1 Outline

1. Intro: The introduction is currently too specific to your method. It should introduce the high-level goals, and why we have those high-level goals. Other methods for defining representations of homology group generators should be introduced here as well, along with an intuitive explanation for why the proposed approach is needed/is better/addresses something new. Since most of the jargon will not yet be defined, the intro should be clear and intuitive.

2. Background: This is where you can get into more details about persistent homology. Notation and jargon can be defined here, along with more details about other methods.

3. Methodology: This is where you explain your method. First define the method, perhaps providing an algorithm. Include a simple illustration to help the reader. The bounds you worked out can also be included here, perhaps as a subsection.

4. Simulation study: This is where you include results of a simulation study. First explain the simulation study design, include tables or graphics of the results, and then analyze the performance. There will need to be a comparison to a competing method.

5. Application: This is where we can include the cosmology example. We can work out the details when we get to it.

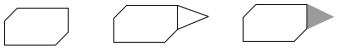
6. Concluding remarks: Remind the reader the main points of your work, any relevant discussion on the method or results, and ideas for improvement or future work.

2 Introduction

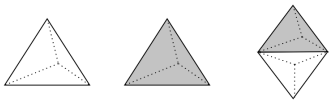
Given a point cloud dataset, it might be useful to investigate its underlying topological structure, e.g. whether the data points form a loop or a void in the space (usually denoted as X.) One approach is via "homology" defined in algebraic topology, as it naturally captures such information. For real life applications, we are generally interested in lower dimensional (0, 1, and 2 dimensions) homological groups because they have intuitive meanings: $H_0(X)$ counts the number of path-connected sets of X, $H_1(X)$ counts the number of holes in X, and $H_2(X)$ counts the number of voids in X. fig. 1 below illustrates what homology groups encode. One advantage of using homology is that it summarizes a given dataset's basic topological features in just a few numbers. An immediate application might be to use those numbers as extra information for the input layer of the neural networks.

Some applications of persistent homology includes network analysis, text mining, astrophysics data analysis, neuroscience, etc. In network data, clique

(a) $H_0 = 1, 2, 1$ resp.



(b) $H_1 = 1, 2, 1$ resp.



(c) $H_2 = 1, 0, 1$ resp.

Figure 1: Some examples of low dimensional homological groups

community are recognized as an important concept to describe cohesive structures of the network, and in [7], the authors proposed to use persistent homology to detect such cliques and their evolution. In [10], the authors developed the so-called Similarity Filtration with Time Skeleton (SIFTS) algorithm that uses H_1 generators to identify the semantic "tie-backs" in text documents. For demonstrations in neuroscience and related fields, one can refer to [8], where the datasets on mouse connectome, electrical and chemical synapses in C. elegans, and genomic interaction data, are analyzed using persistent homology. In addition, the study on small scale subhalo distributions in cold dark matter and warm dark matter models using persistent homology is proposed in [2].

Current methods of finding H_1 loop representations in a given point cloud is not stable, i.e., perturbations of the points and changes in the filtration sequence (defined later) can both affect the final results. Also, the loop representation is not unique by nature (discussed briefly in the next section). Therefore, the question of finding good representatives remains open. One approach is to find a "minimal" representation for a loop, i.e., use as few 1-simplexes as possible to capture a loop feature. For example, a triangle is considered better than a rectangle to capture a hole. Another possible direction is to "thicken" the original loop representative to hopefully capture the underlying manifold better (it is usually the case that the underlying loop manifold is in a band shape rather than thin connected line segments.) In this paper, we propose to use the thickened loop as the loop feature representative. We also check the stability of this representation in some sense defined later.

3 Background

Some key concepts of persistent homology are briefly reviewed in this section. A simplex is a triangle and its extension to other dimensions. For example, a zero-simplex is a point, a one-simplex is a segment, a two-simplex is a solid triangle, and a three-simplex is a solid tetrahedron. An n-simplex σ is the

convex hull of $n+1$ affinely independent points, $a_{0:n} = \{a_0, a_1, \ldots, a_n\}$. The *faces* of σ are the k-simplexes formed from a subset of size $k+1$ of $a_{0:n}$, where $k = 0, \ldots, n$ For example, the faces of a triangle are the three vertices (zero-simplexes), three segments (one-simplexes), and one triangle (two-simplex). A simplicial complex, K, is a collection of simplexes, satisfying two conditions: (i) each face of a simplex σ is also in K, and (ii) non-empty intersection of two simplexes σ_1 and σ_2 is a face of both σ_1 and σ_2. A *filtered simplicial complex* is a sequence of complexes, where each complex contains the previous one (see fig. 2 for an example.)

