Topological Data Analysis

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Rajdeep Haldar

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2020

Supervisor: Prof. Sourish Das

Certificate

This is to certify that this dissertation entitled Topological Data Analysis towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Rajdeep Haldar at Indian Institute of Science Education and Research under the supervision of Prof. Sourish Das, Associate Professor, Chennai Mathematical Institute , Department of Mathematics, during the academic year 2019-2020.

Sourish Das

Prof. Sourish Das

Committee:

Prof. Sourish Das

Prof. Anindya Goswami

This thesis is dedicated to my parents, sister and Jessebel Vistro for nursing me with their constant love, support and encouragement.

Declaration

I hereby declare that the matter embodied in the report entitled Topological Data Analysis are the results of the work carried out by me at the Department of Mathematics, Indian Institute of Science Education and Research, Pune, under the supervision of Prof. Sourish Das and the same has not been submitted elsewhere for any other degree.

Rajdeep Haldar

Rajdeep Haldar

Acknowledgments

I would like to thank my supervisor, Prof. Sourish Das, for guiding me through the course of this project. I am especially grateful to Prof. Priyavrat Deshpande for sharing his expertise and having invaluable discussions with me.

I would like to thank Prof. Uttara Naik Nimbalkar who helped me build the foundations I needed to commence this project.

I would also like to acknowledge Prof. Anirban Chakraborti for the collaborative work with financial data and mention Prof. Arvind Rao for sharing his medical data and expertise. The continuous cooperation of the previously mentioned made interdisciplinary applications of TDA described in this thesis come to fruition.

Finally, with all my heart and soul I would like to thank Jaideep Mahajan, Vishnu N. and all my friends at IISER Pune for the constant moral support.

Abstract

This thesis is a mathematical exposition of the theory behind Topological Data Analysis (TDA) complemented by two applications in medicine and financial realm. We start by establishing the foundation of homology theory, then study the reconstruction of the underlying manifold from point cloud data. Followed by the theory of persistent homology which provides a topological summary of the significant geometrical features of the data. We study its diagram representations, robustness and characterisation via persistence modules. Subsequently, we study persistence landscapes and extend statistical concepts of confidence intervals, convergence and hypothesis testing for topological summaries of the data. Furthermore, we discuss the mapper algorithm, which provides network representations for high dimensional data. Finally we end the thesis with a brief discussion on the interdisciplinary application of TDA implemented in this project.

Contents

Introduction

Topological Data Analysis (TDA) aims to extract underlying geometrical features of the inherent space from which the data has been sampled from. It marries theoretical frameworks of algebraic topology and statistical analysis to draw geometrical inference from the sampled data. [11]

The most popular concepts in TDA are *Persistent Homology* and *Mapper Algorithm*. [12] Persistent Homology provides a topological summary of the data, highlighting the significant homological features of the underlying space, from which the data has been sampled.

The topological summaries obtained from *Persistent Homology* are robust i.e less susceptible to noise and can be given function representations (*Landscapes*). This kind of representation enables us to perform statistics on these topological summaries to define confidence intervals and statistical tests. In turn providing us the ability to address questions like whether two samples are sampled from the same space, based on the underlying geometry?

While *Mapper Algorithm* gives a network visualisation of data represented in arbitrary dimension. The *Mapper Algorithm* is superior to other visualisation techniques such as PCA (Principal Component Analysis), MDS (Multi Dimensional Scaling) etc. as *Mapper* retains high dimensional geometrical structures. These high dimensional structures are lost in the previously mentioned techniques as they rely on embedding/projecting the data to a lower dimensional space.

Data has shape, and TDA acknowledges this fact to extract additional information disregarded by standard statistical techniques. TDA methods can be used as exploratory data analysis tools or to generate additional input features for statistical/machine learning models. Due to the previously mentioned merits, TDA serves as a topic of great interest.

In this thesis we establish the mathematics behind TDA. We focus on Persistence Homology and Mapper Algorithm. We also apply these concepts to medical and financial data to develop novel insights.

Chapter 1

Homology Theory

Homology is a mathematical concept of characterising n dimensional holes in a topological space. The $0, 1, 2$ dimensional holes represent the number of connected components, loops and voids respectively. These homological properties are invariant under continuous deformation (*homotopy*) of a topological space and give us a characterisation which is useful to capture the geometric properties of the underlying space from which the point cloud data has been sampled.

We begin the chapter by establishing the theory for combinatorial spaces called simplicial complexes and towards the end will generalise it for any topological space. The contents of this chapter are based on [1].

1.1 Simplicial Homology

Definition 1.1.1 (Affine Independence)**.** *Let $v_0, \ldots, v_p \in \mathbb{R}^p$ such that $\sum_{i=0}^{p} t_i v_i = 0 \iff t_i = 0 \; \forall \; i$ then $v_0, \ldots, v_p$ are said to be affinely independent.*

Definition 1.1.2 (p - simplex, Δ_p)**.** *A p- simplex σ is the convex hull of $p+1$ affinely independent points $v_0, \ldots, v_p \in \mathbb{R}^p$. It is denoted by $\sigma = [v_0, \ldots, v_p]$. Any k- simplex induced by a proper subset of $\{v_0, \ldots, v_p\}$ with cardinality $|k|$ is a face of σ.*

According to the above definition a 0-simplex is a single point with no faces; a 1-simplex is an edge with the endpoints as its faces; a 2-simplex is a filled triangle with its edges and vertices as faces and so on.

Definition 1.1.3 (Geometrical simplicial complex)**.** *A geometrical simplicial complex K in $\mathbb{R}^d$ is a collection of simplices such that:*

1. *Any face of a simplex in K is also a simplex in K. [Hereditary property]*

2. *Intersection of any two simplices in K is either empty or a common face of both.*

Simplices should be intuitively viewed as generalisation of triangles of arbitrary dimensions and simplicial complexes should be viewed as spaces built by gluing these generalised triangles together only at their faces.

Remark 1.1.1. *The notion of geometric simplices and simplicial complexes can be generalised to abstract simplices and simplicial complexes where each affinely independent point can be replaced by an abstract mathematical object (set representation).*

Once we have a simplicial complex we can look at the algebraic space spanned by the constituent simplices as basis, we will these spaces chains.

Definition 1.1.4 (p-chain)**.** *Let K be a simplicial complex and p be its dimension. A p-chain (C_p) is the formal sum of p-simplices in K.*

$$C_p = \{c : c = \sum a_i \sigma_i\}$$

$a_i \in \mathrm{R}$ where R is any ring, $\sigma_i \in$ (set of p-simplices in K)

Suppose we have a 1- simplex $[ab]$ (line segment $\bar{ab}$), we would like to have a notion of boundary which would output the 0-simplices $[a], [b]$ (vertices a, b). An operator which takes elements in C_p and outputs its boundary elements which would be one dimension less i.e in C_{p-1}.

Definition 1.1.5 (Boundary maps ∂)**.** *$\partial_p : C_p \to C_{p-1}$ is a linear map such that for all $\sigma = [v_0, \ldots, v_p]$*

$$\partial_p(\sigma) = (-1)^j \sum_{j=0}^{p} [v_0, \ldots, \hat{v_j}, \ldots v_p]$$

where $[v_0, \ldots, \hat{v_j}, \ldots v_p] = [v_0, \ldots, v_{j-1}, v_j, \ldots v_p]$

Lemma 1.1.1. *$\partial_p \circ \partial_{p+1}(d) = 0$ for all $d \in C_{p+1}$*

The lemma implies that the image of $p+1$ boundary map sits inside the kernel of the p^{th} boundary map. This sort of structure is called a *chain complex* represented as the following quiver diagram:

$$\begin{tikzcd} 0 \arrow[r] & \dots \arrow[r, "\partial_{p+2}"] & C_{p+1} \arrow[r, "\partial_{p+1}"] & C_p \arrow[r, "\partial_p"] & C_{p-1} \arrow[r, "\partial_{p-1}"] & \dots \arrow[r, "\partial_0"] & 0 \end{tikzcd}$$

Definition 1.1.6 (p -Cycles)**.** *Collection $Z_p \subset C_p$ such that $Z_p = Ker(\partial_p)$.*

These are objects which have zero boundary (no boundary).

Definition 1.1.7 (p -Boundaries)**.** *Collection $B_p \subset C_p$ such that $B_p = Img(\partial_{p+1})$.*

These are objects which are part of boundary of some one higher dimension object.
In the introduction of this chapter, we said homology characterises holes in arbitrary dimension, intuitively holes are objects which have no boundary with empty interior. In other words, holes are objects which have no boundary and also aren't boundary of some one higher dimension object. This motivates us to define the p^{th} homology as follows:

Definition 1.1.8 (p^{th}-Homology group)**.** *The p^{th}-Homology group H_p is the p^{th} cycle group modulo p^{th} boundary group.*

$$H_p = {}^{Z_p}/_{B_p}$$

p^{th} betti number $\beta_p = rank(H_p)$

Continuing with our intuition, which now has been formalised; β_0 represents number of connected components, β_1 represents number of loops, β_2 represents number of voids and so on.

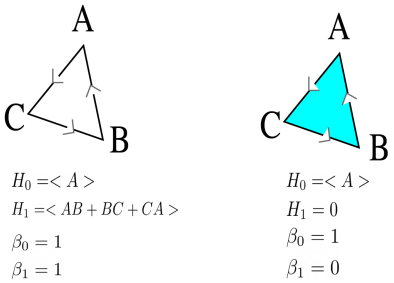

Figure 1.1: The left simplicial complex has a loop generated by the individual edges, represented as a generator in H_1 but it disappears when the simplex $[ABC]$ is introduced in the simplicial complex in the right, as it is a boundary of $[ABC]$.

1.2 Singular Homology

Now that we have established homology theory for simplicial complexes we would like to generalize the computation to any topological space.

Definition 1.2.1 (singular p -simplex)**.** *A singular p -simplex σ in a topological space X is a continuous function from the standard p -simplex to the topological space X.*

$$\sigma : \Delta_p \to X$$

The singular p simplex σ also induces singular $p-1$ simplices by restriction of the domain of σ to the $p-1$ -faces of the standard p simplex.
Let $\Delta_p \sim [v_0, \ldots, v_p]$, define restriction maps $i^j_{p-1} : \Delta_{p-1} \to \Delta_p$ such that $i^j_{p-1}([v_0, \ldots, \hat{v_j}, \ldots v_p]) = [v_0, \ldots, \hat{v_j}, \ldots v_p]$. The singular $p-1$ simplex corresponding to the face $[v_0, \ldots, \hat{v_j}, \ldots v_p]$ will be $\sigma \circ i^j_{p-1}$. Inductively we can define all the singular faces of σ.
Now that we have singular simplices we can use them as building blocks to build our singular complexes, motivated by the definition of simplicial complexes we similarly define singular complexes.

Definition 1.2.2 (Singular complex)**.** *A singular complex $K(X)$ on a topological space X is a collection of singular simplices such that:*

1. *Any singular face of a simplex in $K(X)$ is also a singular simplex in K. [Hereditary property]*

2. *Any two singular simplices having a common face in domain should agree as maps when restricted to those faces.*

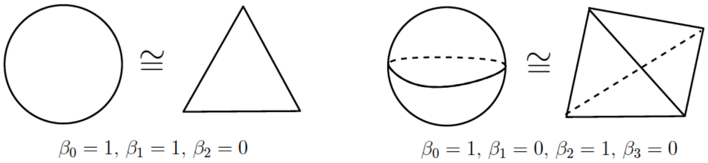
≅
≅
$\beta_0 = 1,\ \beta_1 = 1,\ \beta_2 = 0$
$\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1,\ \beta_3 = 0$

Figure 1.2: Left: Singular homology for S^1; Right: Singular Homology for S^2. Singular homology enables us to compute homology of any general topological space X by calculating the simplicial homology on the inverse image of the singular complex built on that space while constrained by the singular maps.

