Directed Persistence

Greg Bell & Thanos Gentimis

June 8, 2015

Abstract

In this paper we explore the idea of *directed persistence*, an extension of persistent homology, where the original Čech and Rips complexes are determined by use of a skewed metric. The result, for various well defined metrics, is equivalent to the use of directed ellipses replacing the isotropic balls of the traditional construction. We explore this idea experimentally and present a means to identify global directions for features, while exploring the characteristics of the set of directed persistence diagrams as generalized persistence modules.

1 Introduction

Gunnar Carlsson has said that the main motivation behind computational data analysis is that, "data has shape and the shape matters." In this paper we attempt to extend that notion to the shape of the features (holes, voids, etc) and claim that they may have some characteristic direction, which may in turn be of importance. We propose a means of exploring this shape, in the case of point-clouds embedded in some Euclidean space, with a structure which is coarse enough to yield fast computations, but also informative enough to give a global direction of these shapes if such a direction exists. The results presented in this paper are based on synthetic data but a full implemented method will be useful in biomedical applications (protein docking), geological analysis (oil pockets) and others.

The ideas presented here are a mixture of topological tools, with persistent homology being the one used mostly, and geometric tools reflected by the choice of the metric we endow our space of interest with. In that light, this paper can be thought of as an exploration of persistent homology, using different, intrinsic metrics on a specific data set. Using the same rationale as (say) the iso-map technique [21], our method improves the topological reconstruction of the space in question, since not only does it yield a coarsely correct number of features, it also gives an idea of the relative position and direction of these features as they are imbedded in our representation space. This extra information does come at a cost, since at this stage the computations are repeated for each chosen direction. In order to alleviate that problem we propose to examine basic properties of this persistent module, to perhaps extract a faster, more economic way of computing these structures through analyzing its algebraic structure. We will explore this structure following similar ideas to [3, 7] and [4]

In the paper *Persistent Homology of Delay Embeddings and its Application to Wheeze Detection* [11] the authors presented a novel use of persistent homology to detect wheezes, by viewing the topological invariants of the 2 dimensional delay embedding of the signal. The point-cloud that a wheeze yields has a characteristic one dimensional persistent loop, which resembles an ellipse, with major axis forming an angle of 30° to 60° with respect to the x-axis. The point-cloud of regular breathing sounds does not have such a persistent loop and thus a concise and robust to noise algorithm of detection was proposed. During the analysis of the various signals, in some wheezes the holes were present but since they were stretched towards a direction and slim towards the perpendicular direction, it was hard to recover them through normal persistence, since the corresponding lifetime was short. The method described in this paper would make the lifetimes of those persistent holes bigger if the directions of the ellipses used was within the $30° - 60°$ range.

In order to implement the computations of persistence we used javaplex [20], after we preprocessed the distance matrix with our own implementation. During our final stages of our computations we used perseus[16], a persistent homology computational tool based on Discrete Morse Functions, which improved the computational complexity drastically. In the future we will pursue a further extension of our current algorithm using a modification of Perseus.

This paper is organized as follows: In section 2 we present some basic notions about algebraic topology and regular persistent homology. In section 3 we give the basic definitions and theorems related to directed filtrations. In section 4 we present an algorithm for estimating the direction of topological features using directed filtrations, present our results on simulations and give a brief theoretical justification. In section 5 we explore the properties of the space of directed filtration as an object of different categories. Finally in 6 we review and conclude the paper and pose some open problems.

We want to thank Hamid Krim and Harish Chintakunta for various interesting conversations on the subject, Vidit Nanda for helping us with questions about Perseus, Henry Adams for his explanations of the javaplex algorithm, and Primoz Skraba and Peter Bubenik for their ideas and discussions pertaining the algebraic properties of the generalized directed persistent module. Finally we would like to thank the IMA, for its support; this work was begun during a workshop at the IMA while the first author was an IMA New Directions Professor during its Annual Thematic Year on Scientific and Engineering Applications of Algebraic Topology.

2 Preliminaries

We will first review some basic notions from algebraic topology. For more details we refer the reader to [8, 13].

We will be interested in computing the homology of a simplicial complex K with coefficients in a field. Generally speaking we will be thinking of the field as either $\mathbb{Z}_2$ or $\mathbb{Q}$. We will suppress the field and denote the i-th homology vector space of the simplicial complex K by simply $H_i(K)$. The rank of this vector space is called the *i-Betti number* of K and is denoted by $\beta_i = \dim(H_i(K))$. The basic topological structure of K is quantified by the number of independent cycles in each homology space. For example, $\beta_0(K)$ quantifies the number of path components of K, $\beta_1(K)$ quantifies the loops in K, and $\beta_2(K)$ quantifies the number of "voids" in K.

One of the key features of homology is *functoriality*. Homology is a *functor* from the category of topological spaces with continuous maps to the category of vector spaces and linear maps. More precisely, it assigns to each topological space a vector space and for any continuous map $f : X \to Y$ between topological spaces, there is an induced linear map $H_i(f) : H_i(X) \to H_i(Y)$, such that $H_i(f \circ g) = H_i(f) \circ H_i(g)$ and $H_i(I_X) = I_{H_i(X)}$, where I denotes the identity map.

Let $X = \{x_1, x_2, \ldots, x_k\}$ be a finite subset of points in some Euclidean space $\mathbb{R}^n$. Given $r \geq 0$ we construct a simplicial complex from X as the nerve of the union of $B_r(x_i)$. This complex is called the *Čech complex* and is denoted $\check{C}_r(X)$. The Čech Complex over a set of points $X = \{x_1, x_2, ..., x_n\}$ in $\mathbb{R}^n$ is constructed as follows. Let $r > 0$ be a constant and consider the balls of radius r around each point of that set. This subset of $\mathbb{R}^n$ is called a *cover* of X. We then consider an abstract collection of *vertices* $\{v_1, v_2, ..., v_n\}$ where v_i corresponds to the point x_i, for each i. Connect the vertices v_i, v_j by an edge if the balls $B_r(x_i)$ and $B_r(x_j)$ intersect. Add the triangle $[v_i, v_j, v_k]$ if all three of the balls $B_r(x_i), B_r(x_j), B_r(x_k)$ intersect and continue similarly for higher order simplices. This procedure is known as taking the *nerve* of the cover. The output of this procedure is an abstract simplicial complex called the Čech complex (see Figure 1). The nerve theorem tells us that the topological invariants of union of balls of the cover are the same with those of the Čech complex [10]. Thus, this complex is a natural enlargement of the point cloud and can be used to glean its topological properties.