Figure 2: A filtration where we grow the complex by adding a single simplex at each step

There are different ways to define a filtered simplicial complex, and one of them is the Vietoris-Rips complex, which is defined as follows. Two points are connected if the distance between them is less than a given value say δ; and we add a triangle simplex if all three edges have length less than δ; similarly, a tetrahedron is added if all of its edges have length less than δ, etc. Mathematically, the VR complex is defined as

$$\mathrm{VR}(S, \delta) = \left\{ \sigma \;\middle|\; \mathrm{diam}(\sigma) \leq \delta \right\} \tag{1}$$

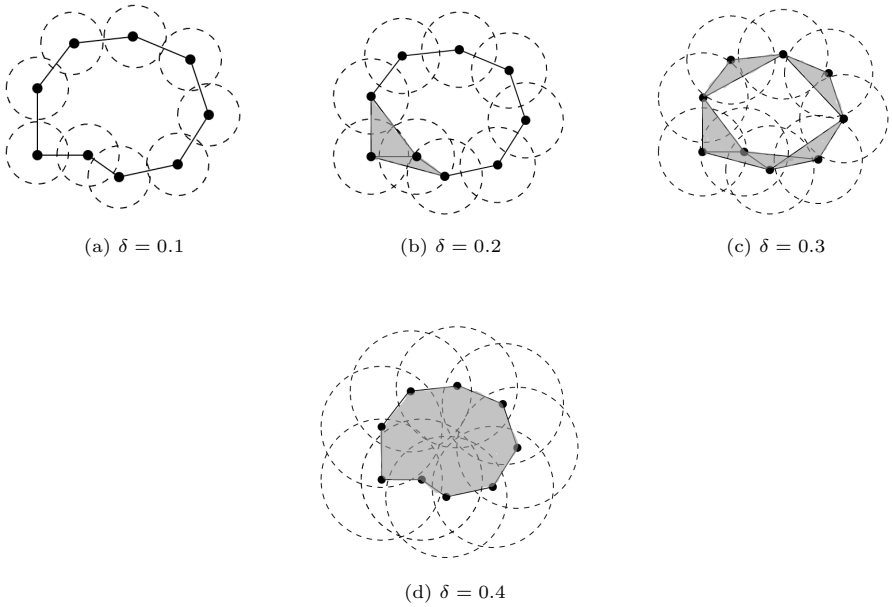
where set S is the original dataset's points, and "diam" is the diameter of a simplex σ. To build a filtration with VR complex, we just need to slowly increase δ. In this paper, a VR filtration is used in the later experiments. Now we are ready to introduce the basic idea of persistent homology, a central concept in TDA. For each hole (or other dimensional features) we find, it is useful to record its "persistence", i.e., how stable this feature is, as we move along the filtration. We use VR-filtration (see fig. 3) as a visual aid to explore this concept.

When $\delta = 0$, we only have the individual points. Suppose we start with $\delta = 0.1$, then a central hole appears. When $\delta = 0.2$, couple more 1-simplexes are added, but the hole persists. The hole keeps shrinking as δ increases. At $\delta = 0.4$, the hole disappears for the first time. In literature, a tuple (b, d) is often used to record such persistence, where b means the "birth time" of this feature in the filtration, and d means the "death time" of this feature. So here, $(b, d) = (0.1, 0.4)$. For a fuller example, suppose fig. 3 is just a snapshot from a bigger point cloud, and we have other loop features discovered somewhere else. fig. 4 is the corresponding persistence diagram.

Next, we put the above paragraph in mathematical terms. In VR-filtration, a monotone increasing sequence $0 < \delta_1 < \delta_2 < \ldots < \delta_k < \infty$ leads to a filtration

$$\mathrm{VR}(S, \delta_1) \subset \mathrm{VR}(S, \delta_2) \subset \ldots \subset \mathrm{VR}(S, \delta_k)$$

A sequence of p-dimensional homology group corresponding to the above filtra-



(a) $\delta = 0.1$ (b) $\delta = 0.2$ (c) $\delta = 0.3$

(d) $\delta = 0.4$

Figure 3: VR filtration; in the last figure, internal line segments are omitted for diagram clarity

tion is then computed,

$$H_p(\mathrm{VR}(S, \delta_1)) \to H_p(\mathrm{VR}(S, \delta_2)) \to \ldots \to H_p(\mathrm{VR}(S, \delta_k))$$

Each homology group can be thought of as a vector space; thus it has a basis, which intuitively speaking are our "loop features". Finally, each feature appears and disappears at certain timestamps along the filtration, and we denote such information as (b, d).

For a more detailed introductory treatment to TDA and persistent homology, refer to [6]. Next comes the question of how to represent each topological feature: we now know the birth and death time of a feature, but we also want to know where it is and how it roughly looks. By definition, homological group H_q is defined as the quotient group of cycles and boundaries, and because the representative of a quotient group element is not unique (refer to any textbook on abstract algebra for the concept of *quotient group*), we do not have a unique representation. To be a bit more intuitive, if two representatives of a hole only defer by the boundary of some triangles, we say they are equivalent. For example, in fig. 5, the blue and orange loops for the central hole are equivalent.

People have come up with different ways to find "good" representatives in some sense. The simplest approach may be: when a loop forms the first time in the filtration, we find an arbitrary head-to-tail connected line-segments that captures this loop and call it the representative no matter how this loop feature may evolve down in the filtration. Another way is to use whatever the matrix

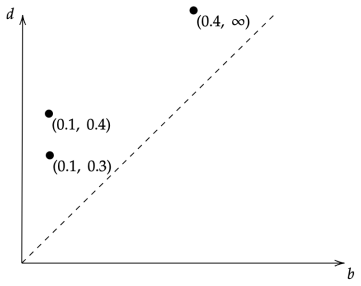


Figure 4: persistent diagram; the hole we discovered above is represented as point $(0.1, 0.4)$, and suppose there is a feature existing even after δ reaches its maximum value, then we simply set $d = \infty$.

