

On topological data analysis for SHM; an introduction to persistent homology

T. Gowdrige, N. Dervilis, K. Worden

Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield
Mappin Street, Sheffield S1 3JD, UK

Abstract

This paper aims to discuss a method of quantifying the 'shape' of data, via a methodology called *topological data analysis*. The main tool within topological data analysis is *persistent homology*; this is a means of measuring the shape of data, from the homology of a simplicial complex, calculated over a range of values. The required background theory and a method of computing persistent homology is presented here, with applications specific to structural health monitoring. These results allow for topological inference and the ability to deduce features in higher-dimensional data, that might otherwise be overlooked.

A simplicial complex is constructed for data for a given distance parameter. This complex encodes information about the local proximity of data points. A singular homology value can be calculated from this simplicial complex. Extending this idea, the distance parameter is given for a range of values, and the homology is calculated over this range. The persistent homology is a representation of how the homological features of the data persist over this interval. The result is characteristic to the data. A method that allows for the comparison of the persistent homology for different data sets is also discussed.

Key words: Topological data analysis; Persistent homology; Simplicial complex.

1 Introduction

Topological methods are very rarely used in structural health monitoring (SHM), or indeed in structural dynamics generally, especially when considering the structure and topology of observed data. Topological methods can provide a way of proposing new metrics and methods of scrutinising data, the most rudimentary and most powerful of which, *persistence homology*, will be discussed in this paper.

SHM has been dominated by the insurgence of machine learning since its introduction to the field [1, 2]. *Topological Data Analysis* (TDA) aims to work alongside the vast work already established to provide a new light into ways that data can be analysed. Previous machine learning operations in SHM have never distinctly considered the data *topology*. This paper aims to bring to light the potential importance of TDA in SHM. TDA generalises well into higher dimensions, so an insight into higher-dimensional data structures can be obtained.

TDA has previously proven useful in other research areas, such as medicine, where virus evolution has been tracked. TDA has also been used to sequence viruses such as Influenza A and HIV, along with many other uses such as clustering diabetes types. In economics, TDA has been used to identify topological patterns in multi-dimensional time series data. It is thought that topological data analysis might be able to give early warning signs of imminent market crashes.

The contents of this paper aim to walk through the process of performing TDA. A brief outline of the methodology is given here, to construct a geometric object called a *simplicial complex* from data. A simplicial complex can be thought of as a higher-dimensional analogue of a graph. The vertices of the simplicial complexes are the data points, and the connections between the vertices are less than a prescribed threshold distance. Therein, the simplicial complex encodes information about connection between the vertices. This geometric object has now attributed shape to the data. Following from the simplicial complex, this information can be manipulated in order to output algebraic

groups that will capture information about the shape and structure of the connection in the simplicial complex, most specifically about the number of k -dimensional holes that will be found within the assumed manifold underlying the data; these groups are called the *homology groups*. A generalisation of the homology will be discussed with respect to pioneering work by Edelsbrunner [3, 4] where the homology can be considered over a range of simplicial complexes; this is called the *persistent homology*, aptly named as this uncovers how the homology persists over a range of values.

The layout of the paper is as follows: Section 2 give the necessary definitions, listed at the beginning to provide the required mathematics that is not usually common to engineering. Section 3 will be devoted to persistent homology, its significance, provide an intuitive understanding and show two common forms of how it can be displayed. Section 4 will show the topology of some known shapes and how the persistent homology links to these. Section 5 will introduce an interesting engineering specific example which will show how this theory can be used. Following this, the paper concludes.

2 Background Theory and Definitions

At the expense of labouring multiple required definitions, the theory pertinent to the paper topic will quickly follow. For more details, the following works can be consulted [5–9].

2.1 Algebra

Definition 2.1. An *equivalence relation*, denoted by \sim , is a binary operation that is *reflexive* ($a \sim a$), *symmetric* (if $a \sim b$ then $b \sim a$) and *transitive* (if $a \sim b$ and $b \sim c$ then $a \sim c$). The equivalence relation provides a partition of a set into elements that share a common property.