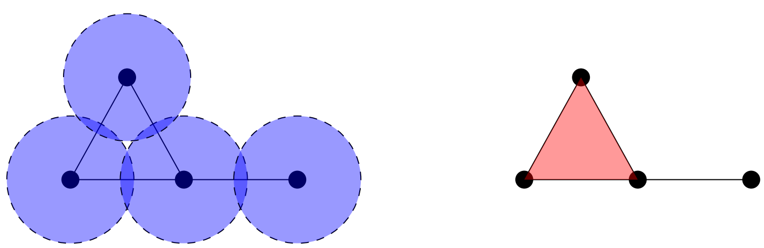

Figure 1: Construction of the Čech complex (left) and the Rips complex (right) on the same set of points with the same r value. Observe that the Rips complex is the flag over the corresponding Čech complex.

An approximation of the Čech complex that is more amenable to computation is the Vietoris-Rips complex, which we describe presently. The input in this case is simply the distances between the points X; an embedding in $\mathbb{R}^n$ is not required. First, we consider an abstract collection of vertices $\{v_1, v_2, ..., v_n\}$ corresponding to the points of X. Then, for a parameter $r > 0$ we connect the vertices v_i, v_j by an edge if the distance between the point x_i and x_j is less or equal to $2r$. This will yield the same vertices and edges as the Čech Complex defined above. Since the actual embedding of the points is not given, one cannot proceed with the Čech construction. Instead, we proceed by adding the triangle $[v_i, v_j, v_k]$ if all the three edges $[v_i, v_j]$, $[v_j, v_k]$, $[v_k, v_i]$ exist. We then continue by adding higher dimensional simplices if all their lower dimensional faces exist. This is the flag complex construction of the one dimensional skeleton. The computations performed in this paper rely on algorithms that utilize this flag complex construction.

If we imagine that X is a finite sampling of an underlying topological space $\mathbf{X}$ with some noise, then an interesting problem is to determine the homology of $\mathbf{X}$ from the sample X. It is clear that this depends greatly on the density of the sample (see [17]). It is also easy to see that for extremely small or extremely large values of r, the homology of any simplicial complex approximation $K(X, r)$ (such as the Čech or Rips complexes described above) will be a poor reflection of that of $\mathbf{X}$. Instead of selecting a particular value of r and computing the homology of a particular $K(X, r)$, one considers all values of r at once and observes the evolution of the topological information as r increases.

More specifically, given an increasing sequence of values $r_1 < r_2 < \cdots < r_k$, put $K_i = K(x, r_i)$. Then, one gets a nested sequence of simplicial complexes with vertex set X:

$$K_1 \subseteq K_2 \subseteq \cdots \subseteq K_i \subseteq \cdots \subseteq K_k.$$

If $H_n^i = H_n(K_i)$, then from the functorial properties of homology one gets a sequence of vector spaces and linear maps, which Zomorodian and Carlsson [22] call a *persistence module*:

$$H_n^1 \to H_n^2 \to \dots H_n^i \to \cdots \to H_n^k.$$

A particular class α may come into existence in $H_n(K_t)$ (for some t). We call t the *birth time of* α. Then it is either mapped to some previously existing class in $H_n(K_s)$ for some $s > t$, or it maps to a non-zero element in the last homology $H_n(K_k)$. In the former case, we call s the *death time of* α. This yields a *barcode for* α: an interval $[t, s]$ representing its birth time and death time. The collection of all such barcodes (for all classes in H_n) is also called a *barcode*. The key philosophy is that classes that correspond to barcodes with longer lives represent "real" topological features of $\mathbf{X}$ and are not simply aberrations caused by noise or sampling. An example of a barcode and corresponding complex is given in the next section.

A more detailed exposition about the ideas of persistent homology can be found in [2,9,22]

3 Directed Filtrations

3.1 Construction Using Ellipses

Let X be a finite set of points in $\mathbb{R}^n$. More formally, one could take a set $X = \{x_1, \dots, x_m\}$ and a collection $\{f_1, \dots, f_n\}$ of functions (called features) such that $f_i : X \to \mathbb{R}$ for each $1 \leq i \leq n$. We will implicitly identify X with its image $F(X) = \{f_1(x), f_2(x), \dots, f_n(x) \mid x \in X\} \subseteq \mathbb{R}^n$. We need not assume that F is one-to-one. If it is not, depending on the particular application of interest one could ignore points with the same features or perturb one feature by a particular (small) ϵ to create two features that are very close. The second method is particularly useful when each point is significant to the application.

Let $r > 0$ be given and let $\vec{u} \in S^{n-1}$. Really, we want to choose a direction for the axis of an ellipse; that is, an element of $\mathbb{R}P^{n-1}$. We represent this axis by a choice of an element of S^{n-1} Choose some $\mathbf{c} = \{c_1, c_2, \ldots, c_{n-1}\} \in (0,1)^{n-1}$ and construct a hyper-ellipsoid in $\mathbb{R}^n$ by making the major axis parallel to $\vec{u}$ with semi-length r and in all directions perpendicular to $\vec{u}$ we take semi-lengths of minor axes equal to $c_i \cdot r$. In other words, if $\{\vec{u}, \vec{u}_1, \ldots, \vec{u}_{n-1}\}$ is an orthonormal basis for $\mathbb{R}^n$, we take the image of the unit disk $D^n \subset \mathbb{R}^n$ under the matrix $[r\vec{u},\ c_1 r\vec{u_1},\ \cdots\ c_{n-1} r\vec{u}_{n-1}]$. Set $E_{\vec{u}}(x, r)$ to be equal to this image centered at $x \in \mathbb{R}^n$. We often fix X and $\mathbf{c}$ while allowing values of r and $\vec{u}$ to vary.

Remark 1. In most cases we will assume that for all $i = 1, 2, \ldots, n-1$, $c_i = c$, for some $0 < c < 1$. All the propositions that we will discuss can be generalized to include distinct c_i's with slight changes in the proofs.

Let $K_{r_i}(\vec{u})$ denote the nerve of the set $\bigcup_{x \in X} E_{\vec{u_1}}(x, r_i)$. There is a natural map induced by inclusion from $K_{r_i}(\vec{u}) \to K_{r_{i+1}}(\vec{u})$, for $i \in \{0, 1, \cdots, k-1\}$.

Thus, for any (finite) increasing sequence $I = \{r_0 < r_1 < \cdots < r_k\}$ we obtain a nested sequence of simplicial complexes

$$K_{r_1}(\vec{u}) \subseteq K_{r_2}(\vec{u}) \subseteq \cdots \subseteq K_{r_k}(\vec{u}) \tag{1}$$

When $\vec{u}$ and I are understood, we simplify notation by writing

$$\vec{K}_1 \subseteq \vec{K}_2 \subseteq \cdots \subseteq \vec{K}_k.$$

We call this a *directed filtration* of X with *direction* $\vec{u}$*, coefficients* $\mathbf{c}$ *and increments* I. Let

$$H_n^1 \to H_n^2 \to \ldots H_n^i \to \cdots \to H_n^k$$

the corresponding sequence of homologies induced by inclusions.