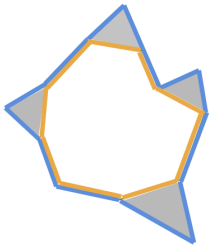


Figure 5: Non-unique representative loops

reduction algorithm (among the first TDA algorithms proposed) reports. The benefit of this approach is that it naturally reports both (b, d) and the representative at the same time, and some libraries incorporate this algorithm. But sometimes we want more control over how to select a nice loop representative, and here we highlight two such approaches.

One method is to find the "minimal" representation [3], i.e., use as few line segments as possible to capture the hole. The algorithm goes roughly as follows. We first build a filtration sequence

$$K_1 \subset K_2 \subset \cdots \subset K_n$$

where K_i is the i-th complex in the filtration and differs from K_{i-1} by only a single simplex σ_i. When a loop first appears in the filtration with the addition of σ_i, let $c_i \subset K_i$ be the shortest cycle, in term of the number of line segments, containing σ_i. Now, suppose we want to find a representation for a loop with a particular (b, d) tuple. We first collect all loop features $\{c_i\}$ born before or at timestamp b that also persist at least until timestamp d. Then, at timestamp d, we find among those c_i's a combination that vanishes. We output such combination as our representation for the loop feature with (b, d). The proof for correctness is fairly technical, and we refer to the paper itself [3].

Couple remarks are in place for Dey et al.'s algorithm. Although each c_i is of fewest line segments, their combinations may not be. But, since each c_i is minimal, we would expect their combination to be still relatively simple. Also, the method is not stable, i.e., the representative loop will change along with the perturbation of sampled points and the order of filtration sequence. The authors also noted these two problems, and provided an example for the second remark here.

Next, Xu and Kehe [9] proposed another method where the designated representative contains an empty region. For example, in fig. 5, the orange loop

may be preferred over the blue one. The main idea is to apply the concept of α-shape developed in [4]. Start with any algorithm that outputs a loop feature representative, which would then be improved by this algorithm (call this initial set of representative points S_X.) Construct the convex hull containing S_X, and collect all points inside this convex hull as set S'_X. Then construct the α-hull and its corresponding α-shape. Name the centers of those green circles making up the α-hull as C_α. Next, pick the centers among C_α that are inside the outer part of the α-shape, and name the set C'_α. Finally, collect the data points touching the circles centered at C'_α, and call them the improved representative points. And now if we connect the improved representative points by line segments, they do not encircle extra data points, achieving the goal of encircling "an empty region". A drawback of this approach is that it only works when the dimension of the topological feature we look for agrees with the dimension of the underlying space, e.g., find a loop in 2D space, or a void in 3D space (but not for finding loops in 3D space).

In the following sections, we turn to the method we propose in this paper, i.e., thicken the line-segment loop. There is a R-package called "TDA" [5] that takes the original data points as input, and returns a line-segment loop for each loop feature discovered. Our idea is quite straight-forward: we start with this R-package output, and then thicken it by some radius of choice (discussed later). There is no need to walk through the algorithm; rather, a main part of this paper is to investigate how stable this thickened loop is. To this end, we first go through some notations.

4 Some notations

One way to measure the difference between two sets in a metric space is by Hausdorff distance, d_{H}, defined as

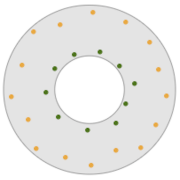
$$d_{\mathrm{H}}(X,Y) = \max\left(\sup_{x\in X} d(x,Y), \sup_{y\in Y} d(y,X)\right) \tag{2}$$

with

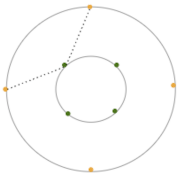
$$d(x,Y) = \inf_{y\in Y} d(x,y) \tag{3}$$

where $d(x,y)$ is the space's endowed metric; similarly define for $d(y,X)$. Intuitively, it measures the extreme distance between the two sets X and Y. Suppose now we obtain two sets of representative points of a given H_1 feature. A natural question is: how different are these two representations? One could directly use Hausdorff distance to answer that question. We provide an easy example in fig. 6, where the shaded region is the underlying loop manifold. Suppose in one sampling instance, we get the green dots; and in another sampling instance, we get the orange dots. Naturally, if the original data points' location are perturbed, the resulting representation of the H_1 feature is also changed. It may be expected that the Hausdorff distance is upper bounded by the thickness of the loop band. However, if we assume the representative loop data points in

fig. 6 are evenly distributed, eq. (2) says the distance should be the line shown in fig. 6b, which is larger than the thickness of the band.



(a) Possible perturbation of representative points



(b) Hausdorff distance between two representatives, each comes from a different sampling instance

Figure 6

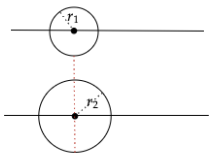


Figure 7: Start with parallel lines, then add balls around some points

If the underlying loop manifold has some width like in fig. 6a, it might be better to add some thickness to the output. Hence, we choose to enlarge the line-segment loop by including balls with radius equal to the birth time of this loop-feature in the VR-filtration, around each representative point. As a side note, enlarging the loop may lead to a bigger Hausdorff distance. In fig. 7, we observe that the new Hausdorff distance (the red dotted line) after adding two balls is bigger than the original distance if $r_2 > r_1$.

A main part of this paper is to find a quantitative way to upper bound the distance between the enlarged representative loops, where each loop is generated from a different set of sampled points. Now we introduce some notations. Suppose each data point is picked randomly in $\mathbb{R}^k$ following some distribution, with c.d.f. F.