Definition 2.2. *Groups* are an extension of the concept of set, to include a binary operator. This set-operator pair is written $G = \{S, \circ\}$. There are four necessary axioms associated with a group, these are: *closure*, *associativity*, *existence of a unique identity*, and *the existence of an inverse* for every element in the group.

Definition 2.3. A *subgroup*, H , of a group, G , is a subset such that $H \subseteq G$, and H satisfies all of the group axioms. The operation and identity of the subgroup is inherited from the parent group.

Definition 2.4. A group is called an *abelian group*, if the result of the operation is independent of the ordering of elements. This is the case when the operation is *commutative*. $x \circ y = y \circ x, \forall x, y \in G$.

Definition 2.5. Let G be a group and H a subgroup of G . A *left coset* of H in G is a subset of G of the form $gH = \{gh \mid h \in H\}$ for some $g \in G$, the set of left cosets of H in G is written G/H . Similarly, a *right coset* of H in G is a subset of G of the form $Hg = \{hg \mid h \in H\}$ for some $g \in G$, and the set of right cosets is written $H \setminus G$ [5].

Definition 2.6. A subgroup H of G is said to be *normal* if $gH = Hg, \forall g \in G$, this is written $H \triangleleft G$. In this case, the set of cosets G/H and $H \setminus G$ are the same by definition. The set of cosets for normal subgroups is called the *quotient group*.

Definition 2.7. A group *homomorphism* is a *mapping* between two groups, $h : G_1 \rightarrow G_2$. The mapping is considered a homomorphism if the identity element in G_1 is mapped to the identity element in G_2 , and the group operation distributes over the homomorphism, $h(u * v) = h(u) \circ h(v), \forall u, v \in G_1$.

Definition 2.8. Given a group homomorphism $h : G_1 \rightarrow G_2$, the *kernel* of h , $\ker(h) \subset G_1$, is the set of elements x such that $h(x) = e$, where e is the identity element. The *image* of h , $\text{im}(h) \subset G_2$, is the set of elements y such that $y = h(x)$ for some x .

Definition 2.9. A *metric space* is defined by a pair (X, ∂_X) . X refers to the set where the elements of the metric space live and ∂_X is the associated metric or distance function between two points in X . For a space to qualify as a metric space, the following criteria must be true: $\partial(x, y) \geq 0$, $\partial(x, y) = \partial(y, x)$, and $\partial(x, z) \leq \partial(x, y) + \partial(y, z)$.

Definition 2.10. An *open ball* is defined as $B_\epsilon(x) = \{y \in X \mid \partial_X(y, x) < \epsilon\}$. This encloses a space around the point x , where all points enclosed are less than the distance ϵ from the point x . This space is often referred to as the ϵ -neighbourhood of x .

Definition 2.11. A *topological space* is represented by a pair, (X, \mathcal{T}) , where X is the set of all the elements and \mathcal{T} is a collection of subsets, referred to as the *topology*. The *open sets* in \mathcal{T} must satisfy the following axioms: The

set of elements and the empty set are elements of the topology, any union of sets in \mathcal{T} is also an element of \mathcal{T} , the intersection of a finite collection of elements of \mathcal{T} is an element of \mathcal{T} .

Definition 2.12. Given two topological spaces, (X, \mathcal{T}_X) and (Y, \mathcal{T}_Y) . These spaces are said to be *homeomorphic* if there exists a bijective continuous map between them.

Now that a topological space has been discussed, when taking homeomorphisms between spaces there are quantities characteristic of that structure called *topological invariants*; these do not change under homeomorphisms. It will be the main purpose of this review to develop a method of calculating a topological invariant called the *persistent homology*.