Since the ellipses are convex, the Nerve Theorem implies the homology of each $\vec{K}_i$ is the same as that of $\bigcup_{x \in X} E_{\vec{u}}(x, r_i)$.

The critical features of the construction are that the sets should be closed and convex (for the Nerve Theorem to apply). Thus, one can easily replace these ellipses with other families of covers sharing these properties and achieve the same outcome that is described above. It should be noted that we have assumed $\vec{u}$ to be a unit vector. There is no harm in taking vectors of non-unit length and by tweaking the constants of multiplication one can recover the original persistence from this construction.

Definition 2. Using the technique in [22] we can construct the multi-set of *directed persistence barcodes*, where an interval $[b, d]$ corresponds to the creation of a homology class at H_n^b that gets mapped to zero at H_n^d. There is also a corresponding persistence diagram, which is a subset of $\mathbb{R}^2$ in which the point (b, d) corresponds to a homology class born in H_n^b that dies in H_n^d. In this set of points we append infinite copies of the diagonal to create the *directed persistence diagram* denoted by $Dgm(X, \vec{u}, I)$. If we are only interested in barcodes/diagrams for a specific homology dimension n we use the notation $Dgm_n(X, \vec{u}, I)$.

Remark 3. If the direction $\vec{v}$ and the parameter increments I are fixed we will use the simpler notation $Dgm(X)$ for the directed persistence diagram. Also since each of these diagrams is equivalent to the directed persistent barcode, which we denote with $Bar(X)$, we will use both terms interchangeably.

Include the figure of your presentation with 4-5 K_r's using directed Rips complexes (the actual ellipses growing) and next to it the persistent diagram that corresponds to this

Remark 4. One could also extend this definition for an $I = [a, b]$ or even $I = [a, +\infty)$. For the remainder of the paper though we will consider only finite directed filtrations unless stated otherwise.

Example 1. Consider a finite point cloud X in $\mathbb{R}^n$. One can perform techniques like Principal Component Analysis (PCA) and compute a set of eigenvalues $\{\lambda \geq \lambda_1 \geq ... \geq \lambda_{n-1}\}$ and corresponding eigenvectors $\{\vec{u}, \vec{u}_1, ..., \vec{u}_{n-1}\}$ that characterize the data. Then these vectors can be extended to an orthogonal basis of $\mathbb{R}^n$ so we can construct a *principal directed filtration* by choosing a $\mathbf{c}$ and a (finite) non-negative increasing sequence $I = \{r_0 < r_1 < \cdots < r_k\}$ as described before.

3.2 Construction Using skewed Metrics

Let us now present an equivalent way of defining directed filtrations, which is a bit more technical and less intuitive but it is more practical especially for computations. Instead of using ellipses on the regular Euclidean distance one can use a skewed distance on $\mathbb{R}^2$ and regular balls. Namely, suppose that we want to construct a planar ellipse entered at zero with major semi-axis of length α parallel to the x axis and minor semi-axis β and angle with the x-axis θ. Then consider the metric:

$$\begin{aligned} d_\theta(A,B)^2 &= \frac{[\cos\theta(x_B - x_A) + \sin\theta(y_B - y_A)]^2}{\alpha^2} \\ &+ \frac{[-\sin\theta(x_B - x_A) + \cos\theta(y_B - y_A)]^2}{\beta^2} \end{aligned}$$

The wanted ellipse is nothing more than $B_\theta(\mathbf{O}, 1)$ i.e. the ball of of radius 1 around the origin, with the metric d_θ. Suppose now that X is a finite set of points in $\mathbb{R}^2$ and consider the function $f_\theta^X : \mathbb{R}^2 \to R^+$ with

$$f_\theta^X(A) = d_\theta(A, X) = \min_{(x) \in X} \{d_\theta(A, x)\}$$

If we the point set X is fixed we will simplify the notation to f_θ. Clearly then

$$\bigcup_{x \in X} E_{\vec{u}}(x, r_i) = f_\theta^{-1}(r_i)$$

where $\vec{u}$ is the vector forming an angle of θ degrees with the x axis and $E_{\vec{u}}(x, r_i)$ are the ellipses defined in the previous sections.

Thus we obtain a sequence of topological spaces

$$X = f_\theta^{-1}(0) \subseteq f_\theta^{-1}(r_1) \subseteq f_\theta^{-1}(r_2) \subseteq \cdots \subseteq f_\theta^{-1}(r_k)$$

From the nerve lemma, we get that the persistent homology chain sequence corresponding to the sequence above is the same as the one obtained from 1

3.3 Properties

Definition 5. Let X be a finite point cloud in $\mathbb{R}^n$, $\mathbf{c}$, and $I = \{r_0 < r_1 < \cdots < r_k\}$ be a sequence of non-negative numbers as before. A direction vector $\vec{u} \in S^{n-1}$ is said to be *regular* if and only if $E_{\vec{u}}(x_i, r_m) \cap E_{\vec{u}}(x_j, r_m)$ is either empty or it contains a non-empty open set. This also implies that if $E_{\vec{u}}(x_i, r_m) \cap E_{\vec{u}}(x_j, r_m) \neq \emptyset$ then $\lambda(E_{\vec{u}}(x_i, r_m) \cap E_{\vec{u}}(x_j, r_m)) > 0$, where $\lambda(\cdot)$ denotes the Lebesgue measure.

Lemma 6. *Let X be a point cloud in $\mathbb{R}^n$, $I = \{r_0 < r_1 < \cdots < r_k\}$ a sequence of non-negative numbers, and $\mathbf{c}$ as before. Suppose that $\vec{v}$ is a regular vector. Then, there exists a $\delta > 0$ such that for all $\vec{u} \in S^{n-1}$ for which $\|\vec{u} - \vec{v}\| < \delta$ the nested sequences $\mathcal{K}(u) = \{K_{r_1}(\vec{u}) \subseteq K_{r_2}(\vec{u}) \subseteq \cdots \subseteq K_{r_n}(\vec{u})\}$ and $\mathcal{K}(v) = \{K_{r_1}(\vec{v}) \subseteq K_{r_2}(\vec{v}) \subseteq \cdots \subseteq K_{r_n}(\vec{v})\}$ are identical.*

Proof. Since $\vec{v}$ is a regular vector it means that whenever $x_i, x_j \in X$, have the property that $E_{\vec{u}}(x_i, r_m) \cap E_{\vec{u}}(x_j, r_m) \neq \emptyset$, then $E_{\vec{u}}(x_i, r_m) \cap E_{\vec{u}}(x_j, r_m)$ contains an open set. Thus, the intersection contains an open ball $B(y, a)$, where $y = y(x_i, x_j, r_m)$ and $a = a(x_i, x_j, r_m)$ will depend only on x_i, x_j, and r_m. With a simple geometric argument one can find a $d = d(x_i, x_j, r_m)$ such that if we perturb the two ellipses by a vector with length at most d in opposite directions they will still intersect in points contained in $B(y, a)$. In other words if we construct the vector $\vec{v} = \vec{u} + \vec{s}$ with $\|\vec{s}\| \leq d$ then $E_{\vec{v}}(x_i, r_m) \cap E_{\vec{v}}(x_j, r_m) \neq \emptyset$ and the intersection contains points in $B(y, a)$. Consider then $\delta_1 = \min\{d(x_i, x_j, r_m) : x_i, x_i \in X, r_m \in I\}$. Such a delta exists since the sets X,I are finite.