Definition 4.1 (probabilistic Hausdorff distance)**.** Suppose we take two sets A, B from a probability space $X \subset \mathbb{R}^k$ equipped with an underlying probability measure P, c.d.f. F, and some metric d (we assume on the outset that F is continuous.) Define the probabilistic Hausdorff distance between A, B via

$$d_{\mathrm{PH}}(A, B) = \frac{1}{2}\left(\frac{1}{P(A)}\int_A d(a, B)dF + \frac{1}{P(B)}\int_B d(b, A)dF\right) \tag{4}$$

where the distance from point to set is defined as in eq. (3). If the underlying probability space X is arbitrary (not $\mathbb{R}^d$), we simply change dF into $d\mu$ and $P(A)$ into $\mu(A)$, where μ is the corresponding probability measure on X.

If there is no pre-specified probability measure on $X \subset \mathbb{R}^d$, we can simply take the uniform p.d.f. If furthermore, set A and B are unbounded, then d_{PH} is undefined. This distance can be thought of as an averaged version of the Hausdorff distance. The motivation is that if the two sets are close in the majority points, we would expect to have a measure that says the difference between the sets is small; however, the original Hausdorff distance only considers the longest pair-wise distance, which can be large.

Proposition 4.1. Some properties of d_{PH},

- $d_{\mathrm{PH}} \geq 0$ and $d_{\mathrm{PH}}(A, B) = d_{\mathrm{PH}}(B, A)$
- $d_{\mathrm{PH}} \leq d_{\mathrm{H}}$
- Given three sets $A, B, C \subset X$, we have

$$d_{\mathrm{PH}}(A, C) \leq \frac{1}{2}\left(d_{\mathrm{PH}}(A, B) + d_{\mathrm{H}}(A, B)\right) + \frac{1}{2}\left(d_{\mathrm{PH}}(B, C) + d_{\mathrm{H}}(B, C)\right)$$

Proof. The first result follows directly from the definition of d_{PH}. For the second bullet point, because

$$\frac{1}{\mu(A)}\int_A d(a, B)d\mu \leq \frac{1}{\mu(A)}\int_A \sup_{a\in A} d(a, B)d\mu = \sup_{a\in A} d(a, B)$$

we have

$$d_{\mathrm{PH}}(A, B) \leq \frac{1}{2}\left(\sup_{a\in A} d(a, B) + \sup_{b\in B} d(b, A)\right) \leq d_{\mathrm{H}}(A, B)$$

Next, for an arbitrarily fixed $b \in B$, and $\epsilon > 0$, we can find some $c \in C$ such that $d(b, C) + \epsilon \geq d(b, c)$. Now observe,

$$\begin{aligned} d(a, C) &\leq d(a, c) \qquad \text{by } d(a, C) := \inf_{c\in C} d(a, c) \\ &\leq d(a, b) + d(b, c) \\ &\leq d(a, b) + d(b, C) + \epsilon \end{aligned}$$

So, we have $d(a, C) \leq d(a, b) + d(b, C) \leq d(a, b) + d_{\mathrm{H}}(B, C)$ for any b. Pass b to infimum, we get $d(a, C) \leq d(a, B) + d_{\mathrm{H}}(B, C)$. With this relation, we now have,

$$\frac{1}{\mu(A)}\int_A d(a, C)d\mu \leq \frac{1}{\mu(A)}\int_A d(a, B) + d_{\mathrm{H}}(B, C)d\mu \leq d_{\mathrm{PH}}(A, B) + d_{\mathrm{H}}(B, C)$$

Similarly, we have

$$\frac{1}{\mu(C)}\int_C d(c, A)d\mu \leq d_{\mathrm{PH}}(B, C) + d_{\mathrm{H}}(A, B)$$

Hence, the third result follows. Since we are not able to show the triangle inequality, i.e., $d_{\mathrm{PH}}(A, C) \leq d_{\mathrm{PH}}(A, B) + d_{\mathrm{PH}}(B, C)$, d_{PH} may not be a metric (we have $d_{\mathrm{PH}}(A, C) \leq d_{\mathrm{H}}(A, B) + d_{\mathrm{H}}(B, C)$ though, by the second property.) But we can just use d_{PH} as a distance measure. □

Next, we discuss one possible way to quantify the bandwidth (or thickness) of a loop M in $\mathbb{R}^2$. The underlying probability space is $\mathbb{R}^2$ with c.d.f. F.

Definition 4.2 (bandwidth of a loop)**.** Suppose manifold M is of a "loop shape". To be more precise, $M \subset \mathbb{R}^2$ is homeomorphic to set $E = D(1) - D(1/2)^\circ \subset \mathbb{R}^2$ (1 and 1/2 are some arbitrary choice, one can use other numbers too) via function $f : X \to M$, where $D(r)$ is the closed disk of radius r and $D(r)^\circ$ is its interior. We call $(\partial M)_i := f(\partial D(1/2))$ the inner boundary of M; and $(\partial M)_o := f(\partial D(1))$ the outer boundary of M. Then, the *average bandwidth* of M is defined via,

$$\tau(M) = \frac{1}{P(M)} \int_M \rho(m, (\partial M)_i, (\partial M)_o) dF \tag{5}$$

where

$$\rho(m, (\partial M)_i, (\partial M)_o) = \inf_{\substack{a \in (\partial M)_i, b \in (\partial M)_o \\ \text{line } (a,b) \text{ passes point } m}} d(a, b) \tag{6}$$

The *peak bandwidth* of M is defined via,

$$\kappa(M) = \sup_{m \in M} \rho(m, (\partial M)_i, (\partial M)_o) \tag{7}$$

We provide two computation examples to illustrate the idea of this "bandwidth".