2.2 Manifolds

Manifolds are continuous surfaces from which the data are assumed to be sampled. Data that are observed are assumed to lie on the surface of a manifold. By understanding the topology of the sampled data points, it is the aim of TDA to extract topological information about the underlying manifold from the sampled data. The manifold is unknown prior to analysis and persistent homology will identify features within the manifold over a range of length scales. Thereby, understanding the shape of the sampled data, it is the conjecture of TDA that the shape of the manifold is also understood. Formally, a manifold is a space that is locally homeomorphic to some n -dimensional Euclidean space, \mathbb{R}^n [6].

2.3 Simplicial Complex

Simplicial complexes are a way of representing data sampled from a manifold; they can be thought of as higher-dimensional analogues of graphs, giving a way of encoding connections between vertices. The dimension that a simplicial complex can capture is restricted by the number of points that are fully connected. Simplicial complexes can be manipulated to output the homology for the data, and following this, the persistent homology.

A simplicial complex is a structure made up of fundamental building blocks called *simplices*. The first four simplices can be seen in Figure 1. Each vertex in the simplex is fully connected to all the other vertices and the space enclosed by the vertices is part of that simplex.

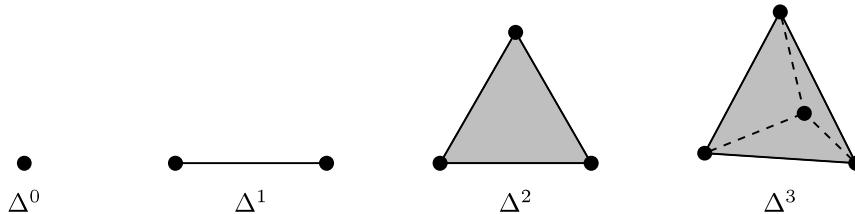


Figure 1: The first four simplices.

There are many ways to construct a simplicial complex from point data. For simplification, only one method will be discussed within this paper, the *Vietoris-Rips* complex. The Vietoris-Rips (VR) complex can be constructed for point data to output a corresponding complex according to the rules $VR_\epsilon(X, \partial_X)$: let (X, ∂_X) be a finite metric space and $\epsilon > 0$ be a fixed value. The abstract simplicial complex is determined by the rules [10]:

1. The vertices, $v \in X$, form the vertices in $VR_\epsilon(X, \partial_X)$.
2. A k -simplex is formed when $\partial_X(v_i, v_j) \leq 2\epsilon$, $\forall i, j \leq k$ for some $\epsilon > 0$.

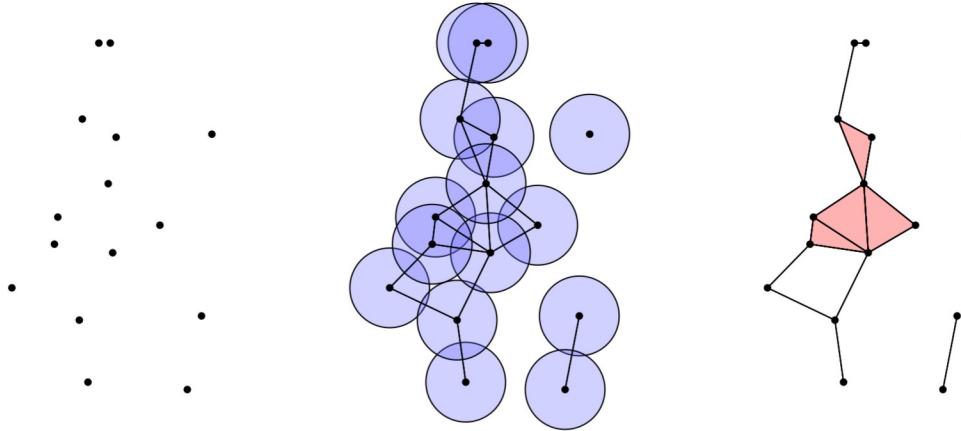


Figure 2: The process of constructing a VR complex from some data.

This process of constructing a simplicial complex for some data attributes shape to the data, by defining connections between arbitrarily close points. Before moving on to homology, which is an algebraic construct, a method of converting the geometric simplicial complex into an algebraic simplicial complex must first be considered. An abstract simplicial complex represents a method of defining the geometric connections, so that algebra can then be applied to the shape.