In the same way we can generalize the concept of chains, cycles, boundaries and finally homology. The only tweak here is the boundary map.

Definition 1.2.3 (singular Boundary maps ∂)**.** $\partial_p : C_p(X) \to C_{p-1}(X)$ *is a linear map such that for all singular* p *-simplices* σ

$$\partial_p(\sigma) = (-1)^j \sum_{j=0}^{p} \sigma \circ \imath_{p-1}^{j}$$

Using this as the boundary map the rest of theory is consistent with simplicial homology and thus we have established the theory of homology for any topological space X.

1.3 Homotopy Invariance

At the beginning of the chapter we said that these homological properties are invariant under continuous deformation and hence allow us to give some kind of characterisation, this invariance is the basis of all topological data analysis. In this section we will formalise it.

Definition 1.3.1 (Homotopic Maps)**.** *Two functions* $f, g : X \to Y$ *are said to be homotopic if* $\exists$ *a continuous function* $H : X \times [0,1] \to Y \ni H(X,0) = f$ *and* $H(X,1) = g$*. We denote* f *and* g *are homotopic as* $f \sim g$*.*

Definition 1.3.2 (Homotopy Equivalence)**.** *Two topological spaces* X *and* Y *are said to be homotopicaly equivalent if* $\exists f : X \to Y$ *and* $g : Y \to X$ *such that* $f \circ g \sim \mathbb{I}_X$ *and* $g \circ f \sim \mathbb{I}_Y$*. We represent homotopicaly equivalent as* $X \overset{h}{\sim} Y$*.*

The above is the mathematical definition for two space being equivalent in terms of continuous deformation. Suppose we have a continuous map $f : X \to Y$ this induces a map between the singular simplices of X to singular simplices of Y ($f_{\#}(\sigma) = f \circ \sigma : \Delta_p \to Y$ where σ is a singular p simplex on X). As this is true for any singular simplex, in turn we have an induced homomorphism $f_{\#} : C_p(X) \to C_p(Y)$ for all p. This can be represented as the following quiver representation:

$$\begin{tikzcd}
0 \arrow[r] & \dots \arrow[r, "\partial_{p+2}"] \arrow[d, "f_{\#}"] & C_{p+1}(X)\dots \arrow[r, "\partial_{p+1}"] \arrow[d, "f_{\#}"] & C_p(X)\dots \arrow[r, "\partial_p"] \arrow[d, "f_{\#}"] & C_{p-1}(X)\dots \arrow[r, "\partial_{p-1}"] \arrow[d, "f_{\#}"] & \dots\dots \arrow[r, "\partial_0"] \arrow[d, "f_{\#}"] & 0 \\
0 \arrow[r] & \dots \arrow[r, "\partial_{p+2}"] & C_{p+1}(Y) \arrow[r, "\partial_{p+1}"] & C_p(Y) \arrow[r, "\partial_p"] & C_{p-1}(Y) \arrow[r, "\partial_{p-1}"] & \dots \arrow[r, "\partial_0"] & 0
\end{tikzcd}$$

It is easy to see that $\partial \circ f_{\#} = f_{\#} \circ \partial$ hence the diagram commutes. Hence $f_{\#}$ takes cycles to cycles and boundaries to boundaries. Hence there is a natural homomorphism $f_* : H_p(X) \to H_p(Y)$. This argument can be extended for compositions and hence if we have $f : X \to Y$ and $g : Y \to Z$ then $g_* \circ f_* = (g \circ f)_*$.

Theorem 1.3.1. *If $f, \tilde{f} : X \to Y$ are homotopic then $f_* = \tilde{f}_*$*

So homotopic maps induce the same homomorphisms between homology groups. The immediate corollary by using the above theorem and the composition property is the following.

Corollary 1.3.2. *If $X \overset{h}{\sim} Y$ then $H_p(X) \cong H_p(Y)$ for all p.*

The betti numbers β_p for all p, give a topological description of a given space. In order to draw topological inference for a space X from the betti numbers, it suffices to compute the homology of a homotopically equivalent space to X. In the next chapter using the data point cloud we estimate the underlying manifold upto homotopy. Hence computing the homology of this reconstruction helps us draw topological inference about the original manifold.

Chapter 2

Topological Inference

Let $\mathbb{X}_n$ be a point cloud consisting of n points in $\mathbb{R}^d$. The underlying assumption of the TDA pipeline is that there is a manifold $\mathcal{M}$ from which the point cloud $\mathbb{X}_n$ has been sampled. Our aim in this chapter will be to estimate the manifold $\mathcal{M}$ upto homotopy and extract its homological information via the estimated topological space as homology is an homotopic invariant. The main results of this chapter are based on the paper [6].

2.1 Reconstruction and Nerve Theorem

Suppose we have a compact subset $K \subset \mathbb{R}^d$, we can define a distance function $d_K : \mathbb{R}^d \to \mathbb{R}$ such that $d_K(x) = \inf_{y \in K} \|x - y\|$.

Definition 2.1.1 (r-offset)**.** *The r-offset K^r of a compact subset $K \subset \mathbb{R}^d$ is defined as the r-sublevel set of the distance function d_K. $K^r = d_K^{-1}[0, r]$.*

The r-offset of K is basically the union of r - radius balls around each point in K.
In order to estimate the underlying manifold $\mathcal{M}$ from the point cloud $\mathbb{X}_n$, we would require a notion of distance between these two spaces, motivating the following definition.

Definition 2.1.2 (Hausdorff Distance D_H)**.** *Given two compact subsets $K, K' \subset \mathbb{R}^d$ the hausdorff distance $D_H(K, K') = \sup_{x \in \mathbb{R}^d} |d_k(x) - d_{k'}(x)|$*

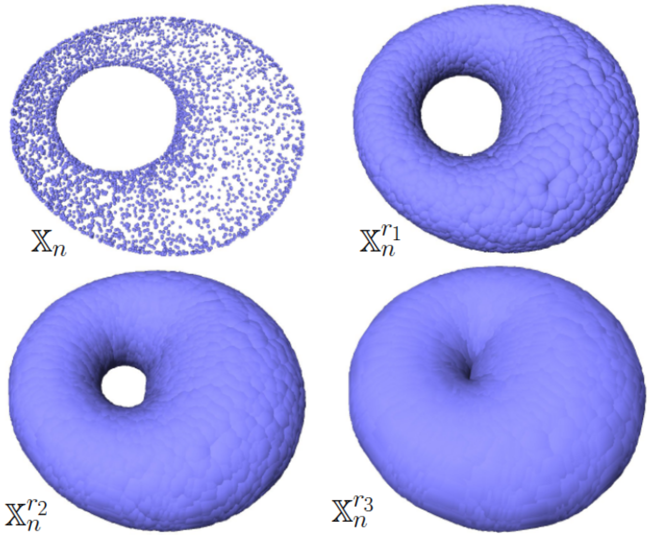

Figure 2.1: r-offsets $\mathbb{X}_n^r$ of a point cloud $\mathbb{X}_n$ sampled from a torus for offset radii $r_1 < r_2 < r_3$. Note that for a certain range of radii the r-offsets will be homotopically equivalent to the underlying torus. **Credits:** [6]

Intuitively when we vary r and look at the r-offsets $\mathbb{X}_n^r$ around the point cloud we hit a sweet spot of radii r such that $\mathbb{X}_n^r$ is homotopicaly equivalent to the underlying manifold $\mathcal{M}$ (fig. 2.1). Turns out the differential properties of the distance function $d_{\mathcal{M}}$ and the hausdorff distance between $\mathbb{X}_n$ and $\mathcal{M}$ are enough to specify this sweet spot.

Definition 2.1.3 (α -critical and α -reach)**.** *Given a distance function d_K for a compact subset $K \in \mathbb{R}^d$ a point $x \in \mathbb{R}^d$ is called α -critical if $\|\nabla d_K(x)\| \leq \alpha$.*
The α reach for d_K: $reach_\alpha(d_K)$ is the maximum r for which there is no α - critical point in $d_k^{-1}(0, r]$.

Theorem 2.1.1 (Reconstruction Theorem)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact set such that $reach_\alpha(d_{\mathcal{M}}) \geq R > 0, \alpha \in (0, 1)$ and $\mathbb{X}_n$ be the point cloud such that $D_H(\mathcal{M}, \mathbb{X}_n) = \epsilon < \frac{R}{5+4/\alpha^2}$. Then for $r \in [4\epsilon/\alpha^2, R - 3\epsilon]$ and $\eta \in (0, R)$, $\mathcal{M}^\eta \overset{h}{\sim} \mathbb{X}_n^r$*

The reconstruction theorem gives us the sweet spot for r for which the r-offsets of the point cloud are homotopicaly equivalent to the underlying manifold. So the homology of the r-offset will capture the homology of the manifold the point cloud has been sampled from. The next theorem will help us relate r offsets to simplicial complexes. This gives us a pragmatic gateway to compute homology of the underlying space. As simplicial complexes are combinatorial spaces they are desired space to work with when dealing with computations.

Definition 2.1.4 (Nerve of a cover). *Given an open cover $\mathcal{U} = (U_i)_{i\in I}$ of a topological space X . The nerve of $\mathcal{U}$ is the abstract simplicial complex $\mathcal{N}(\mathcal{U})$ whose vertices are the U_i's such that $\sigma = [U_{i0}, \ldots, U_{ik}] \in \mathcal{N}(\mathcal{U})$ iff $\cap_{j=0}^{k} U_{ij} \neq \phi$.*

Theorem 2.1.2 (Nerve Theorem). *Let $\mathcal{U} = (U_i)_{i\in I}$ be an open cover of a topological space X by open sets such that intersection of any sub-collection of U_i's is either empty or contractible. Then X and the nerve $\mathcal{N}(\mathcal{U})$ are homotopicaly equivalent.*

The nerve theorem along with the reconstruction theorem implies that the nerve of certain r -offsets of the point cloud $\mathcal{N}(\mathbb{X}_n^r)$ are homotopicaly equivalent to the underlying manifold $\mathcal{M}$. So calculating the simplicial homology of this nerve is sufficient to extract the homology of the underlying manifold. The caveat here though is getting the optimal r and satisfying the regularity assumptions of reconstruction theorem. So, instead of choosing a particular r we can vary the offset radius r and look at the variation of homology of the r- offsets by computing the homologies of the nerves (simplicial complexes) of these offsets. The intuition is homological features which are significant should appear persistently when analysing the variation of homology. This is the motivation behind the theory of persistence homology which we will establish in the next chapter. It suffices to work with these simplicial complexes henceforth we shall only deal with simplicial complexes to study the underlying homological properties.

2.2 Building simplicial complexes from the point cloud

In this section we will restrict our focus two types of simplicial complexes built from the point cloud. The first one is the Čech complex, which is the direct consequence of the idea we were establishing at the end of previous section. The second one is the Rips complex which is superior to Čech complex in terms of computational efficiency while still being closely related to the Čech complex.