(a) A circular loop, call it M_1

(b) A rectangular loop, call it M_2

Figure 8: Bandwidth examples

In fig. 8a, the larger radius minus the smaller radius is d. For any point m inside M_1, $\rho(m, (\partial M_1)_i, (\partial M_1)_o) = d$. So, $\tau(M_1) = d$ and $\kappa(M_1) = d$. In fig. 8b, it requires more computation in the four corners, so an upper bound can be defined as

$$\tau(M_2) < \frac{2b^2c + 2a^2d + 4ab\sqrt{a^2 + b^2}}{2bc + 2ad + 4ab}$$

This is because except the four corners, $\rho(m, (\partial M_2)_i, (\partial M_2)_o)$ are constants a and b. Also, since we assume a uniform distribution, we just need to multiply by the areas. If furthermore $a = b, c = d, a \ll c$, then $\tau(M_2) \approx a$ with

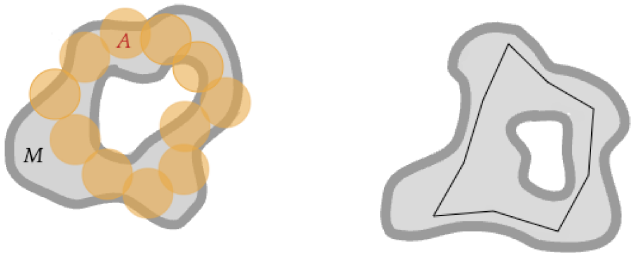
$$\kappa(M_2) = \sqrt{a^2 + b^2} = \sqrt{2}a$$

5 Bound the distance between enlarged loops

The next task is to use the terminology above to bound the distance between two enlarged loops (each generated from a different set of sampled points from the underlying space $X \subset \mathbb{R}^2$.) fig. 9a shows a possible enlarged loop A. We do allow background noises outside the loop manifold M. However, to make the loop M stands out from the background (otherwise it would not make sense to say M is a loop feature because it blends in with the background), we require the distribution density inside M to be higher than its surrounding regions. Also, in our arguments below, we make two assumptions on the outset.

- (A1) assume the representative points of this loop feature are within M;
- (A2) assume the loop representation (the 1-complex polygon) is in between $(\partial M)_o$ and $(\partial M)_i$. To be more precise, the polygon divides X into two path connected components, one containing $(\partial M)_o$ and one containing $(\partial M)_i$ (see fig. 9b).

The first assumption is mild, because by our density requirement, the sampled points near M but on the outside are much more sparse, implying the points inside of M would get connected first under the VR-filtration, thus become the representative points. Furthermore, because the vertices of the 1-complex polygon are the representative points, this polygon is contained in set A.



(a) An enlarged representative loop

(b) Assumption A2

Figure 9

Proposition 5.1. We first bound the distance between enlarged loop A and the underlying manifold M:

$$d_{\mathrm{PH}}(A, M) \leq \frac{1}{2}\left(r \cdot \frac{P(A - M)}{P(A)} + \tau(M)\right) \tag{8}$$

where $\tau(M)$ is the average bandwidth of M.

Proof. Recall

$$d_{\mathrm{PH}}(A, M) = \frac{1}{2}\left(\frac{1}{P(A)}\int_A d(a, M)dF + \frac{1}{P(M)}\int_M d(m, A)dF\right)$$

Now,

$$\begin{aligned}\frac{1}{P(A)}\int_A d(a,M)dF &= \frac{1}{P(A)}\int_{A-M} d(a,M)dF \\ &\leq \frac{1}{P(A)}\int_{A-M} rdF \\ &= r\cdot\frac{P(A-M)}{P(A)}\end{aligned}$$

where r is the birth radius of this loop feature. The first equality is because if $a \in M$, then $d(a,M)=0$; and the inequality comes from our first assumption (A1) that representative points are within M: $d(a,M) \leq d(a, \text{center of the circle}) = r$.) Next,

$$\begin{aligned}\frac{1}{P(M)}\int_M d(m,A)dF &= \frac{1}{P(M)}\int_{M-A} d(m,A)dF \\ &\leq \frac{1}{P(M)}\int_{M-A} \rho(m,(\partial M)_i,(\partial M)_o)dF \\ &\leq \tau(M)\end{aligned}$$

The first inequality is because of the following. For any m, there exist $a \in (\partial M)_i$ and $b \in (\partial M)_o$ such that $d(a,b)=\rho(m,(\partial M)_i,(\partial M)_o)$ (here boundaries are compact sets.) Because the polygon (the loop representation by the VR-filtration) does not cross itself on edges, Jordan curve theorem tells us that line segment (a,b) must cross this polygon (by our second assumption, point a lies in the interior of the polygon, and point b lies in the exterior.) Suppose line (a,b) crosses the polygon at a point x. Then,

$$d(m,A) \leq d(m,\ \text{polygon}) \leq d(m,x) \leq \rho(m,(\partial M)_o,(\partial M)_i)$$

□

As a side note, by our density assumption, $P(A-M)$ should be much smaller than $P(A)$; and intuitively speaking, $\tau(M)$ is a very loose bound.

