2014, 2003–2010. https://doi.org/10.1109/CVPR.2014.257
9. H. Edelsbrunner, J. L. Harer, *Computational topology: An introduction*, Providence: American Mathematical Society, 2010. https://doi.org/10.1090/mbk/069
10. N. F. S. Zulkepli, M. S. M. Noorani, F. A. Razak, M. Ismail, M. A. Alias, Haze detection using persistent homology, *AIP Conf. Proc.*, **2111** (2019), 020012. https://doi.org/10.1063/1.5111219
11. N. F. S. Zulkepli, M. S. M. Noorani, F. A. Razak, M. Ismail, M. A. Alias, Topological characterization of haze episodes using persistent homology, *Aerosol Air Qual. Res.*, **19** (2019), 1614–1624. https://doi.org/10.4209/aaqr.2018.08.0315
12. T. K. Dey, Y. S. Wang, *Computational topology for data analysis*, Cambridge: Cambridge University Press, 2022. https://doi.org/10.1017/9781009099950
13. H. Bando, S. Kaji, T. Yaguchi, Causal inference for empirical dynamical systems based on persistent homology, *JSIAM Lett.*, **14** (2022), 69–72. https://doi.org/10.14495/jsiaml.14.69
14. N. F. S. Zulkepli, M. S. M. Noorani, F. A. Razak, M. Ismail, M. A. Alias, Hybridization of hierarchical clustering with persistent homology in assessing haze episodes between air quality monitoring stations, *J. Environ. Manage.*, **306** (2022), 114434. https://doi.org/10.1016/j.jenvman.2022.114434
15. S. Agami, Comparison of persistence diagrams, *Commun. Stat.-Simul. C.*, **52** (2023), 1948–1961. https://doi.org/10.1080/03610918.2021.1894335

16. B. Di Fabio, M. Ferri, Comparing persistence diagrams through complex vectors, In: *Image analysis and processing-ICIAP 2015*, Cham: Springer, 2015, 294–305. https://doi.org/10.1007/978-3-319-23231-7_27

17. J. P. Wang, Y. H. Dong, Measurement of text similarity: A survey, *Information*, **11** (2020), 421. https://doi.org/10.3390/info11090421

18. P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to data mining*, 2 Eds., London: Pearson, 2019.

19. P. Bubenik, Statistical topological data analysis using persistence landscapes, *J. Mach. Learn. Res.*, **16** (2015), 77–102.

20. P. Bubenik, The persistence landscape and some of its properties, In: *Topological data analysis*, Cham: Springer, 2020, 97–117. https://doi.org/10.1007/978-3-030-43408-3_4

21. S. M. S. S. Musa, M. S. Md Noorani, F. A. Razak, M. Ismail, M. A. Alias, S. I. Hussain, An early warning system for flood detection using critical slowing down, *Int. J. Environ. Res. Public Health*, **17** (2020), 6131. https://doi.org/10.3390/ijerph17176131

22. S. M. S. S. Musa, M. S. Md Noorani, F. A. Razak, M. Ismail, M. A. Alias, S. I. Hussain, Using persistent homology as preprocessing of early warning signals for critical transition in flood, *Sci. Rep.*, **11** (2021), 7234. https://doi.org/10.1038/s41598-021-86739-5

23. M. S. Ismail, M. S. Md Noorani, M. Ismail, F. A. Razak, M. A. Alias, Early warning signals of financial crises using persistent homology, *Physica A*, **586** (2022), 126459. https://doi.org/10.1016/j.physa.2021.126459

24. M. S. Ismail, M. S. Md Noorani, M. Ismail, F. A. Razak, Early warning signals of financial crises using persistent homology and critical slowing down: evidence from different correlation tests, *Front. Appl. Math. Stat.*, **8** (2022), 940133. https://doi.org/10.3389/fams.2022.940133

25. M. E. Aktas, E. Akbas, A. El Fatmaoui, Persistence homology of networks: methods and applications, *Appl. Netw. Sci.*, **4** (2019), 61. https://doi.org/10.1007/s41109-019-0179-3

26. N. Ravishanker, R. J. Chen, An introduction to persistent homology for time series, *WIRES Comput. Stat.*, **13** (2021), e1548. https://doi.org/10.1002/wics.1548

27. A. Hatcher, *Algebraic topology*, Cambridge: Cambridge University Press, 2002.

28. L. Betthauser, P. Bubenik, P. B. Edwards, Graded persistence diagrams and persistence landscapes, *Discrete Comput. Geom.*, **67** (2022), 203–230. https://doi.org/10.1007/s00454-021-00316-1

29. E. F. Beckenbach, R. Bellman, *Inequalities*, Berlin: Springer, 1961. https://doi.org/10.1007/978-3-642-64971-4

30. G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. Medina-Mardones, A. Dassatti, K. Hess, giotto-tda: A topological data analysis toolkit for machine learning and data exploration, *J. Mach. Learn. Res.*, **22** (2021), 1–6.

31. F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, Stochastic convergence of persistence landscapes and silhouettes, In: *Proceedings of the thirtieth annual symposium on computational geometry*, New York: Association for Computing Machinery, 2014, 474–483. https://doi.org/10.1145/2582112.2582128

32. N. Atienza, R. Gonzalez-Díaz, M. Soriano-Trigueros, On the stability of persistent entropy and new summary functions for topological data analysis, *Pattern Recogn.*, **107** (2020), 107509. https://doi.org/10.1016/j.patcog.2020.107509

33. J. Reininghaus, S. Huber, U. Bauer, R. Kwitt, A stable multi-scale kernel for topological machine learning, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, 2015, 4741–4748. https://doi.org/10.1109/CVPR.2015.7299106

34. H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, et al., Persistence images: A stable vector representation of persistent homology, *J. Mach. Learn. Res.*, **18** (2017), 1–35.

35. K. C. Chan, U. Islambekov, A. Luchinsky, R. Sanders, A computationally efficient framework for vector representation of persistence diagrams, *J. Mach. Learn. Res.*, **23** (2022), 1–33.

36. Y.-M. Chung, S. Day, Topological fidelity and image thresholding: A persistent homology approach, *J. Math. Imaging Vis.*, **60** (2018), 1167–1179. https://doi.org/10.1007/s10851-018-0802-4

37. Y.-M. Chung, A. Lawson, Persistence curves: A canonical framework for summarizing persistence diagrams, *Adv. Comput. Math.*, **48** (2022), 6. https://doi.org/10.1007/s10444-021-09893-4

AIMS Press