Thus, we have shown that there is a $\delta_1 > 0$ so that any simplex in $K_{r_i}(\vec{v})$ will also be in $K_{r_i}(\vec{u})$. For the opposite inclusion, we must show that there is a $\delta_2 > 0$ so small that whenever $E_{\vec{v}}(x_i, r_m) \cap E_{\vec{v}}(x_j, r_m) = \emptyset$ then for all $\vec{u}$ with $\|\vec{u} - \vec{v}\| < \delta_2$ we have $E_{\vec{u}}(x_i, r_m) \cap E_{\vec{u}}(x_j, r_m) = \emptyset$. Clearly $\delta = \min\{\delta_1, \delta 2\}$ will satisfy the requirements of the theorem. Since $E_{\vec{v}}(x_i, r_m) \cap E_{\vec{v}}(x_j, r_m) = \emptyset$ and these sets are closed, one can apply the fact that the (metric) space is normal to find μ so small that $N_\mu(E_{\vec{v}}(x_i, r_m)) \cap N_\mu(E_{\vec{u}}(x_j, r_m)) = \emptyset$. As above, we minimize this value over all (finitely many) indices and apply a simple geometric argument to assert the existence of a $\delta_2 > 0$ so that the when $\vec{u}$ and $\vec{v}$ are δ_2-close, we have $E_{\vec{u}}(x_i, r_m) \subset N_\mu(E_{\vec{v}}(x_j, r_m))$ and so, $K_{r_i}(\vec{u}) \subset K_{r_i}(\vec{v})$, when $\|\vec{u} - \vec{v}\| < \delta_2$. A pictorial representation of the proof is found in Figure 3.

□

Lemma 7. *Let X be a point cloud in $\mathbb{R}^n$, $I = \{r_0 < r_1 < \cdots < r_k\}$ a sequence of non-negative numbers, and $\mathbf{c}$ as before. The set of all non-regular directions in S^{n-1} has Lebesgue measure 0. Similarly, the set of all non-regular directions in B^n has measure 0.*

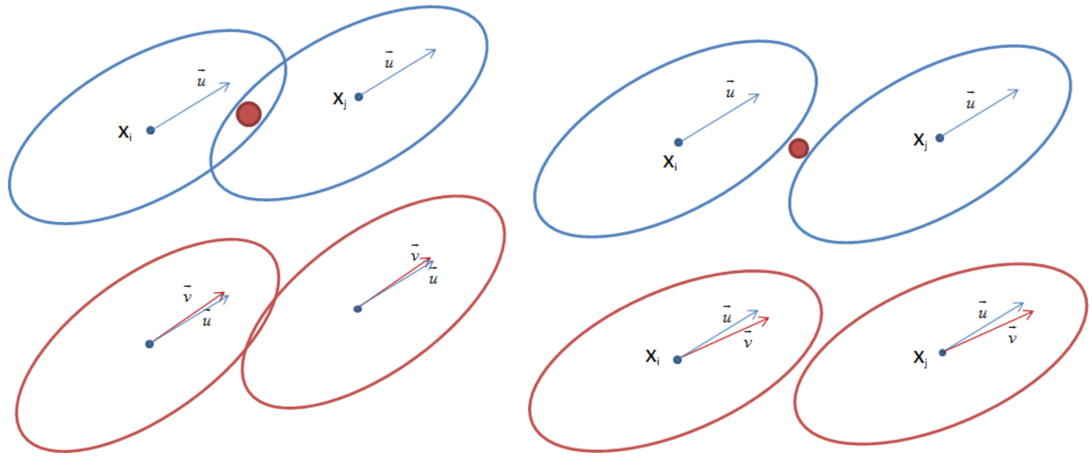

Figure 2: Choosing delta

Proof. It suffices to show this for a single pair of points and fixed $0 < c < 1$. The general case follows from the fact that a finite union of measure-zero sets has measure zero.

To this end, suppose that x_1 and x_2 are points in $\mathbb{R}^n$. Given a principal direction vector $\vec{a}$, we construct ellipses centered at x_1 and x_2 (respectively) $E_1(\vec{a})$ and $E_2(\vec{a})$ by taking all non-major axes with length equal to $c\|\vec{a}\|$.

Suppose that we would like $E_1(\vec{a})$ and $E_2(\vec{a})$ to be tangent. Then, the point of tangency t must occur at the midpoint of the segment from x_1 and x_2. So, we would like to determine all $\vec{a}$ for which the ellipse $E_1(\vec{a})$ contains the point t. Instead of fixing the vector from x_1 to t and rotating $\vec{a}$ to find such vertices, we imagine fixing $\vec{a}$ and rotating the space (including the vector from x_1 to t) and recovering $\vec{a}$ by taking an inverse rotation.

From this interpretation it is clear that we are interested in the set determined by the intersection of an $n-1$-sphere and the boundary of an n-dimensional ellipsoid. These meet in a disjoint union of spheres of dimension at most $n-2$. Since $\vec{a}$ and $-\vec{a}$ generate the same ellipse, the possibilities for $\vec{a}$ lie on a sphere of dimension $n-2$. This set has n-dimensional Lebesgue measure 0. □

Proposition 8. *"For every regular direction there is a small incremental rotation that yields an isomorphic persistence module."*

4 Estimating a direction of topological features

In this section we will explain how one can use the idea of directed filtrations to estimate a *direction* of a topological feature (hole, void, etc) similar to the idea of Principal Component Analysis in data analysis. Our presentation will focus on $\mathbb{R}^2$ but the theory can be generalized to higher dimensions. The problem that we will try to answer can be stated as follows:

4.1 Problem Statement

Problem Statement 1. Suppose that X is a set of points in $\mathbb{R}^2$ sampled with bounded error from a topological space K. Suppose, furthermore, that this topological space has a 1 dimensional "hole", i.e. that is separates $\mathbb{R}^2$ into two connected components a bounded one and unbounded one and that this hole has a direction in the sense that the bounded component defined above has a specific principal component vector $\vec{u}$. Computing that vector is a PCA optimization problem.