Proposition 5.2. Now we bound the distance between two enlarged loops A and B, i.e., $d_{\mathrm{PH}}(A,B)$. Use the notation C_A to denote the collection of centers from enlarged loop A, and r_A to denote the radius of those circles; similarly for C_B and r_B. There are three different upper bounds,

$$\begin{aligned}\frac{1}{2}\Bigg(&\frac{P(M)}{P(A)}\cdot\tau(M)+\frac{P(A-M)}{P(A)}(r_A+\kappa(M)) \\ &+\frac{P(M)}{P(B)}\cdot\tau(M)+\frac{P(B-M)}{P(B)}(r_B+\kappa(M))\Bigg)\end{aligned} \tag{9}$$

$$\kappa(M)+\frac{1}{2}\left(r_A\cdot\frac{P(A-M)}{P(A)}+r_B\cdot\frac{P(B-M)}{P(B)}\right) \tag{10}$$

$$|r_A - r_B| + d_{\mathrm{H}}(C_A, C_B) \tag{11}$$

each involves a different main term.

Proof. First observe,

$$\begin{aligned}
\frac{1}{P(A)} \int_A d(a, B) dF &= \frac{1}{P(A)} \left(\int_{A \cap M} d(a, B) dF + \int_{A - M} d(a, B) dF \right) \\
&\leq \frac{1}{P(A)} \Bigg[\int_{A \cap M} \rho(a, (\partial M)_i, (\partial M)_o) dF \\
&\qquad + \int_{A - M} r_A + \rho(c_a, (\partial M)_i, (\partial M)_o) dF \Bigg] \\
&\leq \frac{P(M)}{P(A)} \cdot \tau(M) + \frac{P(A - M)}{P(A)} (r_A + \kappa(M))
\end{aligned}$$

The first term in the first inequality follows from the same reasoning as the previous proposition. Unfortunately, the information of B is not encoded in this term because we have no idea where B is relative to A, so we can only treat a as some point in M, not related to B. The second term follows from the triangle inequality: we first travel from a to the center (call it c_a) of any circle that contains it, and then from that center to B. The same type of arguments work for the other term in d_{PH}.

For the second bound, we increase $\rho(a, (\partial M)_i, (\partial M)_o)$ to $\kappa(M)$ in the second inequality above, and then simplify from there. Next, again by triangle inequality, $d(a, B) \leq \max(0, r_A + d(c_a, c_b) - r_B)$, where c_b is the nearest center of B to c_a.

$$\begin{aligned}
\frac{1}{P(A)} \int_A d(a, B) dF &\leq \frac{1}{P(A)} \int_A \max(0, r_A - r_B + d(c_a, c_b)) dF \\
&\leq \frac{1}{P(A)} \int_A |r_A - r_B| + d(c_a, c_b) dF \\
&\leq \frac{1}{P(A)} \int_A |r_A - r_B| + d_{\mathrm{H}}(C_A, C_B) dF \\
&= |r_A - r_B| + d_{\mathrm{H}}(C_A, C_B)
\end{aligned}$$

□

The first upper bound involves the average bandwidth of M; the second bound involves the peak bandwidth of M, and the third bound is mainly dominated by the Hausdorff distance between the collection of centers.

There are three things to mention here. First, by our density assumption, $P(A - M) \ll P(A)$ and $P(B - M) \ll P(B)$, which implies the term involving this factor is small. Second, the first two upper bounds rely on assumption (A2), while the third one does not. Third, we cite the theorem below proved by Chazal et al. [1],

Theorem (Chazal et al.)**.** Generate the VR-filtration from two totally bounded spaces X and Y, we have,

$$d_b(\mathrm{dgm}(H(\mathrm{Rips(X)})), \mathrm{dgm}(H(\mathrm{Rips(Y)}))) \leq 2d_{\mathrm{GH}}(X, Y)$$

where d_b is the bottleneck distance between the two resulting persistence diagrams, and d_{GH} is the Gromov-Hausdorff distance.

In our case, X, Y can be thought of as the two sets of finitely many sampled points (so they are automatically totally bounded.) If we sample many points, then X and Y would be dense in the loop manifold region, implying d_{GH} can be small too. This means the birth and death radius of that loop feature in the persistence diagram is close to each other, or in other words, r_A is close to r_B. This is helpful because in our first and second bound, due to $P(A-M)/P(A) \approx 1$, $P(B-M)/P(B) \approx 1$, a stable r_A, r_B implies the bound itself is stable. The third upper bound, although has a simpler form and avoids using assumption (A2), does not enjoy the stability property because $d_{\mathrm{H}}(C_A, C_B)$ is unstable.