Definition 2.13. An *abstract simplicial complex*, K , consists of a set of vertices, $\text{vert}(K)$ and a set of abstract simplices, $\text{simp}(K)$, such that [11]:

1. Every simplex $\Delta^k \in \text{simp}(K)$ is a non-empty subset of $\text{vert}(K)$, or the simplices are a union of the vertices.
2. For all vertices, $v \in \text{vert}(K)$, there is also an abstract simplex $\{v\} \in \text{simp}(K)$.
3. For every non-empty abstract simplex, Δ^k , a non-empty subset of Δ^k is also an abstract simplex. This is referred to as a *face* of Δ^k .
4. For an abstract simplex $\Delta^k \in \text{vert}(K)$, the dimension of Δ^k is $\dim(\Delta^k) = |\Delta^k| - 1$, where $|.|$ denotes the number of elements in the set.

This notion of an abstract simplicial complex is pivotal to TDA. Abstract simplicial complexes allow for the mapping between categories: a mapping from a geometric realisation of a simplicial complex which is embedded in Euclidean space, to an abstract simplicial complex. The abstract simplicial complex form can be algorithmically analysed with computer packages to output results for that simplicial complex.

2.4 Homology

The homology groups, $H_k(X)$ are invariants for the data set, X , where k refers to the dimension of the feature of the homology group. The homology groups encode information about the number of k -dimensional holes in the data.

Definition 2.14. The *orientation* of the vertices of a simplex, Δ^k , is an equivalence class of orderings of the vertices under the equivalence relation that two orderings are the same if they differ by an *even permutation*. An even permutation is one that can be expressed as a composition of even permutations [7]. There are only two possible orientations, a positive and a negative.

Definition 2.15. To each standard simplex Δ_i^k of a simplicial complex K , an abelian group, $C_k(K)$, called a *chain group* is associated to it. The k^{th} chain group is the set of all k -simplices in the simplicial complex K [12].

Now, consider a K containing l_k of any standard simplex Δ^k for all values of k . The k -chain group of K , $C_k(K)$ is the free abelian group generated by the oriented k -simplices of K . This means any element $c_k \in C_k(K)$ can be thought of in abstract as,

$$c_k = \sum_{i=1}^{l_k} f_i \Delta_i^k, \quad f_i \in \mathbb{Z}$$

where the following criteria are satisfied [8]:

1. A negation of simplices. $\Delta_i^k + (-\Delta_i^k) = 0, \forall i, k.$

2. A linearity over the elements. $\sum_{i=1}^{l_k} f_i \Delta_i^k + \sum_{i=1}^{l_k} g_i \Delta_i^k = \sum_{i=1}^{l_k} (f_i + g_i) \Delta_i^k$ where $f_i, g_i \in \mathbb{Z}.$

Following the previous two definitions, a boundary operator can be formulated. The boundary operator provides an algebraic formulation for the exterior region of a simplicial complex. Now that simplicial complexes are being represented by chain groups, the problem has become entirely algebraic.

Definition 2.16. The *boundary operator*, ∂_k , maps between chain groups,

$$\partial_k: C_k(K) \rightarrow C_{k-1}(K)$$

Given an oriented simplex $\Delta^k = [v_0, \dots, v_k]$, a positive sign is assigned to every member of the even permutation class of Δ^k and a negative sign to every member of the odd permutation class. The boundary operator must now obey the rules [8]:

1. For an oriented simplex,

$$\partial \Delta^k = \sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k]$$

where $[v_0, \dots, \hat{v}_i, \dots, v_k]$ represents the face of the simplex with the i^{th} vertex omitted. Note that every successive omission changes the orientation of the face.