Since the "inside" of a topological hole is very hard to compute in general, performing PCA on it to compute the principal direction is inherently a hard problem. Our goal is to compute that vector $\vec{u}$ approximately by solving a simpler optimization problem. The function whose values we will try to optimize is that of maximum persistence, [18], described below:

Definition 9. Given a persistence diagram $Dgm(X, \vec{u_1}, I) = \{(a_j; b_j) : j \in J\}$ we define the *Maximum Persistence* as:

$$MP(Dgm(X, \vec{u_1}, I)) = \max_{(a,b) \in Dgm(X, \vec{u_1}, I)} \{|b - a|\}$$

We define an equivalence relation on the set $\mathcal{P}$ as follows:

$$\begin{aligned} Dgm(X, \vec{u_1}, I) &\sim Dgm(X, \vec{u_2}, I) \Leftrightarrow \\ MP(Dgm(X, \vec{u_1}, I)) &= MP(Dgm(X, \vec{u_2}, I)) \end{aligned}$$

Consider the set of equivalence classes of the persistence diagrams under this relation $\widetilde{\mathcal{P}_{MP}}$. The Maximum persistence function induces a total ordering on this set as follows:

$$\begin{aligned} Dgm(X, \vec{u_1}, I) &\leq_{MP} Dgm(X, \vec{u_2}, I) \Leftrightarrow \\ MP(Dgm(X, \vec{u_1}, I)) &\leq MP(Dgm(X, \vec{u_2}, I)) \end{aligned}$$

We are now ready to state our first relaxed version of the original problem.

Problem Statement 2. Suppose that X is a set of points in $\mathbb{R}^2$ sampled with bounded error from a topological space K. Suppose, furthermore, that this topological space has a 1 dimensional "hole", i.e. that is separates $\mathbb{R}^2$ into two connected components a bounded one and unbounded one and that this hole has a direction in the sense that the bounded component defined above has a specific principal component vector $\vec{u}$. If we assume that $\vec{u} \in S^1$ then each vector is completely determined by its angle with the x-axis θ and we write $\vec{u} = \vec{u}(\theta)$. We claim that this vector can be approximated by the solution of the optimization problem:

$$\theta' = \frac{\pi}{2} - argmin_{\theta \in [0,\pi)} \{MP(Dgm_1(X, \vec{u}(\theta), I)\}$$

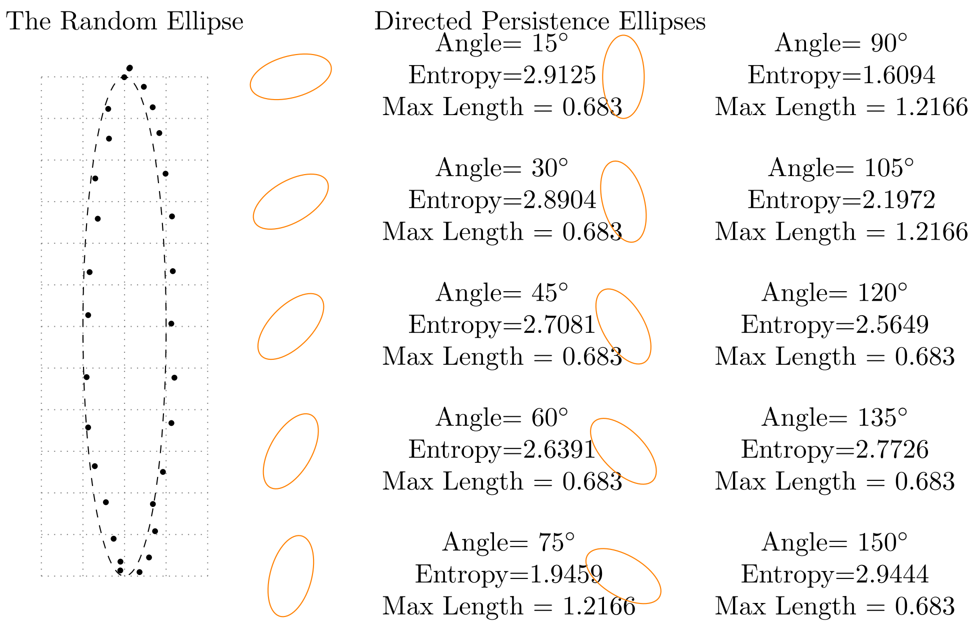

Figure 3: Using maximum persistence to determine the direction of a topological feature (hole).

A pictorial representation of the problem can be found in Figure 4. To validate our results we did the following experiments:

I'm going to replace this with a better version from Jordan.

A theoretical justification of our methods for the ideal case of rectangles follows.

Lemma 10. *Consider a rectangle, G in $\mathbb{R}^2$ such that the major axis is parallel to the x-axis, with side length M and the minor axis is parallel to the y-axis with side length N (so in particular $N < M$). Let $0 < c < 1$ be the ratio of the directed ellipses with angle $0° \leq \theta \leq 90°$. Then, the formula for the maximal persistence in terms of θ is given by:*

$$MP(\theta) = \begin{cases} \frac{N}{2c\cos\theta} & \text{if } 0 \leq \theta \leq \arctan(c) \\ \frac{N}{2\sin\theta} & \text{if } \arctan(c) \leq \theta \leq 90° \end{cases} \quad \text{(if } N < cM\text{)}$$

and

$$MP(\theta) = \begin{cases} \frac{M}{2\cos\theta} & \text{if } 0 \leq \theta \leq \arctan\left(\frac{N}{M}\right) \\ \frac{N}{2\sin\theta} & \text{if } \arctan\left(\frac{N}{M}\right) \leq \theta \leq 90° \end{cases} \quad \text{(if } N > cM\text{).}$$

Proof. □

Corollary 11. *For the rectangle described in the previous lemma*

Cases	*Max MP*	*Min MP*
$\dfrac{N}{M} < c$	$\theta = \arctan(c)$	$\theta = \dfrac{\pi}{2}$
$\dfrac{N}{M} > c$	$\theta = \arctan\left(\dfrac{N}{M}\right)$	$\theta = \dfrac{\pi}{2}$

Thus, independent of N, M, and c, the minimum of the maximum persistence is obtained at $\theta = \frac{\pi}{2}$.

4.2 Experimental Results

In order to validate our method we carried out the following experiments.

Experiment 1. In our first experiment we considered 20 ellipses with varying angles θ_0 on the plane. We sampled n points from those ellipses with uniform error of at most 2% of the "diameter" of the ellipse $\sqrt{a^2+b^2}$ where a, b are the semi-axis of the ellipse. We varied n for 100 to 200 with an incremental step of 10. We then use our optimization method to predict find θ_1, which is a prediction of θ_0 using the maximum persistence. Then we computed the difference from the original angle $Err_1 = \theta_0 - \theta_1$ and averaged over the 20 iterations. In figure $exp1$ we plotted the average errors with respect to the value of n for the MP optimization problem.