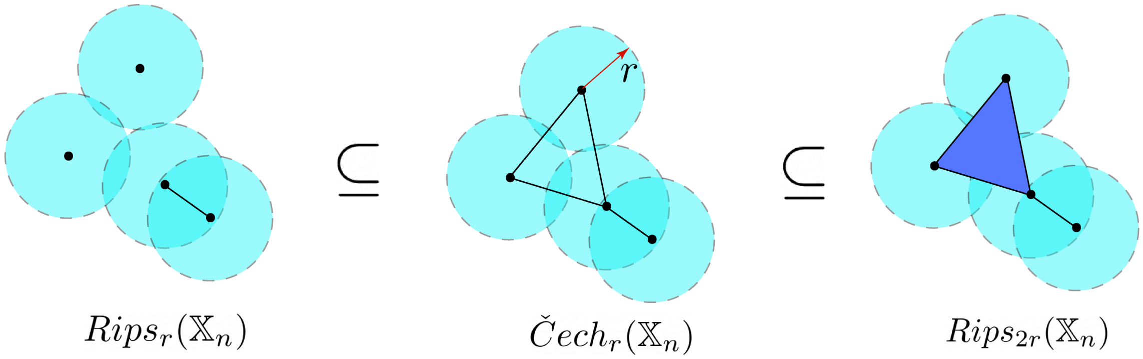

Figure 2.2: The Rips and Čech complexes for a point cloud $\mathbb{X}_n$.

2.2.1 Čech Complex

Definition 2.2.1 (Čech Complex (r))**.** *The Čech Complex of radius r on the point cloud $\mathbb{X}_n \subset M$ is the simplicial complex defined as:*

$$\check{C}ech_r(\mathbb{X}_n) = \{\sigma = [x_0, \ldots, x_k] : \cap_{i=0}^{k} \overline{B(x_i, r)} \neq \phi; x_i \in \mathbb{X}_n\}$$

where M is a metric space.

Remark 2.2.1. *It is clear that $\check{C}ech_r(\mathbb{X}_n)$ is nothing but $\mathcal{N}(\mathbb{X}_n^r)$, the nerve of r-offset of the point cloud $\mathbb{X}_n$.*

So by previous results it follows that $\check{C}ech_r(\mathbb{X}_n)$ is homotopicaly equivalent to the r-offset and hence if we want to observe the change in homology with varying r-offsets it suffices to observe the homology of $\check{C}ech_r(\mathbb{X}_n)$ with varying r.

2.2.2 Rips Complex

Definition 2.2.2 (Vietoris-Rips complex(α))**.** *Given a point cloud $\mathbb{X}_n \subset M$ and a metric space (M, ρ) the Rips complex of radius r is defined as:*

$$Rips_r(\mathbb{X}_n) = \{\sigma = [x_0, \ldots, x_k] : \rho_M(x_i, x_j) \leq r; \forall (i, j); x_i \in \mathbb{X}_n\}$$

It can be easily shown that the Rips and Čech Complex have the following relationship.

Lemma 2.2.1.

$$Rips_r(\mathbb{X}_n) \subseteq \check{C}ech_r(\mathbb{X}_n) \subseteq Rips_{2r}(\mathbb{X}_n)$$

The r-Čech complex is sandwiched between the Rips complex $r \to 2r$ [fig. 2.2], hence for practical purposes when looking at variation of homology we tend to work with rips complexes as they are much easier to compute and still capture the homological variation of the r-offsets due to the sandwich property. In the next chapter we extend this idea to establish the theory of persistent homology.

Chapter 3

Persistent Homology

As discussed at the end of previous chapter, we would like to build various simplicial complexes and see the variation of homology. Intuitively the significant homological features persistently show up. That is the idea of persistent homology, it measures the strength of a particular topological feature by computing how *long* a topological feature lasts. We will formalise all these heuristics in this chapter.

Definition 3.0.1 (Filtrations)**.** *A filtration of a simplicial complex K is a nested family of sub-complexes $(K_r)_{r\in I}$ where $I \subseteq \mathbb{R}$ such that if $r, r'inI$ and $r \leq r'$ then $K_r \subseteq K_{r'}$ and $K = \cup_{r\in I} K_r$*

More generally a filtration of a topological space M is a nested family of subspaces $(M_r)_{r\in I}$. For example, $f : M \to \mathbb{R}$ is a function then $M_r = f^{-1}(-\infty, r]$ for all $r \in I$ defines a sub-level filtration. Similarly $M_r = f^{-1}[r, \infty)$ defines a super level filtration. This can be extended to simplices too. Let K be a simplicial complex with vertex set V and $f : V \to \mathbb{R}$ then f can be extended to K as for any $[v_0, \ldots, v_k] \in K$, $f[v_0, \ldots, v_k] = \max\limits_{i\in\{0,\ldots,k\}}(f(v_i))$.
Then $K_r = \{\sigma \in K : f(\sigma) \leq r\}$ forms a sub-level filtration. Similar construction can be done for super-level filtration.

Remark 3.0.1. *It is evident from the definition that the Čech Complex(r) with increasing r, forms a filtration of simplicial complexes. Similarly the Vietoris-Rips complex(r) with varying r also forms a filtration. Due to the sandwich property of Rips and Čech complexes, it suffices to work with the Rips filtration to capture the homology variation of the offsets.*

Remark 3.0.2. *The concept of sub-level/super-level filtration from a function f is quite general. The r- offsets for a compact space $X \subset \mathbb{R}^d$ can be thought of a sub-level filtration of the function $d_X : \mathbb{R}^d \to \mathbb{R}$. By the nerve theorem we know that the r-offsets and the Čech complex are homotopicaly equivalent. In turn the Čech filtration or the Rips filtration (sandwich property) is actually a special case of the sub-level filtrations for any function f.*

Once we have a filtration for every K_r in the filtration there exists H_p^r homology groups of K_r for all p. If $K_i \subseteq K_j$ then there is a natural induced homomorphism $F_p^{i,j} : H_p^i \to H_p^j$ by the inclusion between the homology groups. The image of these homomorphism tell us which p^{th} homological features that existed at i^{th} state also exist at j^{th} state. Hence these images are called the p^{th} persistent homology groups from i to j.

$$0 = H_p^0 \xrightarrow{F_p^{0,1}} H_p(K_1) \ldots \xrightarrow{F_p^{i,i+1}} H_p^{i+1} \xrightarrow{F_p^{i+1,i+2}} \ldots \xrightarrow{F_p^{n-1,n}} H_p^n = H_p(K)$$

Alternatively we can define the p^{th} persistent homology group from i to j as:

Definition 3.0.2.

$$H_p^{i,j} = {}^{Z_p(K_i)}\!\big/\!_{B_p(K_j) \cap Z_p(K_i))}$$

The rank of the above group is known as the p^{th} persistent betti number $\beta_p^{i,j} = rank(H_p^{i,j})$.

This is similar to the usual definition of homology except the quotienting, in the usual homological definition we wanted to consider cycles which were not part of any boundary, now that we have a concept of filtration and want to retain cycles that persisted from the i^{th} state to the j^{th} state we would want the cycles in the i^{th} state which are not part of the boundary in the j^{th} state.

Computing these persistent homology groups gives us information of when a feature is born and is dead. We would like to find out the features which were born early but died at a later stage. These will be the features we will call *significant*. We incorporate all this information into a single diagram called the persistence diagram which is a topological summary of the underlying space.

3.1 Persistence diagram and Barcodes

Let $\mu_p^{i,j}$ be the number of independent p-dimensional classes that are born at K_i and die at K_j. We then have

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

The $(\beta_p^{i,j-1} - \beta_p^{i,j})$ represents the number of features that died at the j^{th} state and were present at the i^{th} state. The second term $(\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$ represents the number of features that died at the j^{th} state and were present at the $i - 1^{th}$ state; Hence the difference represents the number of features born at the i^{th} state and died at the j^{th} state. Drawing each point (a_i, a_j) with multiplicity $\mu_p^{i,j}$ we get the p^{th} persistence diagram $Dgm_p(\mathcal{F})$ where $\mathcal{F}$ is the filtration.

So each point in the persistence diagram represents a topological feature and it's coordinate are representative of when that feature was born and died respectively.

Lemma 3.1.1 (Fundamental lemma of persistence homology)**.**

$$\beta_p^{k,l} = \sum_{i \leq k} \sum_{j \geq l} \mu_p^{i,j}$$

Hence the persistence diagram encodes all the information about the homology. An alternative representation of persistence diagrams is persistence barcodes where we draw a graph with the horizontal axis representing the index of the filtration and the vertical axis representing the persistent homological features. For each homological feature a horizontal bar of length $j - i$ starting from i and ending at j. Both of these representations are equivalent one can obtain one from other and vice versa.

3.2 Elucidating via examples

Remark 3.2.1. *The persistence diagram for the Rips or Čech filtration would capture the variation of homology of the sub-level sets of the distance function, hence it encapsulates information about the geometry of the underlying manifold. This addresses the original question of estimating the topological features of the underlying manifold, but persistence homology can be used to see variation of homology of any filtration in particular any sub-level/super set filtration of any function f. Henceforth we will consider a general class of*

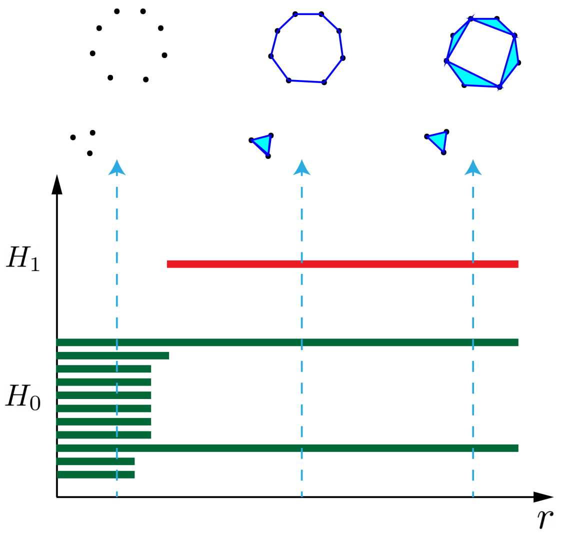

Figure 3.1: Persistent Homology on the $Rips_r(\mathbb{X}_n)$ where $\mathbb{X}_n$ is a point cloud sampled from a circle with an isolated cluster besides it.

filtrations for our upcoming results.

Rips Persistent Homology on Point cloud data

Let us consider a point cloud $\mathbb{X}_n$ sampled from a circle with an isolated cluster besides it (fig. 3.1). $Rips_0$ is the simplicial complex containing only the vertices corresponding to the n points in the point cloud; As we go further in the filtration higher dimension simplices are added to the complex resulting in death and birth of certain homological features. In the above example we start out with 11 points and corresponding to them 11 connected component or H_0 features after a certain r in the filtration only 2 connected components or H_0 features remain corresponding to the circle and the isolated cluster respectively. Also 1 H_1 feature corresponding to the loop of the circle is born at a later stage and persists till the end. Hence when we take *Rips* persistent homology of a point cloud data the longer barcodes corresponding to the persisting homological features represent the underlying topology of the data.