2. Thinking of a simplicial complex K , as the abstract sum of all the standard simplices required to construct it, $K = \sum_{i,k} \Delta_i^k$, $\Delta_i^k \subset K$. The boundary operator is a linear function over all the simplices in the complex.

$$\partial(K) = \partial\left(\sum_{i,k} \Delta_i^k\right) = \sum_{i,k} \partial(\Delta_i^k), \quad \Delta_i^k \subseteq K$$

Definition 2.17. A mapping between successive chain groups with the boundary operator,

$$\dots \rightarrow C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

is called a *chain complex* [12].

For the purposes of calculations, it is said that $C_i(K) = 0$ for $i > \dim(K)$. It is also the case that $C_i(K) = 0$ for $i < 0$. In both these cases, 0 represents the zero group.

Definition 2.18. For a simplicial complex K , elements of the chain group, $z_k \in C_k(K)$, are called *k-cycles* if $\partial z_k = 0$. The group of *k-cycles*, $Z_k(K)$, is given by the kernel of the boundary map,

$$Z_k(K) = \ker(\partial_k: C_k(K) \rightarrow C_{k-1}(K)) = \{z_k \in C_k: \partial_k z_k = 0\}$$

and $Z_k(K)$ is a subgroup of $C_k(K)$ [13].

Definition 2.19. For a simplicial complex K , elements of the chain group, $b_k \in C_k(K)$, are called *k-boundaries* if there exists a $(k+1)$ -chain group, $C_{k+1}(K)$, such that $\partial C_{k+1}(K) = b_k$. The group of *k-boundaries*, $B_k(K)$, is given by,

$$B_k(K) = \text{im}(\partial_{k+1}: C_{k+1}(K) \rightarrow C_k(K)) = \{b_k \in C_k: \exists b'_k \in C_{k+1}, b_k = \partial b'_k\}$$

and $B_k(K)$ is a subgroup of $C_k(K)$ [13].

The interesting part of this analysis, which fundamentally results in the homology groups, is that when the composition of two boundary operations is analysed, applying it twice gives a result of 0. That is the boundary of a boundary is empty.

Lemma 2.20.

$$\partial_{k-1} \circ \partial_k = 0$$

This is a very important result, a proof is supplied in many texts [6, 8]

Corollary 2.21. A very important result follows on from Lemma 2.20. For a simplicial complex, K , any element of the boundary group $b_k \in B_k(K)$ has the property $\partial_k b_k = 0$. Therefore, $B_k(K) \subseteq Z_k(K)$ where $Z_k(K)$ is the group of k -cycles. Since both $Z_k(K)$ and $B_k(K)$ are abelian, a property inherited by being subgroups of $C_k(K)$. $B_k(K)$ is a normal subgroup of $Z_k(K)$, and therefore it can divide the parent group.

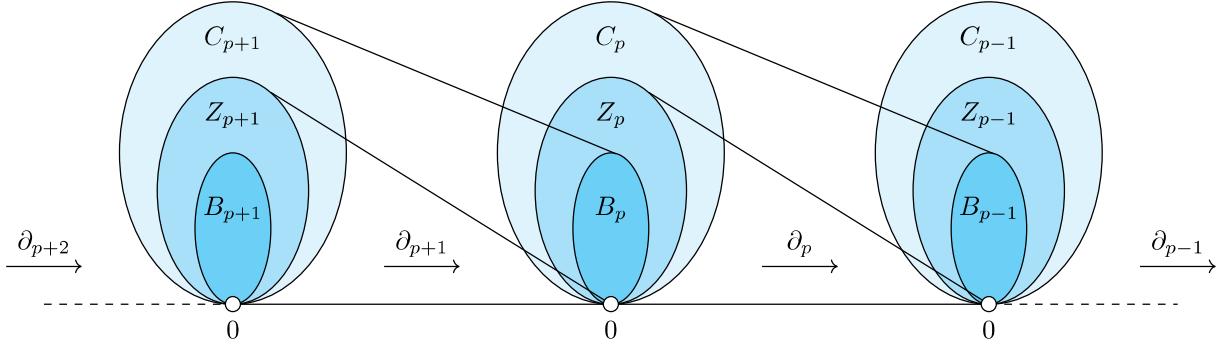


Figure 3: A chain complex with boundary maps between the chain groups.