Report-comment on results,insert figure,make appropriate labels and connect things

Experiment 2. In our second experiment we again considered 20 ellipses with varying angles θ_0 on the plane. We sampled 100 points from those ellipses with uniform error ϵ% of the "diameter" of the ellipse $\sqrt{a^2+b^2}$. This time we let the error vary from 1% to 10% with increments of 1. We then use our optimization method to predict find θ_1, which is a prediction of θ_0 using the maximum persistence. Then we computed the difference from the original angle $Err_1 = \theta_0 - \theta_1$ and averaged over the 20 iterations. In figure $exp1$ we plotted the average errors with respect to the value of n for the MP optimization problem.

Report-comment on results,insert figure,make appropriate labels and connect things

5 Categorical Characterization

In this section we present two categorical views of directed persistence as generalized persistence modules. We also describe how to compute the interleaving distance between two directed persistence modules.

A persistence module is a a collection of vector spaces V_i and linear maps between them $V_i \to V_{i+1}$. When they were introduced by Zomorodian and Carlsson [22], the indexing of the vector spaces was taken to be the natural numbers. Later, Chazal, et al [5] extended this definition in order to allow persistence modules to be indexed by the real numbers. Following this, Bubenik, de Silva, and Scott [1] replaced the indexing set by any preordered category, defined below. We give their development of generalized persistence presently.

A *preordered set* is a pair $(P, \leq)$, where P is a set and $\leq$ is reflexive and transitive. We can view a preordered set P as a category $\mathbf{P}$ by taking the objects to be the elements of P

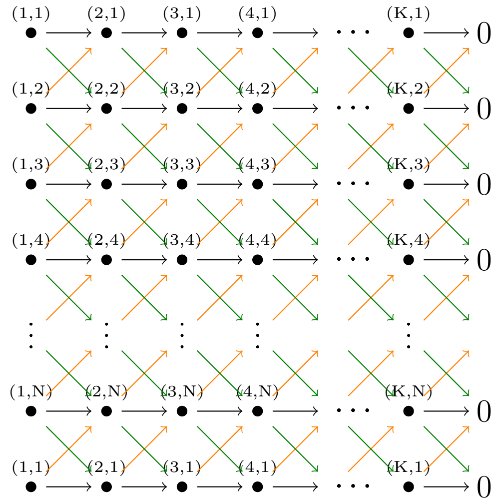

Figure 4: The category $\mathbf{C_{KN}}$

and by insisting that there be a unique morphism $s \to t$ whenever $s \leq t$ in P. A *generalized persistence module* is a functor $\mathbf{P} \to \mathbf{D}$ from a preordered category $\mathbf{P}$ to any category $\mathbf{D}$. Usually we will think of $\mathbf{D}$ as the category **Vect**.

When $\mathbf{P}$ is the category $\mathbf{n}$ of the first n natural numbers with its usual order, we recover the persistence modules of Zomorodian and Carlsson. When $\mathbf{P}$ is the category $\mathbf{R}$ of real numbers with its usual order, we recover the generalized persistence of Chazal et al.

Let N and K be positive integers. Put the set C_{KN} equal to $\{(i,j) \in \mathbb{N}^2 \mid i \leq K,\ j \leq N\}$. Define a preorder on C_{KN} by $(i,j) \leq (i'j')$ if there exist non-negative integers k and ℓ such that $i' = i + k + \ell$ and $j' = j \pm k \pmod N$. Figure 5 shows an illustration of the category $\mathbf{C_{KN}}$ corresponding to this preordered set.

Consider a finite set of features in $\mathbb{R}^2$. Fix some ratio $0 < c < 1$, a direction vector $\vec{u} \in \mathbb{R}P^1$, and a sequence $I = \{r_1 < r_1 < \cdots < r_K\}$. Let $N \in \mathbb{N}$ be so large that with $\theta = \frac{\pi}{N}$, there is an inclusion $E_{\vec{u}}(x, r_i) \subset E_{\vec{u}+\theta}(x, r_{i+1})$ for each $x \in X$ and each $i = 1, 2, \ldots, K$. Then, by functoriality, one obtains linear maps $H_*(K_{\vec{u}}(X, r_i)) \to H_*(K_{\vec{u}+\theta}(X, r_{i+1}))$, for each i. By symmetry, we also have inclusions $E_{\vec{u}}(x, r_i) \subset E_{\vec{u}-\theta}(x, r_{i+1})$ and linear maps $H_*(K_{\vec{u}}(X, r_i)) \to H_*(K_{\vec{u}-\theta}(X, r_{i+1}))$. Thus, we can associate to a collection X, $0 < c < 1$, I, and N the preordered category $\mathbf{C_{KN}}$ and define a functor to **Vect** by $(i,j) \mapsto H_*(K_{\vec{u}+(j-1)\theta}(X, r_i))$. This shows that directed persistence defines a generalized persistence module.

A *translation* on a preordered set $(P, \leq)$ is a function $\Gamma : P \to P$ that is monotone and satisfies $x \leq \Gamma(x)$ for all $x \in P$. A shift map induced by Γ is the composition $F(x) \to F\Gamma(x)$ where $F : \mathbf{P} \to \mathbf{D}$ is a functor. Let F and G be persistence modules, i.e. functors $\mathbf{P} \to \mathbf{D}$. A *natural transformation* $\varphi : F \implies G$ is a choice of $\mathbf{D}$-morphism $\phi_x : F(x) \to G(x)$ for each object x of $\mathbf{P}$. such that the following diagram commutes for every $\mathbf{P}$-morphism $\alpha : x \to y$.

Let Γ and K be transpositions on $\mathbf{P}$ and suppose that F and G are functors $\mathbf{P} \to \mathbf{D}$. Then, a (Γ, K)-*interleaving between* F *and* G consists of natural transformations $\varphi : F \implies G\Gamma$ and $\psi : G \implies FK$ such that $(\psi\Gamma)\phi = F\eta_{K\Gamma}$ and $(\varphi K)\psi = G\eta_{\Gamma K}$.

Proposition 12. *Let $X, c, \vec{u}$, and I be fixed as above. Suppose that θ is an angle $0 < \theta < \pi$ for which $E_{\vec{u}}(x, r_i) \subset E_{\vec{u}+\theta}(x, r_{i+1})$ for each $x \in X$ and $i \in \{1, 2, \dots, K\}$. Then, the persistence diagrams $H_*(K_{\vec{u}})$ and $H_*(K_{\vec{u}+\theta})$ are (Γ, Γ)-interleaved, where $\Gamma : n \mapsto n + 1$.*

Let $C = \mathbf{R}P^1 \times \overline{\mathbb{R}}_+$ be the closed, half-infinite cylinder. Define a preorder on C by $(\vec{u}_1, x_1) \leq (\vec{u}_2, x_2)$ if and only if $x_1 < x_2$ or $x_1 = x_2$ and $\vec{u}_1 = \vec{u}_2$. This relation is reflexive and transitive, so it defines a preorder.