Persistent homology on sub-level filtration of a function

Let $f : \mathbb{R} \to \mathbb{R}$ be a function as described in fig. 3.2, considering the sub-level filtrations $M_r = f^{-1}(-\infty, r]$ for $r \in I$. For $r < a_1$ the homology space is trivial, at $r = a_1$ though an H_0 feature

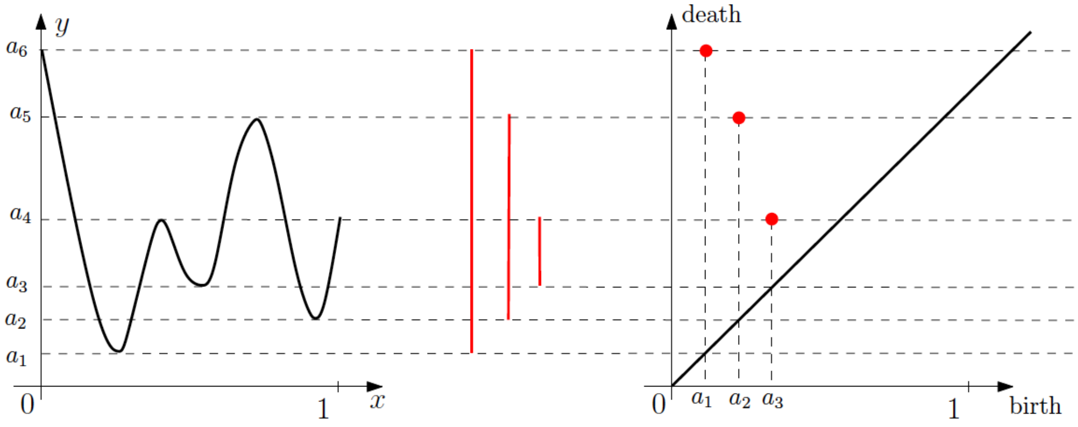

Figure 3.2: Persistent Homology on the sub-level filtration of a function on the left and its corresponding barcode and persistence diagram on the right. **Credits:** [6]

is born the homology feature remains unchanged until $r = a_2$ where a second connected component is born, similarly another connected component H_0 feature is introduced at a_3. At a_4 though there is death of the connected component born at a_3 as it collapses into the connected component born at a_1, likewise the second connected component born at a_2 dies of at a_5. All this information is encoded and represented in the form of barcodes and persistent diagram in fig. 3.2. Each point in the persistence diagram represents the birth -death of a topological feature.

Relation with Morse Theory

Notice that from the previous example it is evident that the birth or death of homological features of the sub-level filtration only occur at the maximas and minimas. This is actual a general property of *Morse functions*. Morse functions are functions with non degenerate critical points and all its critical points being isolated (no two critical points are in each other's neighbourhood). We would briefly touch upon the main results of morse theory as it is an extensive topic on its own and the whole thesis can be dedicated to it. So, we will not delve deep.

Suppose one has a morse function $f : \mathcal{M} \to \mathbb{R}$ defined on the manifold $\mathcal{M}$, there is a result in morse theory stating that the homology of the sub-level filtration of f varies only at the critical values (value of f at the critical points)[4]. Hence it is enough to compute the homologies of the sub-level filtration M_r for $r \in \mathcal{A}$ where $\mathcal{A} := \{\textit{Set of all critical values}\}$,

thereby reducing computation.
Another such result relates the homology of the underlying manifold $\mathcal{M}$ to the critical points of f. The result states that the homology groups of $\mathcal{M}$ are isomorphic to the homology of a chain complex build upon the critical points of f (*Morse complex*) [4].
So basically morse theory gives us a correspondence between the critical points of a morse function and the homology of the underlying manifold. So to draw any topological inference about the underlying manifold it is enough to study chain complexes build on these critical values giving us a reduction. The main results of morse theory on manifolds can be translated to discrete spaces like simplicial complexes as discussed in [3]. Once we have established the notion of morse functions, critical points and morse complex for simplicial complexes we can reduce homology computations on filtrations of simplicial complexes. Suppose we have a filtration K we could look at the reduced complex for each simplicial complex $K_r \in K$ and these gives a reduced filtration cutting computation costs heavily. This sort of optimisation by using morse theory on filtrations has been discussed in great detail in [9].

3.3 Stability

When working with real life data, the point cloud is susceptible to lots of noise. We would like the persistence diagrams acquired to be robust to such kind of noise, i.e given small perturbations to the point cloud, the persistence diagram should not vary much.
If we want to quantify the notion of change, we need to define distances on the space of persistence diagrams. The space of persistence diagrams is a metric space and several distances can be defined on it. The main results of this section are sourced from [10].
Metrics:

1. Haudorff Distance as defines earlier.
2. Bottleneck distance (Wassertein distance with $p \to \infty$)
3. Wassertein distance

Definition 3.3.1 (Bottleneck and Wassertein distance)**.** *The bottleneck and p^{th} wassertein* *distance between two persistence diagram $Dgm(\mathcal{F}_1)$ and $Dgm(\mathcal{F}_2)$ are:*

$$d_B(Dgm(\mathcal{F}_1), Dgm(\mathcal{F}_2)) = \inf_{\gamma} \sup_{x \in Dgm(\mathcal{F}_1)} \|x - \gamma(x)\|_{\infty}$$

$$d_{W_p}(Dgm(\mathcal{F}_1), Dgm(\mathcal{F}_2)) = \inf_{\gamma} \sum_{x \in Dgm(\mathcal{F}_1)} \|x - \gamma(x)\|^p$$

where γ is a bijective mapping between the first and second diagram.

It might be the case that the cardinality of $Dgm(\mathcal{F}_1)$ and $Dgm(\mathcal{F}_2)$ don't match, hence the notion of finding a bijection γ falls apart.In order to circumvent this issue we introduce the diagonal set Δ into the persistence diagrams before computing the distances. Δ intuitively represents the space where a topological feature borns and dies instantaneously. The diagonal set is infinite hence the problem of finding a bijection is now resolved. What we have basically done is match the unmatched off diagonal points to a point in the diagonal.

Definition 3.3.2 (homological critical values)**.** *Let X be a topological space and f a real function on X. A homological critical value of f is a real number a for which there exists an integer k such that, for all ϵ the map $H_k(f^{-1}(-\infty, a - \epsilon)) \rightarrow H_k(f^{-1}(-\infty, a + \epsilon))$ induced by natural inclusion is not isomorphic.*

Basically homological critical values are points where the homology of the sub-level sets changes, for morse functions these are the general critical values of that function. As we know that the homology of the sub-level sets changes only at critical values of the morse function.

Definition 3.3.3 (Tame functions)**.** *A function $f : X \rightarrow \mathbb{R}$ is tame if it has a finite number of homological critical values and the homology groups $H_k(f^{-1}(-\infty, a])$ are finite for all dimensions k and all $a \in \mathbb{R}$*

Tame functions are a more general class of functions. Morse functions on compact manifolds and Piece wise linear functions on simplicial complexes are tame.

Let $Dgm(f)$ be the persistent diagram obtained from the sub-level filtration of $f : X \rightarrow \mathbb{R}$. If X is a triangulable space then there exists a simplicial complex K such that $K \overset{h}{\sim} X$, hence the function f can be extended to the simplicial complex K as mentioned at the beginning section 3. The theoretical results remain the same, but from a practical perspective computations are always done on the triangulation of the space X.

Theorem 3.3.1 (Stability theorem)**.** *Let X be a triangulable space and $f, g : X \to \mathbb{R}$ be tame functions then:*

$$d_B(Dgm(f), Dgm(g)) \leq \|f - g\|_\infty$$

The above theorem implies that if there is a small perturbation to the function f then the persistant diagram of the sublevel sets of f would only perturb by a little bit, and that perturbation is bounded by the L_∞ distance between f and its perturbation.

As mentioned earlier the Čech filtration is a special case of the sub-level filtrations, and the Rips filtration has the sandwiching property. Hence the above stability theorem can be used to prove the following theorem.

Theorem 3.3.2. *Let X, Y be compact metric spaces and $Dgm(Filt(X)), Dgm(Filt(Y))$ be the persistent diagrams of the Rips filtration of X and Y respectively, then:*

$$d_B(Dgm(Filt(X)), Dgm(Filt(Y))) \leq 2D_H(X, Y)$$

The two theorems above imply that persistent diagrams are not susceptible to high variances when dealing with noise .

3.4 Persistence modules

We have already established the concept of persistent homology in the previous sections which involves computing the persistence homology groups $H_p^{i,j}$ for a filtration $K = \cup_{r \in I} K_r$ for all $i, j \in I$ which can be a bit tedious when working with large filtrations; Persistence Module is a compact encoding of the persistence homology vector spaces into a single algebraic object (graded module over a polynomial ring), in turn reducing computations. The results mentioned in this section have been discussed in great detail in [8].
We can calculate homology with coefficients from any base ring R, but we will restrict our attention to homologies computed over coefficients from a field F for reasons explained later in this section.

Definition 3.4.1 (Persistence module)**.** *The persistence module $\mathfrak{M}$ is a collection of vector spaces M^i, $i \in I$ (Some indexing set) together with linear maps $\phi^i : M^i \to M^{i+1}$. The linear maps can be composed to define maps $\phi^{i,j} : M^i \to M^j$.*

Suppose we have a filtration $K = \cup_{r\in I} K_r$, the collection of the p^{th} homologies H_p^r for each simplicial complex K_r, along with the induced homomorphisms $F^{i,j} : H_p^i \to H_p^j$ forms a persistence module H_p^*.

Definition 3.4.2 (Graded Ring). *A graded ring R is a ring which can be decomposed as the direct sum of abelian groups R_i i.e $R \cong \bigoplus_i R_i$, such that for any i, j there is a bilinear product $R_i \otimes R_j \to R_{i+j}$.*

The polynomial ring $R[t]$ over R is a graded ring with $R_i = at^i$ where $a \in R$.

Definition 3.4.3 (Graded Module). *A graded module over a graded ring R is a module $M = \bigoplus_i M_i$ with a direct decomposition such that there is a action of R on M such that $R_i \otimes M_j \to M_{i+j}$.*

Once we have a persistence module H_p^* generated by the homology spaces of a filtration K, we can associate with a graded module $\alpha(H_p^*)$ over the graded polynomial ring over F as follows:

$$\alpha(H_p^*) = \bigoplus_i H_p^i \tag{3.1}$$

The action of $F[t]$ on $\alpha(H_p^*)$ is defined as follows:
Let $(m^0, m^1, \dots)$ be an element in $\alpha(H_p^*)$ then $t \otimes (m^0, m^1, \dots) = (0, F^1(m^1), F^2(m^2), \dots)$. So action of t can be thought of pushing the homology groups to the next stage via the induced homomorphisms in a filtration. We are trying to capture the information of time (index of the filtration) through the action of polynomial ring $F[t]$ on the graded module. Now that we have a graded module over $F[t]$ which is a P.I.D we can apply the structure theorem for P.I.Ds.

Remark 3.4.1. *When we want to apply the structure theorem we would want a simple classification, hence we work with $F[t]$ as its only graded ideals are ideals of the form (t^n) unlike $R[t]$ making its classification quite complex.*

Applying the structure theorem over the graded module $\alpha(H_p^*)$ we get a decomposition as follows:

$$\alpha(H_p^*) \cong (\bigoplus_{i=1}^{n} \sum^{\beta_i} F[t]) \oplus (\bigoplus)_{j=1}^{m} \sum^{\alpha_j} {}^{F[t]}\!/\!_{(t^{k_j})} \tag{3.2}$$

Definition 3.4.4 (l-life intervals). *The l life intervals is an ordered pair (i, j) such that $i < j$ and belong to $\mathbb{Z} \cup \infty$*

In real life we work with only finite filtration K, so all the birth-death pairs of homological features are also l-life intervals. Corresponding to each l life interval (i, j) we can associate the module $M(i, j) = \sum_{k=i}^{j-1} F[t]/(t^{j-k})$. So for a set of l life intervals $\{(i_1, j_1), (i_2, j_2) \ldots, (i_n, j_n)\}$ we can associate the module $M = \bigoplus_{m=1}^{n} M(i_m, j_m)$ which is a direct sum of all the modules associated with each l life interval.