Next, instead of just enlarge the loop by adding balls, we can put a "band" of width r around the representative loop (see fig. 10.) In this case, the loose upper bound can be derived as,

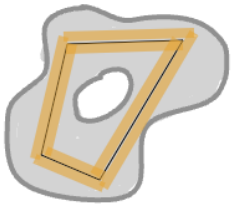


Figure 10: Add a band around the representative loop

$$\begin{aligned}
\frac{1}{P(A)}\int_A d(a,B)dF &= \frac{1}{P(A)}\left[\int_{A\cap M} d(a,B)dF + \int_{A-M} d(a,B)dF\right] \\
&\leq \frac{1}{P(A)}\Bigg[\int_{A\cap M} \max(\rho(a,(\partial M)_i,(\partial M)_o) - r_A, 0)dF + \\
&\qquad \int_{A-M} r_A + \max(\rho(c_a,(\partial M)_i,(\partial M)_o) - r_A, 0)dF\Bigg] \\
&\leq r_A \cdot \frac{P(A-M)}{P(A)} + \frac{1}{P(A)}\int_A |\kappa(M) - r_A| dF \\
&= r_A \cdot \frac{P(A-M)}{P(A)} + |\kappa(M) - r(A)|
\end{aligned}$$

which gives

$$\frac{1}{2}\Bigg(r_A \cdot \frac{P(A-M)}{P(A)} + |\kappa(M) - r(A)| \\ + r_B \cdot \frac{P(B-M)}{P(B)} + |\kappa(M) - r(B)|\Bigg) \tag{12}$$

6 Discrete case

We first go back to the definition of the probabilistic Hausdorff distance. Now suppose set A and set B only have finite number of discrete points. Instead of taking the integral, we simply take the direct average.

Definition 6.1 (discrete version of d_{PH})**.** In the discrete setting here, we define

$$d_{\mathrm{PH}}(A,B) = \max\left(\frac{1}{n}\sum_{a\in A} d(a,B),\ \frac{1}{m}\sum_{b\in B} d(b,A)\right)$$

where A has n points, and B has m points.

Again, we start with two sets of representative points C_A and C_B as before. We still want to enlarge the loop. But in real data sets, we should not simply report back those circles like in figure 4(a). This is because we have no prior knowledge about the underlying probability distribution of X, so if we are given those circles, we can only treat them as full solid disks (with points uniformly distributed in each.) However, this may lead us to mistakenly add extra points that should not exist (see figure 6.)

Figure 11: If we take in the full disks, then we accidentally add too many extra points to the empty or low background noise region

Hence, the modified enlarged loop is A' = (original A)$\cap$(sampled points of X), a discrete set. In this way, even if we include some points outside of M, by our density assumption (A1), there shouldn't be too many. The question now is: with the discrete setting, would our upper bound still be reasonable? We have

this question because in the discrete setting, $A' \subset A$ and $B' \subset B$, which implies the distance $d(a, B')$ may be bigger than $d(a, B)$. Before answering this question, we first discuss a little about the empirical c.d.f.

Definition 6.2. The empirical c.d.f. from n independently sampled points $X_1, X_2, ..., X_n$ is defined via,

$$F_n(\boldsymbol{x}) = \frac{1}{n} \sum 1(X_i \leq \boldsymbol{x})$$

where $\boldsymbol{x}$ is a vector in $\mathbb{R}^d$.

Theorem 6.1 (DKW's inequality)**.** Let F_n be the empirical c.d.f. defined above, for any $\epsilon > 0$ and $z > 0$, there exists a corresponding constant $C_{\epsilon,d}$ such that

$$P(\|F_n - F\|_\infty > z) \leq C_{\epsilon,d} e^{-(2-\epsilon)nz^2}$$

A nice corollary of DKW's inequality is that $F_n \overset{a.s.}{\to} F$ uniformly (one can show this using a useful fact: $X_n \overset{a.s.}{\to} a$ if and only if $\sum_n P(|X_n - a| > \epsilon) < \infty$ for any $\epsilon > 0$.) A weaker conclusion from this is: $F_n \overset{d}{\to} F$ a.s. Now, let's see the connection between the discrete d_{PH} and its continuous version. Because F_n is defined on all $\mathbb{R}^d$, we can still use this c.d.f. in the continuous formula.

Proposition 6.1. In the discrete setting, we sample N points in total from X, where n of them are in A'. Then we sample again for another N points in total, and m of them are in B'. If N is large, $d_{\mathrm{PH}}(A', B')$ is roughly bounded by (again, a very loose one)

$$\max(r_A, r_B) + d_{\mathrm{H}}(C_A, C_B)$$

Proof. Write the empirical c.d.f. from the first round of sampling as F_N, and the empirical c.d.f. from the second sample as G_N. First observe

$$\int_A d(a, B') dF_N = \frac{1}{N} \sum_{a \in A'} d(a, B')$$

Moreover,

$$\begin{aligned} \int_A d(a, B') dF &\leq \int_A r_A + d(c_a, c_b) dF \\ &\leq \int_A r_A + d_{\mathrm{H}}(C_A, C_B) dF \\ &= (r_A + d_{\mathrm{H}}(C_A, C_B)) P(A) \end{aligned}$$

where c_a is the center of any circle in A that contains point a, and c_b is the center of B that is the nearest to c_a. The first inequality is by triangle inequality. Next,

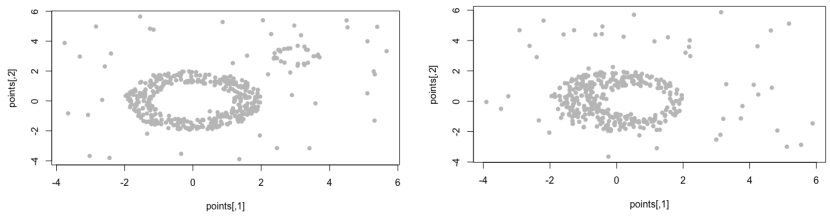
$$\int_A d(a, B') dF_N = \int_{\mathbb{R}^d} 1_A \cdot d(a, B') dF_N \to \int_{\mathbb{R}^d} 1_A \cdot d(a, B') dF = \int_A d(a, B') dF$$