Fix a set of features X and a ratio c of major-axis to minor-axis length. Let $\vec{u}$ and $\vec{v}$ be regular directions for X and fix a sequence $I = r_1 < r_2 < \cdots < r_K$. Define a translation $\Gamma : C \to C$ by $\Gamma(\vec{u}, r) \mapsto (\vec{v}, r')$, where $r' \in I$ is minimal with respect to the property that $E_{\vec{u}}(X, r) \subset E_{\vec{v}}(X, r')$. Next, define $\omega : \mathbf{Trans_C} \to [0, \infty]$ by $\omega_\Gamma = \max\{r' - r \mid \Gamma(\vec{u}, r) = (\vec{v}, r')\}$. Not sure we need a max. Does a single value work for all values of r?

Clearly $\omega_I = 0$, where I is the identity translation and if Γ_1 and Γ_2 are translations, then $\omega_{\Gamma_1\Gamma_2} \leq \omega_{\Gamma_1} + \omega_{\Gamma_2}$. Thus, ω defines a sublinear projection and we can define the interleaving distance between two directed persistence modules E_1 and E_2 as the infimum of ω, where Γ is a translation from E_1 to E_2 and $\omega_\Gamma \leq \omega$.

More generally, we can view directed persistence as a functor from $\mathbb{R}$ and define the interleaving distance as above, where $\Gamma(\vec{u}, r) = (\vec{v}, r')$ where r' is minimal with respect to $E_{\vec{u}}(X, r) \subset E_{\vec{v}}(X, r')$.

Lesnick [14] shows that "in the case of 1-D persistence, the interleaving distance is equal to the bottleneck distance on tame persistence modules." But, I don't know what tame means.

None of the above depends on the fact that our ambient space is $\mathbb{R}^2$. Indeed, in $\mathbb{R}^d$, we consider the set C of pairs $(\vec{u}, x) \in \mathbb{R}P^{d-1} \times [0, \infty)$. Define a preorder on C by $(\vec{u}, x) \leq \vec{v}, y)$ if and only if $x < y$ or $x = y$ and $\vec{u} = \vec{v}$.

Not sure if we still want to pursue this line of thinking or not.

There are various ways one could define and analyze a generalized directed persistence diagram. One way is to treat it as a multidimensional persistent diagram where one parameter is the scalar increment and the other is the angle. One then gets the diagram shown in Figure 6:

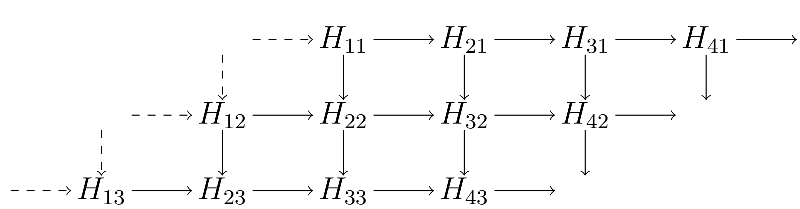

Figure 5: Directed filtrations as multidimensional persistent diagrams

One could then proceed to "complete" this diagram using "pull-backs" and "push-outs" as described in [19]. In the same paper it is proven that this completion method terminates in finite steps and one can now have a classic multidimensional persistence diagram in the sense of [4]. One could use similar analysis to [14] on this multidimensional persistence module.

Another approach is using the categorical formulations of generalized persistence diagrams found in [1,5]. A small description (copy-paste) from Peter's paper is needed here

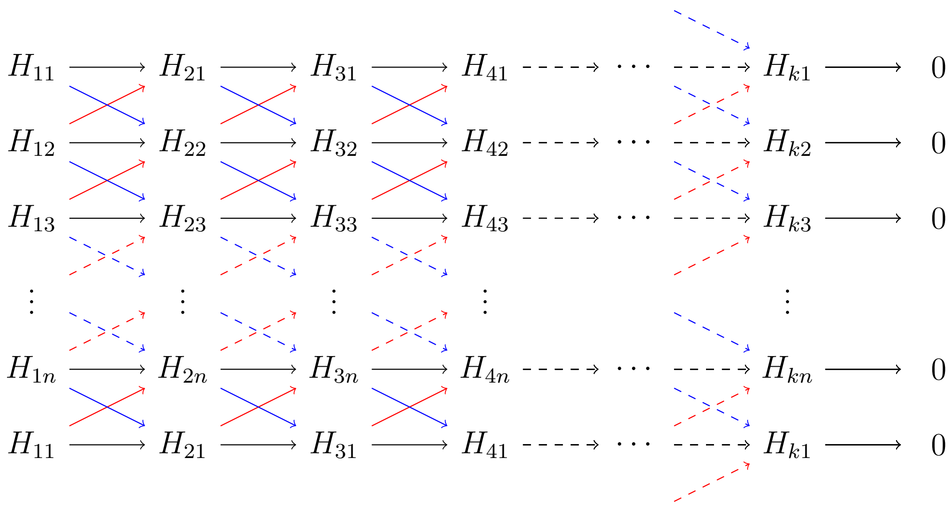

Figure 6: The generalized persistence module for directed filtrations of a set $X \in \mathbb{R}^2$

Definition 13. The degree p Wasserstein distance between two persistence diagrams $\mathrm{Dgm}(X)$ and $\mathrm{Dgm}(Y)$ is defined as:

$$d_{W_p}(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)) = \inf_{\gamma} \left(\sum_x \|x - \gamma(x)\|_\infty^p \right)^{\frac{1}{p}}$$

where the infimum is over all bijections $\gamma : \mathrm{Dgm}(X) \to \mathrm{Dgm}(Y)$ and the sum is over all points in X. We will content ourselves with the degree 1 Wasserstein distance which we will denote with d_W.

Definition 14. The bottleneck distance between two persistence diagrams $\mathrm{Dgm}(X)$ and $\mathrm{Dgm}(Y)$ is defined as:

$$d_B(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)) = \inf_{\gamma,\ x} \|x - \gamma(x)\|_\infty$$

where the infimum is over all bijections $\gamma : \mathrm{Dgm}(X) \to \mathrm{Dgm}(Y)$ and over all points in X.

Definition 15. Let X be a point cloud in $\mathbb{R}^n$, $I = \{r_0 < r_1 < \cdots < r_k\}$ and $\mathbf{c}$ as before. Define the set of all directed persistence diagrams $\mathcal{P} = \{\text{Dgm}(X, \vec{v}, I), \forall \vec{v} \in \mathbb{R}P^n\}$. This is a metric space with the Wasserstein distance or the Bottleneck distance.

For more information about the space of persistence diagrams, see [6, 15].

Definition 16. Suppose that

$$\mathcal{K} = \{K_1 \subseteq K_2 \subseteq \cdots \subseteq K_k\}$$

and

$$\mathcal{M} = \{M_1 \subseteq M_2 \subseteq \cdots \subseteq \vec{M}_k\}$$

are two filtrations over a finite set $X \in \mathbb{R}^n$ indexed by $I = \{r_i := i \cdot \epsilon\,,\ i = 1, 2, ...k\}$. We say that the two sequences are ϵ interleaving if there exist inclusion maps: $\alpha : K_i \to M_{i+1}$ and $\beta : M_i \to K_{i+1}$, for $1 \leq i \leq k-1$.