Notice that the module M is actually has the same structure as the decomposition of $\alpha(H_p^*)$ it can be shown that there is one to one correspondence with the birth-death times and the graded module $\alpha(H_p^*)$. Notice that that if my birth-death time is (i, ∞) then it appears as the free part $\sum_{m=i}^{\infty} F[t]$ in the graded module, while (i, j) is represented as a torsion element, including elements which are present at i^{th} stage but die at the j^{th} stage or in other terms get annihilated when t^{j-1} acts upon them. Now that we know that the persistent homology can be encoded in a single algebraic structure, we would like to actually compute it.

Everything remains the same except we change our boundary map and chain complex cleverly.

Definition 3.4.5 (p^{th} persistence chain). *Let PC_p the p^{th} persistence chain for a filtration K be a free module spanned by all the p-simplices in K over the polynomial ring $F[t]$*

Let $B()$ be the birth function which tracks when a particular simplex is introduced in the filtration. Suppose $\sigma \in K_r$ and $\sigma \notin K_i \forall i < r$ then $B(\sigma) = r$.

Definition 3.4.6 (persistence boundary map $P\partial_p()$). *The persistence boundary map $P\partial_p : PC_p \to PC_{p-1}$ is defined as follows:*

$$P\partial_p(\sigma) = (-1)^i \sum_{i=0}^{p} \sigma^{\hat{i}} t^{B(\sigma)-B(\sigma^{\hat{i}})}$$

where σ is a p- simplex $[v_0, \ldots, v_p]$ and $\sigma^{\hat{i}} = [v_0, \ldots, \hat{v}_i, \ldots, v_p]$

Calculating the homology of the p^{th} persistence chain using the p^{th} persistence boundary map we get the graded homology module $\alpha(H_p^*)$.

3.5 Persistence Landscape

In section 3.3 we saw that a metric can be defined on the space of persistence diagrams, but the space of persistence diagrams is not a complete metric space, which is a desired property when we want to do any statistical analysis. In this section we would establish a one to one continuous function representation of the persistence diagrams. These continuous functions known as *persistence landscapes* actually belong to the L_p norm space which we know is complete. After acquiring these function representations, we can borrow the standard statistical framework for *Banach Spaces* and solidify the notion of probability and confidence intervals on the space of persistence landscapes. This section is inspired by the work of Bubenik in [7].

3.5.1 Landscape representation

As discussed in section 3.1 the p^{th}persistence diagram is characterised by $\mu_p^{i,j}$ the number of features born at i^{th} stage and died at j^{th} stage, but all this information is encoded in the betti numbers $\beta_p^{i,j}$ (number of features present at the i^{th} stage which persist atleast till the j^{th} stage). Thus we will work with the persistent betti numbers to define our landscape representation.

Lemma 3.5.1. *let $\tilde{i} \leq i$ and $\tilde{j} \geq j$, then $\beta_p^{i,j} \leq \beta_p^{\tilde{i},\tilde{j}}$.*

The above is a natural outcome of composition of maps, the rank decreases or remains the same after composition.
We already know that the betti numbers $\beta_p^{i,j}$ are valid only when $i \leq j$ or basically above the diagonal. Hence we would perform a change in coordinate axis by rotating by 45^{o} and making the diagonal as the horizontal axis.

$$x = \frac{i+j}{2} \text{ and } y = \frac{j-i}{2} \tag{3.3}$$

So, $\beta_p^{i,j} = \beta_p^{x-y,x+y}$. If we keep x fixed and use the previous lemma we observe that $\beta_p^{x-y,x+y}$ is a decreasing function in y.

Definition 3.5.1 (Persistence landscape)**.** *The persistence landscape* $\Lambda : \mathbb{N} \times \mathbb{R} \to (\mathbb{R} \cup \infty)$

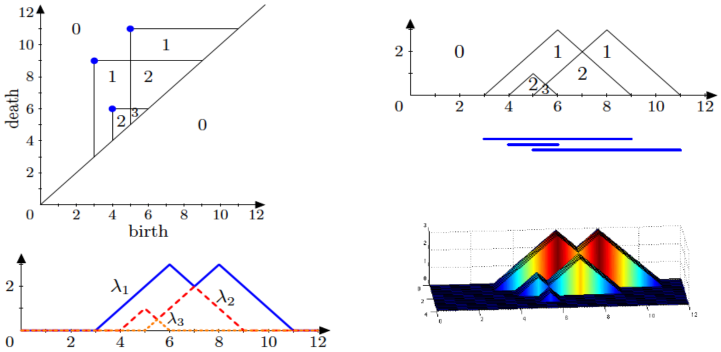

Figure 3.3: Top-Left: Persistence Diagram along with values of $\beta^{i,j}$; Top-Right: Persistence Diagram rotated by 45°; Bottom-Left: Persistence Landscape function corresponding to the above persistence diagram; Bottom-Right: 3-D graph of persistence landscape function. **Credits:**[7]

corresponding to a p^{th} persistence diagram is defined as follows:

$$\Lambda(k,x) = \sup(y \geq 0 \mid \beta_p^{x-y,x+y} \geq k) \tag{3.4}$$

For notational convenience we will denote $\Lambda(k,x) = \lambda_k(x)$ for a fixed k. From lemma 3.5.1 it can be seen that $\lambda_k \geq \lambda_{k'}$ for all $k' \geq k$. The correspondence between persistence landscape and the persistence diagrams is clear from fig. 3.3. Notice that for $k = 1$, Λ traces out triangles corresponding to the dominant homological features. Given a persistence diagram we can obtain a persistence landscape and the reverse direction also holds, given a persistence landscape one can obtain a persistence diagram.

3.5.2 Statistics on Landscapes

Normed space

The persistence landscape Λ is a continuous function from $\mathbb{N} \times \mathbb{R}$ with the measure induced by the product of counting and the standard lebesgue measure. The L_p^{th} norm on Λ is as follows:

$$\|\Lambda\|_p = (\sum_{i=1}^{\infty} \|\lambda_k\|_p^p)^{p^{-1}} \tag{3.5}$$

With the correspondence between the persistence diagram it isn't hard to see the relationship between the norm of landscape and the metrics on persistence diagram. Let $d_b(D,\phi), d_{W_2}(D,\phi)$ be the length of the longest barcode ($\sup_m |j_m - i_m|$) and the sum of squared length of the barcodes ($\sum_{m=1}^{n} |j_m - i_m|^2$) respectively; Where $\{(i_1,j_1),(i_2,j_2)\ldots,(i_n,j_n)\}$ are birth-death pairs of the diagram D.

Remark 3.5.1. *As $\lambda_k \geq \lambda_{k'}$ for all $k' \geq k$, $\|\Lambda\|_\infty = \|\lambda_1\|_\infty$.*

Lemma 3.5.2. *Let Λ be the persistence landscape of a persistence diagram D then the following holds:*

$$\begin{aligned} &(1)\ \ \|\Lambda\|_1 = \frac{d_{W_2}(D,\phi)}{4} \\ &(2)\ \ \|\Lambda\|_\infty = \|\lambda_1\|_\infty = \frac{d_b(D,\phi)}{2} \end{aligned} \tag{3.6}$$

The above lemma can be easily proved by referring to the figure and corresponding the length of the barcodes with the persistence landscapes. (1) is basically the area of the triangle corresponding to each birth-death point and (2) is the peak of λ_1 which will correspond to the height of the triangle corresponding to the longest barcode length.
A consequence of lemma 3.5.2 is that when ever we have a finite persistence diagram i.e the number of birth-death points their corresponding lengths are both finite. Then both $\|\Lambda\|_1$ and $\|\Lambda\|_\infty$ are finite, in turn every $\|\Lambda\|_p$ for all $1 \leq p < \infty$ is finite *(via the decreasing property of L_p norms)*. Hence for any real life case where the persistence diagram is finite $\Lambda \in L_p(\mathcal{S})$ where $\mathcal{S} = \mathbb{N} \times \mathbb{R}$ for all $1 \leq p < \infty$.

Probability in the space of Landscapes

Once we know that $\Lambda \in L_p(\mathcal{S})$ we can borrow the framework of probability and statistics for Banach spaces. We would like to establish a notion of sample mean, expectation, confidence intervals, etc. for persistence landscapes.
Suppose we have n point clouds $\mathbb{X}_1, \ldots, \mathbb{X}_n$ corresponding to them we would get persistence landscapes $\Lambda_1, \ldots, \Lambda_n$, we define the sample mean landscape $\bar{\Lambda}$ as the point-wise mean:

$$\bar{\Lambda}(k,x) = n^{-1} \sum_{i=1}^{n} \Lambda_i(k,x) \tag{3.7}$$

Now we would like to define a probability space on the space of landscapes and would want to extend the results like law of large numbers and central limit theorem for $\bar{\Lambda}$. $\bar{\Lambda}$ isn't a real random variable though it is a random variable in the space of $L_p(\mathcal{S})$.
Let $X : (\Omega, \mathcal{F}, P) \to L_p(\mathcal{S})$ be a $L_p(\mathcal{S})$ r.v. (random variable) then for any functional $f \in L_p^* \cong L_q$(Dual), the composition $X^f = f(X) : (\Omega, \mathcal{F}, P) \to L_p(\mathcal{S}) \xrightarrow{f} \mathbb{R}$ defines a real r.v.

Definition 3.5.2 (Expectation for a L_p random variable). *For a L_p r.v. X the expectation $E(X) \in L_p(\mathcal{S})$ is a function such that for any $f \in L_p^*$:*

$$f(E(X)) = E(X^f) \tag{3.8}$$

Note that in general the expectation $E(X)$ might not exist but if $E(X^{\|\|}) < \infty$ then $E(X)$ always exists.

Theorem 3.5.3 (Law of large numbers). *Let $X_1, \ldots, X_n$ be i.i.d $L_p(\mathcal{S})$ r.v.s then*

$$\bar{X} \xrightarrow{a.s} E(X_i) \tag{3.9}$$

where $\bar{X} = n^{-1}(\sum_i^n X_i)$.

So applying the above theorem to persistence landscapes we get $\bar{\Lambda} \xrightarrow{a.s} E(\Lambda_i)$ where $\Lambda_i, \ldots, \Lambda_n$ are i.i.d $L_p(\mathcal{S})$ r.v. corresponding to the n random sampled point clouds $\mathbb{X}_1, \ldots, \mathbb{X}_n$ from the same manifold $\mathcal{M}$.

Definition 3.5.3 (Gaussian L_p r.v). *Let G be a L_p r.v., G is said to be gaussian if for all $f \in L_p^*$*

$$G^f \sim N(0, Var(G^f))$$

.

A gaussian G variable is completely determined by its *covariance structure*. The covariance structure of G is collection of expectations $E[(G^f - E(G^f)(G^g - E(G^g)]$ for all $f, g \in L_p^*$.