This is because $1_A \cdot d(a, B')$ is a bounded function, and is continuous except on ∂A, which has $F(\partial A) = 0$; then by $F_N \xrightarrow{d} F$, the convergence follows (for details of such convergence result, please refer to Durret's Probability section 3.10.) Furthermore, $n/N = P_N(A) \to P(A)$ by the similar reasoning (P_N is the probability measure induced from F_N.) So, if N is large, $1/n \sum_{a \in A'} d(a, B')$ is roughly bounded by $r_A + d_{\mathrm{H}}(C_A, C_B)$. Same argument gives the bound for $1/m \sum_{b \in B'} d(b, A')$. □

We can compare this bound with quantity ($\star\star\star$) from proposition 3.2. Since A' and B' are subsets of A and B, it's even harder to provide a tight bound. So instead of $|r_A - r_B|$, we have $\max(r_A, r_B)$ here. Although the bound given here is very loose, it at least confirm that our enlarged loop would not be flying everywhere. We present couple examples to show that the actual distance between the enlarged loops can be much smaller than the bound here.

7 Examples

We show two examples that confirm the actual averaged Hausdorff distance is much lower than the theoretical bound we provide.

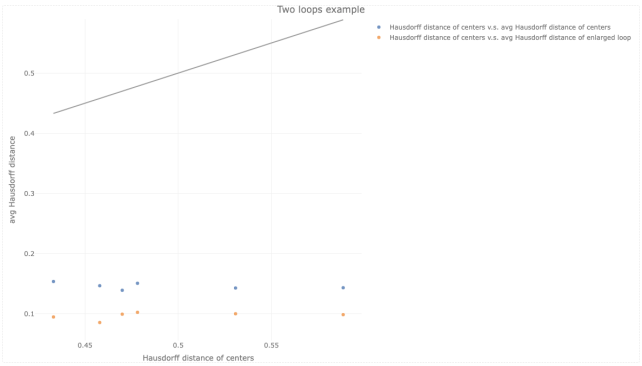


(a) Two circle example (b) Loop with non-uniform thickness

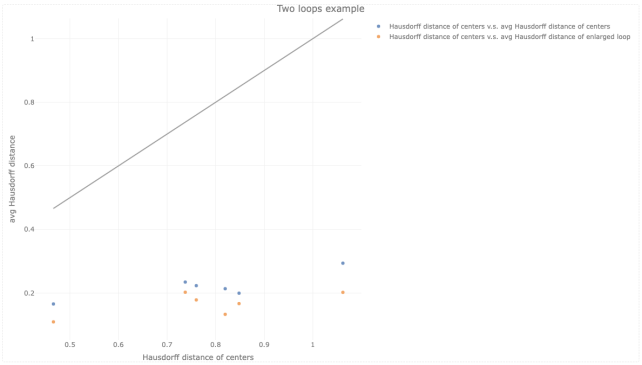
Figure 12

We make a note that the upper bound proposed in the discrete case above, though a very loose one, is not unreasonably large. For example, in figure 1(c), we sample four points each time (the sampled points are evenly distributed along the inner and outer boundary respectively.) It is easy to check that the discrete version of d_{PH} can be exactly $d_{\mathrm{H}}(\text{green dots}, \text{orange dots})$, which is a main term in our upper bound. If the sampled points are densely located along the boundary, then r_A, r_B would be small, which implies the upper bound term $\max(r_A, r_B)$ is small.

On the other hand, in real data sets, the actual distance between the two enlarged sets are much smaller than our theoretical upper bound (see figure 8.) This is because for example, the distance $d(a, B)$ for $a \in A$ can be much smaller in reality, because often times the enlarged circles overlap, meaning a is probably right inside B, or at least close to B. We cannot utilize this observation in our



(a) Comparison between the original Hausdorff distance and averaged Hausdorff distance between representative-point set and enlarged loop for Figure 7(a)-like sample



(b) Comparison for Figure 7(b)-like sample

Figure 13

theoretical bound proof because we have no idea about the exact location of those circles.

References

[1] Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

[2] Jessi Cisewski-Kehe, Brittany Terese Fasy, Wojciech Hellwing, Mark R Lovell, Paweł Drozda, and Mike Wu. Differentiating small-scale subhalo distributions in cdm and wdm models using persistent homology. *Physical Review D*, 106(2):023521, 2022.

[3] Tamal K Dey, Tao Hou, and Sayan Mandal. Persistent 1-cycles: Definition, computation, and its application. In *Computational Topology in Image*

Context: 7th International Workshop, CTIC 2019, Málaga, Spain, January 24-25, 2019, Proceedings 7, pages 123–136. Springer, 2019.

[4] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 1983.

[5] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*, 2014.

[6] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017.

[7] Bastian Rieck, Ulderico Fugacci, Jonas Lukasczyk, and Heike Leitte. Clique community persistence: A topological visual analysis approach for complex networks. *IEEE transactions on visualization and computer graphics*, 24(1):822–831, 2017.

[8] Ann E Sizemore, Jennifer E Phillips-Cremins, Robert Ghrist, and Danielle S Bassett. The importance of the whole: topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673, 2019.

[9] Xin Xu and Jessi Cisewski-Kehe. Emt: Locating empty territories of homology group generators in a dataset. *Foundations of Data Science*, 1(2):227–247, 2019.

[10] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959, 2013.