Corollary 2.21 brings together a lot of the background theory that has been covered up until this point. It is beneficial to break this down into bite-size chunks. The first stage is understanding that $\partial_k b_k = 0$, as this is essentially the application of two boundary operators, as the element b_k is formed from the boundary operation of a previous chain element. This can be rewritten as $(\partial_k \circ \partial_{k+1})c_{k+1}$, then using Lemma 2.20 it follows that this result is equal to zero. If all the elements in $B_k(K)$ are mapped to zero after the application of ∂_k then it must be a subset of the group that is mapped to zero, by Definition 2.18 this is $Z_k(K)$. From Definitions 2.18 and 2.19, it is known that $Z_k(K) \subset C_k(K)$ and $B_k(K) \subset C_k(K)$. Since the chain groups $C_k(K)$ are defined to be abelian, then so are their subgroups, $B_k(K)$ and $Z_k(K)$ by Definition 2.3. It has also been proven that $B_k(K) \subset Z_k(K)$, and since $B_k(K)$ is abelian it is known that $B_k(K) \triangleleft Z_k(K)$ and therefore the set of cosets $Z_k(K)/B_k(K)$ is the quotient group.

Definition 2.22. The *homology groups*, $H_k(K)$, are the quotient groups,

$$H_k = Z_k(K)/B_k(K)$$

Setting aside all this mathematical rigour, the k^{th} homology groups can simply be thought of as the cycles in C_k that are not boundaries of the elements within C_{k+1} . That an element of C_k is a cycle, means it encloses a k -dimensional region. The fact this is not a boundary means the interior is not part of the underlying space. This is where the idea of counting the k -dimensional holes springs from. A generalisation of the rule is that $H_k(K)$ counts the k -dimensional holes in the simplicial complex, K . $H_0(K)$ is the only real exception to this rule, as this encodes information about the number of path-connected components in K . $H_1(K)$ encodes information about 1D holes; these can be visualised as circular holes. $H_2(K)$ encodes information about 2D holes; these can be visualised as cavities. $H_k(K)$ encodes information about kD holes.

Definition 2.23. $H_k(K)$ is a vector space and the elements are the *homology classes* of K . The homology class of a cycle $z_k \in Z_k(K)$ is the coset $c_k + B_k(K) = \{c_k + b_k : b_k \in B_k(K)\}$. Cycles are said to be *homologous* if they are in the same homology class [13].

Definition 2.24. The *Betti number*, β_k , is the dimension of the k^{th} homology group of a simplicial complex,

$$\beta_k = \dim(H_k(K))$$

The Betti number represents the number of k -dimensional holes in a simplicial complex [9].

Betti numbers will be the primary topological invariants used throughout this paper. Betti numbers will be used in persistent homology and they are vital in visualising spaces. Examples of Betti numbers and how they can be visualised are provided in Section 4. Following this, Betti numbers will be used in the engineering examples in Section 5.

3 Persistent Homology

3.1 Understanding

When the distance, ϵ , is smaller than some feature scale, the properties of that feature can be captured. This means topological invariants that are described at a length scale greater than ϵ can be captured. A problem arises here, as usually the feature scale is not known prior to analysis.

Obtaining the homology for a single value of ϵ provides very limited information, this notion is almost redundant, due to potential varying feature length scales in the manifold. For this reason, it is vital to consider what homological features persist as ϵ is varied. The goal of *persistent homology* is to track the homology classes as ϵ is varied. This process of varying ϵ does not bias any disk size, as all are being considered. This process will give an initial value, ϵ_{\min} , where a specific homological feature comes to life and ϵ_{\max} , where the feature is no longer considered for that simplicial complex. This range of values $[\epsilon_{\min}, \epsilon_{\max}]$ is called the *persistence interval* for that homological feature. Each persistence interval is attributed a Betti number. Following this, the set of all persistence intervals is descriptive for that manifold, giving information about in which dimension a hole exists in the data and over what range of values it persists for. Regardless of triangulation of a simplicial complex, or construction method, the information obtained from persistence is the same; i.e. the persistence of a space is considered a topological invariant.