Theorem 3.5.4 (Central limit theorem #1). *Let $X_1, \ldots, X_n$ be i.i.d $L_p(\mathcal{S})$ r.v.s then*

$$\sqrt{n}[\bar{X} - E(X_i)] \xrightarrow{weak} G \tag{3.10}$$

where G has the same covariance structure as X_i for any i.

Replacing X_i with Λ_i we get central limit theorem for persistence landscapes. As discussed earlier though composing a $L_p(\mathcal{S})$ r.v. with $f \in L_p^*$ we get a real valued r.v. We know $\bar{\Lambda}$ is a $L_p(\mathcal{S})$ r.v composing it with f we get a real valued r.v. we can then use the central limit theorem for real r.v.s.

Theorem 3.5.5 (Central limit theorem #2)**.** *Let $X_1, \ldots, X_n$ be i.i.d $L_p(\mathcal{S})$ r.v.s and $f \in L_p^*$ then*

$$\sqrt{n}[\overline{X^f} - f(E(X_i))] \sim N(0, Var(X_i^f)) \tag{3.11}$$

where $\overline{X^f} = n^{-1}(\sum_i^n X_i^f)$.

Using thm 3.5.5 we can form confidence intervals and hypothesis tests for persistence landscapes. Replace X_i by Λ_i and $f = \|\|$ then we get real valued random variables $\Lambda_1^{\|\|}, \ldots, \Lambda_n^{\|\|}$ and the above theorem statement becomes:

$$\sqrt{n}[\overline{\Lambda^{\|\|}} - \|(E(\Lambda_i)\|] \sim N(0, Var(\|\Lambda_i\|))) \tag{3.12}$$

Using the above the $(1-\alpha)$ confidence interval for $E(\Lambda_i)$ becomes:

$$\overline{\Lambda^{\|\|}} \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}} \tag{3.13}$$

Here $S_n = (n-1)^{-1} \sum_{i=1}^n \Lambda_i^{\|\|}$ is the sample variance of $\{\Lambda_1^{\|\|}, \ldots, \Lambda_n^{\|\|}\}$ and $z_{\alpha/2}$ is the $\alpha/2$ critical value of the standard normal distribution.
Similarly if we have two different sample of persistence landscapes $Y = \{\Lambda_1^{\|\|}, \ldots, \Lambda_n^{\|\|}\}$ and $\tilde{Y} = \{\tilde{\Lambda_1}^{\|\|}, \ldots, \tilde{\Lambda'_n}^{\|\|}\}$ want to test whether they belong to the same population or not? i.e they are from the same underlying manifold or not? Then we can define the following Z -statistic which will follow $N(0,1)$ for hypothesis testing.

$$Z = \frac{\overline{\Lambda^{\|\|}} - \overline{\tilde{\Lambda}^{\|\|}}}{\frac{S_n}{\sqrt{n}} + \frac{S_{n'}}{\sqrt{n'}}} \tag{3.14}$$

With this we conclude this chapter. To summarize we started out by formalising the concept of persistent homology. Subsequently, we looked at persistent diagrams and barcode representations capturing the birth-death time of all homology features. The homology features with high birth-death difference are the significant features of the data. Furthermore, we looked at the robustness of these diagrams when introduced to noise. We also looked at a compact representation of the information obtained from persistent homology in terms of a single algebraic object known as the persistent module. Finally we explored the correspondence between persistent diagrams and persistence landscapes which enabled us to perform standard statistical analysis. We addressed the randomness induced by sampling from the underlying manifold in terms of probability distributions on the space of landscapes.

Chapter 4

Mapper Algorithm

We already looked at how persistent homology helps us draw inference about the significant geometrical features of the data. In this chapter we look at another celebrated tool in the realm of topological data analysis called the mapper algorithm. The mapper algorithm provides us network visualisations of data of arbitrary dimensions, while retaining the major structures in the data.

Recall that the nerve theorem 2.1.2 tells us that if we have a nice enough cover *(contractible intersections)* of a space X then the nerve of the cover is homotopicaly equivalent to X, but what happens when we relax the condition of contractible intersections? The corresponding nerve should capture some summary information of the space to some degree even if not homotopicaly equivalent, this is the motivation behind the mapper algorithm.

Definition 4.0.1 (Refined pull back)**.** *Let $f : X \to \mathbb{R}^d$ be a continuous real valued function and let $\mathcal{U} = (U_i)_{i\in I}$ be a cover of $\mathbb{R}^d$. The pull back cover of X induced by $(f,\mathcal{U})$ is the collection of the open sets $(f^{-1}(U_i)_{i\in I})$. The refined pull back is the collection of connected components of the open sets $f^{-1}(U_i)_{i\in I}$.*

Notice that if $f : X \to \mathbb{R}^d$ is a continuous function then the *refined pull back* gives an open cover of X. Once we have an open cover for X we can look at the nerve of the refined pull back. The following summarises the mapper algorithm pipeline:

Input:

1. A data set X with a metric or dissimilarity measure between data points.

2. A lens function $f : X \to \mathbb{R}(or\mathbb{R}^d)$
3. Cover $\mathcal{U}$ of $f(X)$

For each $U \in \mathcal{U}$ decompose $f^{-1}(U)$ into clusters $cls = \{C_{U_1}, \ldots, C_{U_k}\}$ using any standard clustering algorithm like DBSCAN. Compute the nerve of this cover of X defined by cls.(computing nerve of the refined bull back of f)

Output:

1. A simplicial complex (nerve).
2. 1-skeleton (*restriction of the simplicial complex to its 1 and 0 simplices only*) of the nerve to get a graph representation.

The graph is a representation of how the underlying space looks with respect to the function f hence the name *lens function*. For the mapper algorithm we require two inputs the lens function f and the cover of the image space.

4.1 Example

The above fig. 4.1 elucidates mapper algorithm applied to a set of points sampled from the surface of a "pair of pants". It is evident that the output of the mapper algorithm is dependent on the choice of the filter function; When we use width function w as the filter function the "eight shape" structure of the space is captured on the other hand when height function h is used the "flare shape or inverted-y" structure is captured. The mapper algorithm gives representation of the space w.r.t the filter/lens function. One can use varying filter functions to get an idea of an overall geometry of the underlying space.

4.2 Choice of lens and cover

Choice of cover $\mathcal{U}$: The general approach is to segment the image space $f(X)$ into n regular intervals with some overlap o between them. As n increases the number of nodes in the graph increase and more detailed the graph gets, as o increases more edges between the nodes form due to excessive overlapping. Generally the value of o is chosen to be ~ 0.25 corresponding to 25 percent overlap between the intervals. If the image space is d -dimensional then the

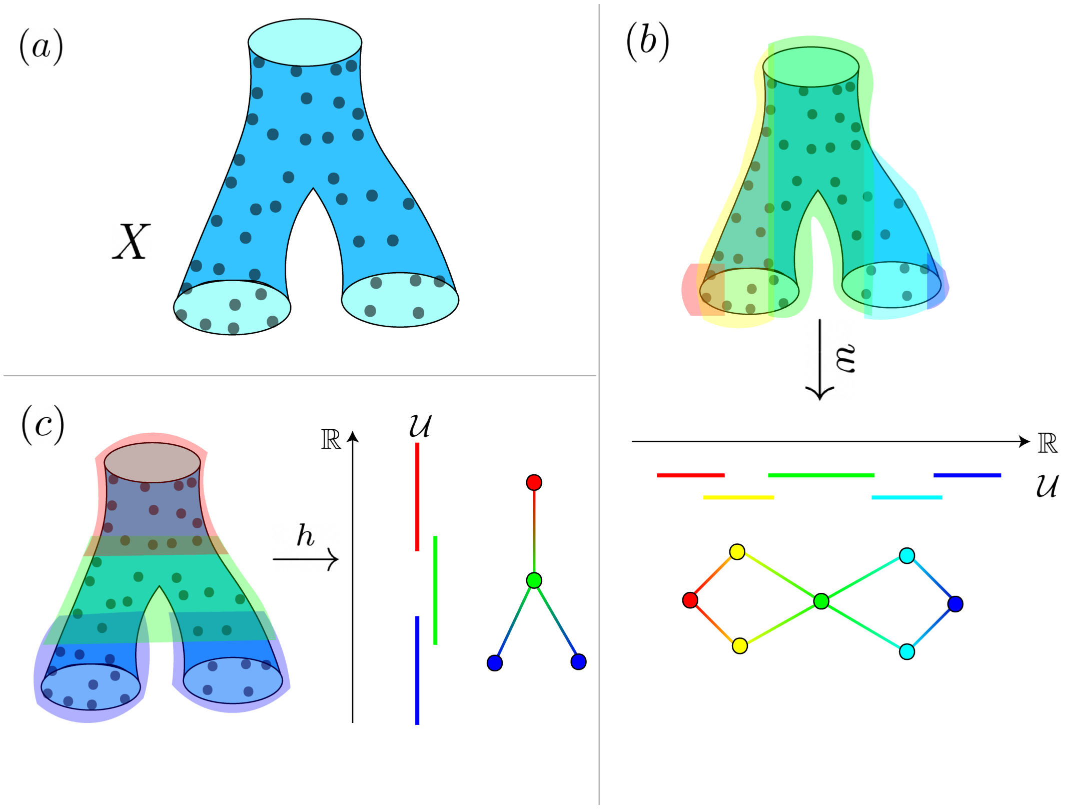

Figure 4.1: (a) Point cloud X sampled from the surface of "pair of pants"; (b) Mapper algorithm on X with $w : X \to \mathbb{R}$ width as the filter function; (c) Mapper algorithm on X with $h : X \to \mathbb{R}$ height as the filter function.

inputs to n and o are d -dimensional vectors specifying the number of regular intervals and overlap for each coordinate space.

Choice of lens: There are a lot of standard go-to statistical lens functions used to look at the data; kNN distance, projection on one of the coordinates, L_p norms, T-SNE, ISOMAP, PCA components to name a few. The standard rule of thumb is to try various lens functions to get an understanding of the layout of the underlying space.

Mapper is generally used as an exploratory analysis tool and is superior to other dimensionality reduction techniques as the clustering is done in the original metric space from where the point cloud resides rather than on some projection or embedding, in turn reducing loss of information.