The proof for the following propositions can be found in [5].

Proposition 17. *Suppose that $\mathcal{K}$ and $\mathcal{M}$ are two ϵ-interleaving filtrations on a finite set $X \in \mathbb{R}^n$ and $\text{Dgm}(\mathcal{K})$,$\text{Dgm}(\mathcal{M})$ are the corresponding diagrams of these two filtrations. Then*

$$d_B(\text{Dgm}(\mathcal{K}), \text{Dgm}(\mathcal{M})) < 2\epsilon$$

Proposition 18. *Suppose that $\mathcal{K}$ and $\mathcal{M}$ are two ϵ-interleaving filtrations on a finite set $X \in \mathbb{R}^n$ and $\text{Dgm}(\mathcal{K})$,$\text{Dgm}(\mathcal{M})$ are the corresponding barcodes of these two filtrations. Then*

$$d_W(\text{Dgm}(\mathcal{K}), \text{Dgm}(\mathcal{M})) < 2\epsilon$$

Lemma 19. *Let $x_0 \in \mathbb{R}^2$, $\epsilon > 0$ and $0 < c < 1$. Set $I = \{r_i := i \cdot \epsilon\ :\ i = 1, 2, ...k\}$ Let $\vec{e_1}$ denote the first standard basis vector $\langle 1, 0\rangle$. Consider*

$$f_{\vec{e_1}}^{-1}(0) \subseteq f_{\vec{e_1}}^{-1}(\epsilon) \subseteq f_{\vec{e_1}}^{-1}(2\epsilon) \subseteq \cdots \subseteq f_{\vec{e_1}}^{-1}(k\epsilon),$$

defined above. Then, there exists some $\delta > 0$ so that for any $\vec{u} \in S^1$ with $d(\vec{e_1}, \vec{u}) < \delta$, then we have

$$f_{\vec{e_1}}^{-1}(i \cdot \epsilon) \subseteq N_\epsilon(f_{\vec{u}}^{-1}((i+1)\epsilon))$$

and

$$f_{\vec{u}}^{-1}(i \cdot \epsilon) \subseteq N_\epsilon(f_{\vec{e_1}}^{-1}((i+1)\epsilon)),$$

for all $i = 0, 1, 2, \ldots, k-1$.

Corollary 20. *If $\mathcal{K}(\vec{e_1})$ and $\mathcal{K}(\vec{u})$ are the persistence modules corresponding to the filtrations obtained from $f_{\vec{e_1}}^{-1}(\cdot)$ and $f_{\vec{u}}^{-1}(\cdot)$, respectively. Suppose that $\vec{u}$ has the property that $d(\vec{e_1}, \vec{u}) < \delta$, where δ is chosen as in the previous theorem. Then, $K(\vec{e_1})$ and $K(\vec{u})$ are ϵ-interleaved.*

This proof follows from the proof in [5].

Computationally interesting in the following sense: if you want to find a principal direction of a feature up to a tolerance of ϵ, the lemma allows you to find an increment in the direction (the δ) that should be chosen to find the directed feature to ϵ tolerance.

6 Conclusions

References

[1] Peter Bubenik, Vin de Silva, and Jonathan Scott, *Metrics for generalized persistence modules*, arXiv preprint arXiv:1312.3829 (2013).

[2] Gunnar Carlsson, *Topology and data*, Bulletin of the American Mathematical Society **46** (2009), no. 2, 255–308.

[3] Gunnar Carlsson and Vin de Silva, *Zigzag persistence*, Found. Comput. Math. **10** (2010), no. 4, 367–405. MR2657946 (2012b:68329)

[4] Gunnar Carlsson and Afra Zomorodian, *The theory of multidimensional persistence*, Discrete Comput. Geom. **42** (May 2009), no. 1, 71–93.

[5] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot, *The structure and stability of persistence modules*, CoRR **abs/1207.3674** (2012).

[6] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer, *Stability of persistence diagrams*, Discrete Comput. Geom. **37** (2007), no. 1, 103–120. MR2279866 (2008i:68130)

[7] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov, *Vines and vineyards by updating persistence in linear time*, Proceedings of the twenty-second annual symposium on computational geometry, 2006, pp. 119–126.

[8] James F. Davis and Paul Kirk, *Lecture notes in algebraic topology*, Graduate Studies in Mathematics, vol. 35, American Mathematical Society, Providence, RI, 2001. MR1841974

[9] Herbert Edelsbrunner and John Harer, *Persistent homology-a survey*, Contemporary mathematics **453** (2008), 257–282.

[10] Herbert Edelsbrunner and John L. Harer, *Computational topology*, American Mathematical Society, Providence, RI, 2010. An introduction. MR2572029 (2011e:00001)

[11] Saba Emrani, Thanos Gentimis, and Hamid Krim, *Persistent homology of delay embeddings and its application to wheeze detection*, IEEE Signal Processing Letters **21** (2014), 459–463.

[12] Robert Ghrist, *Barcodes: the persistent topology of data*, Bull. Amer. Math. Soc. (N.S.) **45** (2008), no. 1, 61–75. MR2358377 (2008i:55007)

[13] Allen Hatcher, *Algebraic topology*, Cambridge University Press, Cambridge, 2002. MR1867354 (2002k:55001)

[14] Michael Lesnick, *The theory of the interleaving distance on multidimensional persistence modules*, arXiv preprint arXiv:1106.5305 (2011).

[15] Yuriy Mileyko, Sayan Mukherjee, and John Harer, *Probability measures on the space of persistence diagrams*, Inverse Problems **27** (2011), no. 12, 124007.

[16] Konstantin Mischaikow and Vidit Nanda, *Morse theory for filtrations and efficient computation of persistent homology*, Discrete and Computational Geometry **50** (2013September), 330–353.

[17] Partha Niyogi, Stephen Smale, and Shmuel Weinberger, *Finding the homology of submanifolds with high confidence from random samples*, Discrete Comput. Geom. **39** (March 2008), no. 1, 419–441.

[18] Jose Perea and John Harer, *Sliding windows and persistence: An application of topological methods to signal analysis*, arXiv preprint arXiv:1307.6188 (2013).

[19] Primoz Skraba and Mikael Vejdemo-Johansson, *Persistence modules: Algebra and algorithms*, arXiv preprint arXiv:1302.2015 (2013).

[20] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams, *Javaplex: A research software package for persistent (co)homology*, 2011.

[21] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), no. 5500, 2319.

[22] Afra Zomorodian and Gunnar Carlsson, *Computing persistent homology*, Discrete Comput. Geom. **33** (2005), no. 2, 249–274.