When varying ϵ , the simplicial complex goes from k disconnected vertices to a fully connected $(k - 1)$ -dimensional simplex (the filtration is likely to not be taken this far, as computational time would be unnecessarily large), where there are k points sampled from a manifold. The features that persist over the longest relative change are more likely to represent features of the space. This procedure will then give a strong indication of the likelihood of the topological invariants associated to that space.

When calculating persistence, a minimum persistence interval length is specified. Given a persistence interval $[\epsilon_i, \epsilon_j]$, the length of the interval can be calculated by $|\epsilon_j - \epsilon_i| > l$ where l represents the interval length threshold. Interval lengths shorter than l will not be considered in the analysis; this is useful as it could be argued that for very small persistence intervals, whether they even 'persist' or not.

The persistence intervals obtained can be represented in two ways: barcodes or persistence diagrams, both having their merits.

3.2 Diagrams

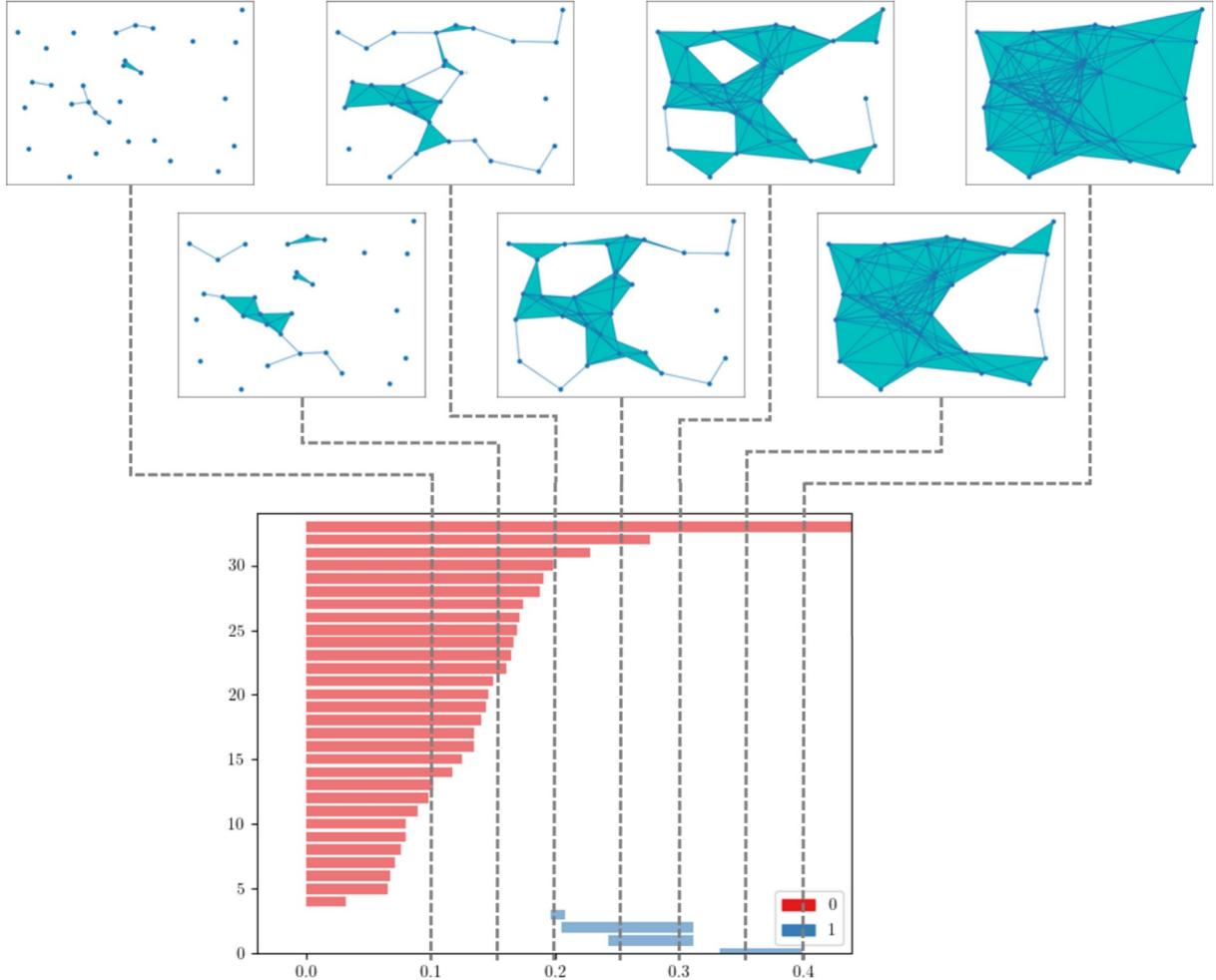


Figure 4: Persistence barcode with realisations showing which simplicial complexes are present at some values of ϵ .

In a barcode representation, the x -axis refers to the value of ϵ . As ϵ increases, the barcode shows which features persist. The set of intervals are plotted with each interval beginning at ϵ_{\min} and ending at ϵ_{\max} . The colour of the interval on the barcode refers to the Betti number, β_k [14]. The value of the y -axis can simply thought as an indexing of the intervals in the barcode. An