4.3 Reeb Graph

In this section we view the output of mapper algorithm as a discretization of an object called the *Reeb Graph*. Given a continuous function $f : X \to \mathbb{R}$, its structure can be visualised by the variation of its level sets. Let $x, y \in X$, we can define an equivalence relation $\sim$ on X as follows:
$x \sim y$ if and only if $f(x) = f(y) = a$ and x, y belong to the same connected component of $f^{-1}(a)$.
The quotient space $R(f) = {}^{X}\!/\!_{\sim}$ is known as the Reeb Graph of X w.r.t f. As the $R(f)$ is just the quotient space of X it preserves *connectedness* of X, also it can be shown that any loops in $R(f)$ correspond to loops in the original space X. This leads to the following properties of the Reeb Graph:

1. $\beta_0(R(f)) = \beta_0(X)$
2. $\beta_1(R(f)) \leq \beta_1(X)$

The above properties suggest that Reeb graph $R(f)$ is a reduction of the original space X based on the contours of f. Even though $R(f)$ doesn't fully capture the topology of the original space X it preserves certain geometrical structures. (fig. 4.2)

Reeb Graph and Mapper Algorithm

It can be seen that the 1-skeleton of the nerve obtained from the mapper algorithm is nothing but just the discretization of the Reeb graph. The mapper algorithm performs clustering on the original point cloud $\mathbb{X}_n$ to obtain components of the inverse image of a filter function f and draws edges between them based on overlapping. The output of Mapper can be thought of an estimation of $R(f)$ where the domain of f is extended to the underlying manifold $\mathcal{M}$ from which point cloud $\mathbb{X}_n$ is sampled from.
As discussed in section 3.2, for a *Morse function* f the homology of the sub-level sets only change at critical values of f. Consequently, if we look at the Reeb Graph $R(f)$, its structure changes at contours of critical values of f. So, when estimating the Reeb Graph via the mapper algorithm it is enough to form a cover $\mathcal{U}$ for $f(\mathbb{X}_n)$, such that critical values of

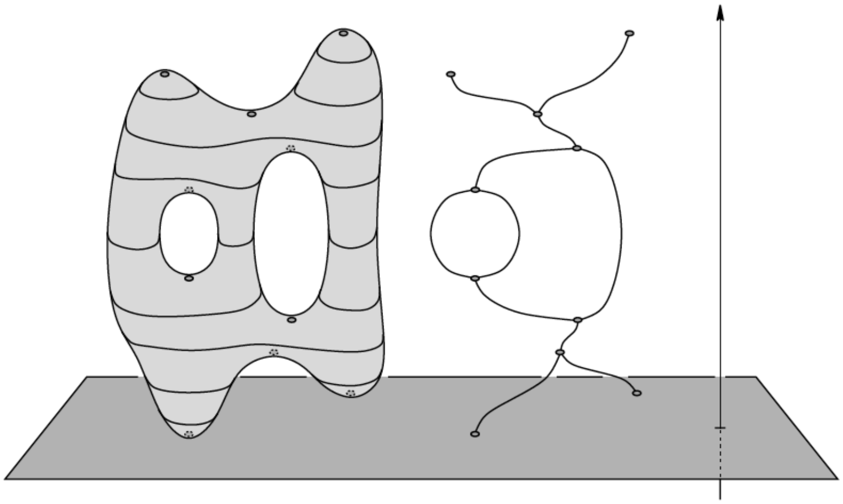

Figure 4.2: Reeb Graph $R(f)$ of X, where f is the height function and X is a two-holed torus. **Credits:** [2]

$f(\mathcal{M})$ are isolated i.e no two critical values belong to the same U_i for any i.
In real life we don't know the distribution of the critical values of f on the underlying manifold $\mathcal{M}$, hence the general thumb rule is to keep the parameter n large enough to segregate the critical values of f when constructing the cover $\mathcal{U}$ of the image set of f.

Chapter 5

Applications

This chapter will focus on the interdisciplinary applications of TDA addressed in this project. The first section will discuss the application of persistence homology on oncology data to quantify interaction between malignant cells and T-cells; The second section will be dedicated to application of TDA on US S&P500 and JPN N-225 stock prices data to represent state of market using barcodes. These representations can then be used to summarise the evolution of the market. All the coding has been done in `R` using the `TDA` package [13].

5.1 Oncology Data

The aim of this collaborative work was to quantify and identify regions of tissue where the cancer cells and T-cells interact; This is cardinal to measure the effectiveness of a drug trial as more interaction is indicative of the potency of a given drug. Identification of regions with high interactions enables us to perform targeted treatment.

Methodology:

- Input data is a matrix with coordinates of cells in the x-y plane along with their respective cell type (M-Malignant cell , T-Immune cell)
- The data is a subset of $\mathbb{R}^2$ so we built a triangulation K of $\mathbb{R}^2$ over the data as shown in fig. 5.1 (b).

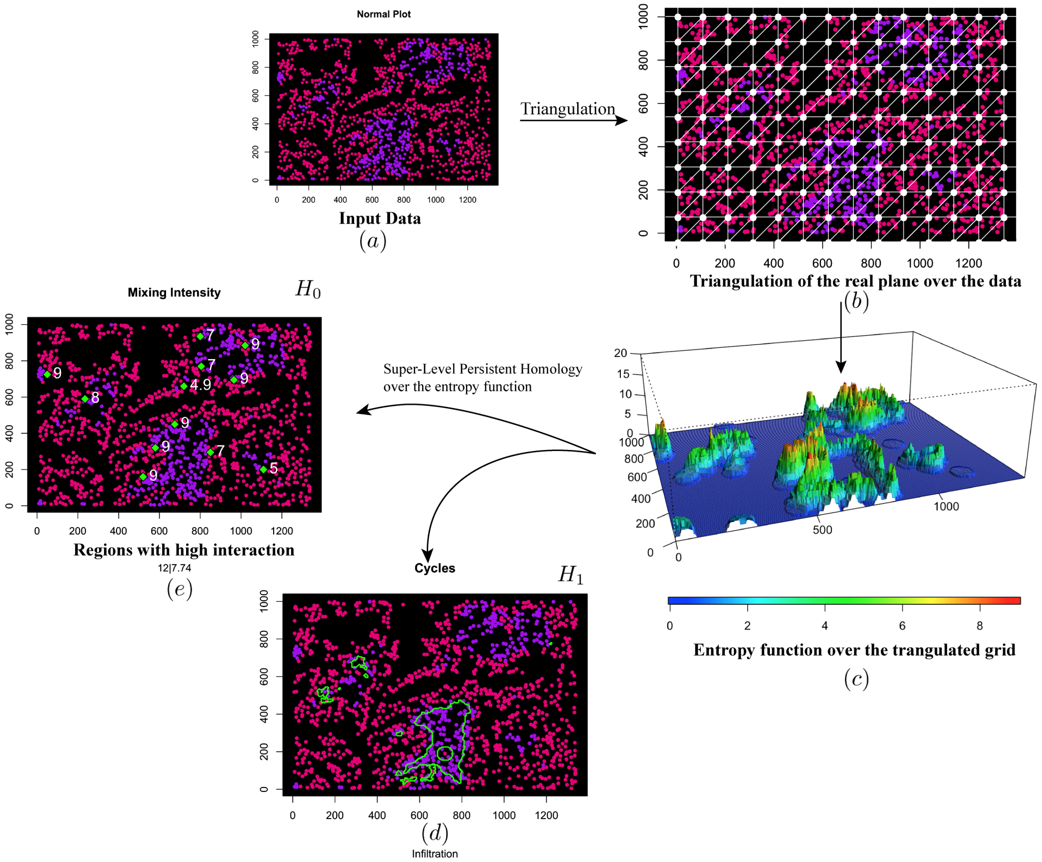

Figure 5.1: Persistence Homology on the super-level filtration of the entropy function. **Credits:** [21]

- The triangulation K is a simplicial complex with vertex set V. We define entropy function $S_r : V \to \mathbb{R}$ which measures interaction between T and M cells in a neighbourhood of v for all $v \in V$.

- $$S_r(v) = \min(n_t(v), n_m(v))(-\log_2 p_m(v) - \log_2 p_t(v))$$

  Where $n_t(v), n_m(v)$ are number of T-cells and M-cells within the distance r from v. $p_t(v) = {}^{n_t(v)}/_{n_t(v)+n_m(v)}$ is the frequency of T- cells within distance r and similarly $p_m(v)$ is the frequency of M-cells. The entropy function is zero if either of the frequencies are zero and takes high values when the frequencies are uniform and number of cells are high.

- S_r is the measure of interaction between different types of cells in a neighbourhood [fig. 5.1 (c)]. The persistent homology of the super-level filtration of S_r will provide us *persistent* H_0 (connected components) and H_1 (loops) features where the value of S_r is high; In other words regions or loops where interaction between the T and M cells is significant.

- In order to calculate the super-level filtration of S_r we first extend the function to the triangulation K as mentioned in chapter 3. For any $[v_0, \dots, v_k] \in K$, $S_r[v_0, \dots, v_k] = \min\limits_{i \in \{0,\dots,k\}} (S_r(v_i))$.
- Then $K_r = \{\sigma \in K : f(\sigma) \geq r\}$ forms a super-level filtration. After computing the persistent homology in this filtration, the persistent H_0 features along with their birth values (max S_r value of that component) are obtained [fig. 5.1 (e)] denoting the significant regions of M-T cells interaction.

The initial problem of quantifying and identifying regions interaction between malignant cells and T-cells was addressed but the method is quite general and can be thought of a density based clustering technique where the density is representative of the interaction between different labelled data points.

5.2 Financial Data

It is well known that the state of market can be studied from the correlation matrix between the constituent stocks, as stocks tend to be highly correlated during a crash and act quite independently in a calm period. Computing the rips persistent homology on the distance matrix induced by the correlation matrix we obtain a more comprehensive barcode representation of the market which capture the internal fluctuations within the market. We can use these barcode representations to generate a summary of the evolution of market and also as an input to mapper giving us a network representation of the same.

Methodology:

- The data in hand is the US S&P500 and JPN N225 stock prices. For a particular market we have N time series corresponding to the prices of the N stocks. We take the log-returns of these time series to obtain N log-return time series.

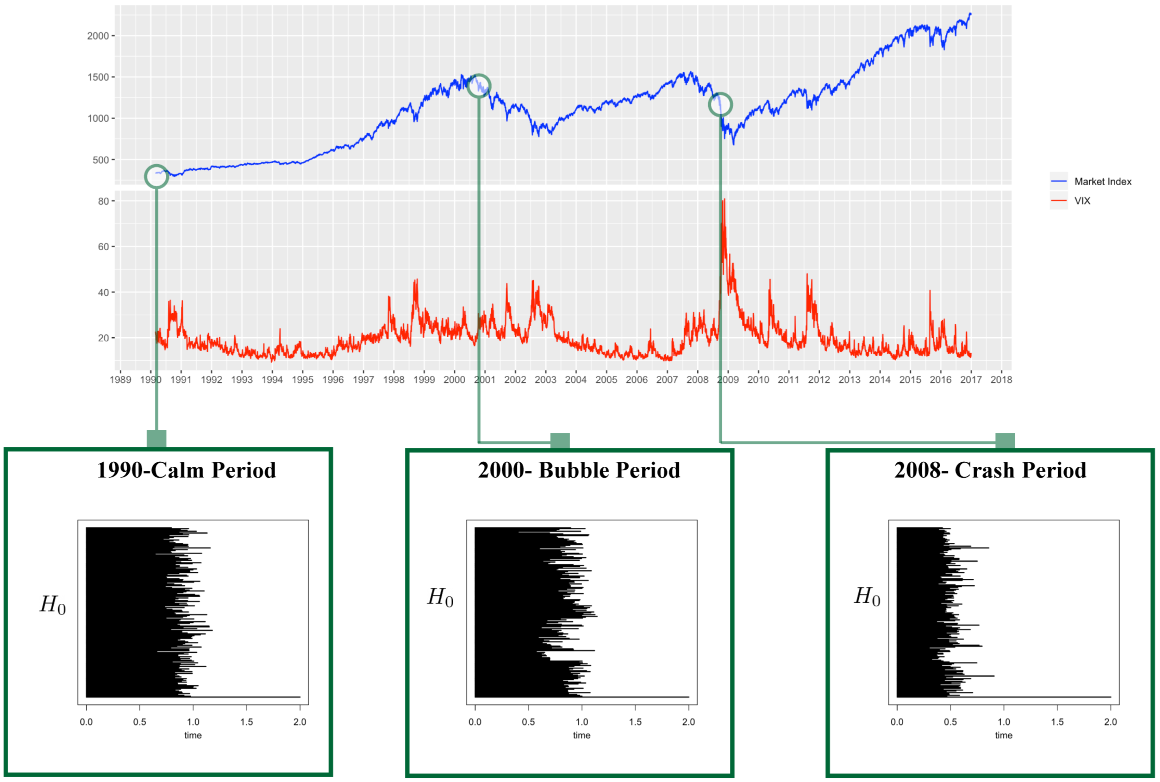

Figure 5.2: The historical values of SP 500 index and Volatility index (VIX) are plotted against time from 1990-02-27 to 2016-12-29 in panels (1) and (2). (3) Time periods corresponding to (a) calm (1990-01-22 to 1990-03-19), (b) bubble(2000-09-06 to 2000-10-31) and (c) crash(2008-08-21 to 2008-10-16) are chosen, with their corresponding barcode diagrams. Where the barcode diagrams are calculated by performing rips persistent homology on the distance matrix D corresponding to the respective time frames.**Credits:** [20]

- We segment the log-return times series into T time frames τ_k where $k = \{1, \dots, T\}$.
- For each time frame τ_k calculate the correlation matrix $C(\tau_k)$ between the N stocks. Now a distance matrix $D(\tau_k)$ between the N stocks is induced from $C(\tau_k)$ as follows $D(\tau_k)(i,j) = \sqrt{1 - C(\tau_k)(i,j)}$.
- Once we acquire the distance matrix $D(\tau_k)$ we can perform persistent homology on the rips filtration built over the point cloud of these N stocks.
- We only focus on the H_0 homology features; The 0^{th} barcode representation $B(\tau_k)$ will comprise of N components with their death times. The birth time will be 0 as

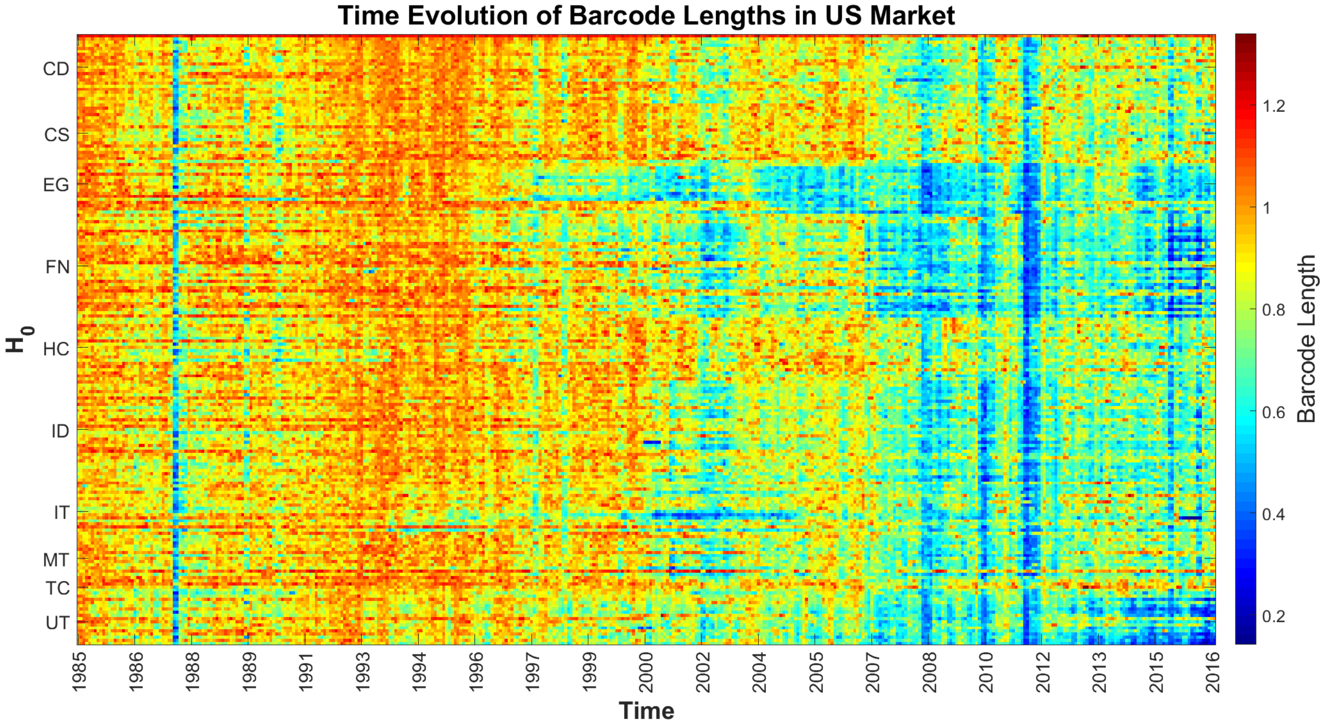

Figure 5.3: The time series entire time period from 1985-01-03 to 2016-12-30 was divided into 402 windows of length 40 which shifted by 20 on each step and correlation matrix and distance matrix was calculated from the 194 stock return time series corresponding to each of those windows. The barcode lengths from the persistence homology of each frame is then stored in the column corresponding to the frame number creating 194x402 matrix M, where M_{ij} = Length of barcode of i^{th} stock on j^{th} time step. The entire 194x402 matrix is coloured according to the value of each element using a colormap which created the visual representation that is being presented in this figure. The stocks are ordered in such a way that the ones which are in the same sector come together which is creating the block behavior in the diagram. **Credits:** [20]

the $Rips_0$ complex is just the simplicial complex only N vertices corresponding to the N stocks. As we go further in the filtration stocks closer to each other in terms of the distance matrix D start collapsing into single components resulting in death of H_0 features.

- Hence during a crash when market is highly correlated the barcodes will be short lived and when there is a calm period the barcodes are long; While in the bubble period sectoral correlation is captured in form of a groove in the barcode diagram [fig. 5.2].

- For a particular time frame τ_k the barcode representation $B(\tau_k)$ (N dimensional vector) is representative of the state of the market.

- We concatenate all the $B(\tau_k)$ vectors for each time frame τ_k, $k = \{1, \ldots, T\}$. To get a $N \times T$ summary matrix. We generate a heat plot as shown by colouring the summary matrix based on its index values.

- The various crashes are captured in the heat plot in the form of vertical blue stripes due to the short barcode lengths during crashes. While the sectorial behaviours are captured via horizontal blue stripes. Hence the formation of bubbles and distinction between exogenous and endogenous crashes can be made looking at the heat plot. The heat plot provedes a comprehensive summary of the evolution of market. [fig. 5.3]

- We also build a network representation of how certain time frames are related to each other in terms of market behaviour. We consider each time frame τ_k to be a point in $\mathbb{R}^{N-1}$ dimensional space where its coordinates are the $N-1$ components of $B(\tau_k)$ after dropping the first component. Then we use various filter functions such as l_2 norm, mean correlation, entropy, etc. and appropriate mapper parameters to get a network summary of the evolution of the market.

Remark 5.2.1. *Note that we only take $N-1$ components of the barcode vector $B(\tau_k)$ as the first component will be the length corresponding to the component when every stock collapses into a single H_0 feature. Hence the length of this component will always be the same for any time frame τ_k so we don't consider this component as an input to mapper.*

Conclusions and Further Readings

In this thesis, we explored the mathematics behind the two most celebrated techniques in TDA, Persistent Homology and the Mapper Algorithm. We learned the theory behind persistent homology and saw how it can be used to extract homological information from the given data. We looked at how to encode this information in terms of persistent diagrams and barcodes; Followed by performing statistics using the concept of persistence landscapes. During the end, we addressed the mapper algorithm as a superior visualisation technique compared to its statistical alternatives. Finally, we finished the thesis with some applications of persistent homology on medical and financial data implemented during this project.

TDA is a vast topic marrying concepts of algebraic topology and statistics. Hence, it is unfeasible to present a detailed study of all the aspects the field has to offer. Although, one can refer to the following sources for further reading.

- A detailed description of building various types of complexes from data and manifold reconstruction can be found in, [14].
- One can refer to Bubenik's paper on the properties of persistence landscape to get a comprehensive understanding of the same, [19].
- Recently, the homological study of statistical functions like the distance to measure and other kernel density estimators is a topic of interest in mode detection, [15].
- The mapper algorithm has hyper-parameters such as the proximity value for clustering, the number of intervals, etc. A statistical analysis of parameter selection has been done in, [16].
- The classic hierarchical clustering using HDBSCAN can be justified mathematically by homologies of Rips Filtration, as shown in, [17, 18].

Besides exploratory data analysis, recently TDA has been integrated with numerous supervised learning techniques for prediction and simulation, based on the underlying geometry of the data. TDA's use in machine learning is still inchoate though having immense potential.

Bibliography

[1] Allen Hatcher, Algebraic Topology, Cambridge University Press 2001.

[2] Herbert Edelsbrunner and John L. Harer, Computational Topology An Introduction, American Mathematical Society 2009.

[3] Kevin P. Knudson, Morse Theory Smooth and Discrete, World Scientific 2015.

[4] J. Milnor, M. Spivak and R. Wells, Morse Theory, Princeton University Press 1963.

[5] Robert Ghrist, Elementary Applied Topology, Createspace Independent Pub 2014.

[6] Frédéric Chazal and Bertrand Michel, *An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists*, (2017) arXiv:1710.04019 [math.ST]

[7] Peter Bubenik, *Statistical Topological Data Analysis using Persistence Landscapes*, Journal of Machine Learning Research 16 (2015) 77-102.

[8] Afra Zomorodian and Gunnar Carlsson, *Computing Persistence Homology*, Discrete Comput Geom (2005) 33: 249

[9] Konstantin Mischaikow and Vidit Nanda, *Morse Theory for Filtrations and Efficient Computation of Persistent Homology*, Discrete Comput Geom (2013) 50: 330-353.

[10] Cohen-Steiner, D., Edelsbrunner, H. and Harer, J., *Stability of Persistence Diagram* Discrete Comput Geom (2007) 37: 103.

[11] Larry Wasserman, *Topological Data Analysis*, Annual Review of Statistics and Its Application 2018 5:1, 501-532.

[12] Gunnar Carlsson, *Topology and Data*, Bull. Amer. Math. Soc. 2009 46 (2), 255-308.

[13] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci and Clment Maria, *Introduction to R package TDA*, (2014) arXiv:1411.1830 [cs.MS]

[14] Jean-Daniel Boissonnat, Frédéric Chazal and Mariette Yvinec, Computational Geometry and Topology for Data Analysis, 2016.

[15] Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertr Michel, Aless, ro Rinaldo and Larry Wasserman; *Robust Topological Inference: Distance To a Measure and Kernel Distance*, Journal of Machine Learning Research, 2018 18(159):1–40.

[16] Mathieu Carrire, Bertrand Michel, Steve Oudot; *Statistical Analysis and Parameter Selection for Mapper* , Journal of Machine Learning Research, 19(12):139, 2018.

[17] J.F. Jardine, *Stable components and layers*, (2019) arXiv:1905.05788 [math.AT].

[18] J.F. Jardine, *Data and homotopy types*, (2019) arXiv:1908.06323 [math.AT].

[19] Peter Bubenik, *The persistence landscape and some of its properties*, (2018) arXiv:1810.04963 [math.AT].

[20] A. Chakraborti, Sourish Das, Hrishidev U, Rajdeep Haldar *[Ongoing work]*, (2019) .

[21] A. Rao, Sourish Das, P. Deshpande, Rajdeep Haldar *[Ongoing work]*, (2019) .