example of a barcode can be seen in Figure 4, with vertical dotted lines showing the intersections with the intervals, showing which features are present and the corresponding simplicial complex is found at the end of the dotted line. A few notable tricks to reading the barcodes are: the length of the interval represents how long the feature persists for. The length of the interval can be thought of as having a higher probability that this feature is characteristic of the manifold, as the longer the interval, the more prominent that feature is in the data.

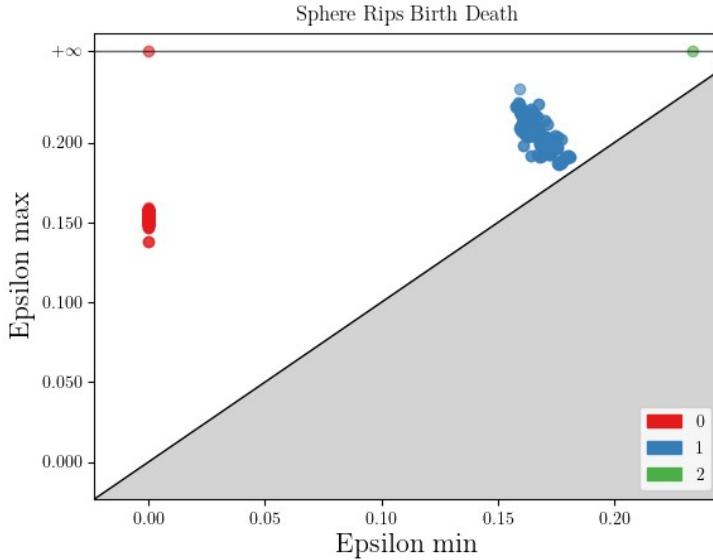


Figure 5: Birth-death diagram.

The other method of visualising the set of persistence intervals is the persistence diagram, or *birth-death* diagram. In this representation, ϵ_{\min} is plotted on the x -axis and ϵ_{\max} is plotted on the y -axis, with each interval represented by the point $(\epsilon_{\min}, \epsilon_{\max})$. Intuitively, there is a line defined by $y = x$ below which the points will not be plotted; this line has the interpretation that the feature must first exist before it can die. Reading these diagrams is as intuitive as reading the barcodes; the vertical height of the point from the line $y = x$ is analogous to the length of the interval, that is the further a point is from the line $y = x$ the more the feature persists. An example of a birth-death diagram can be seen in Figure 5. The Birth-Death diagram shows a more structured method of comparing persistence intervals as the y -axis is not so arbitrary when compared to the barcodes.

On both the barcode and birth-death diagram, it can be seen that features persist to ∞ . This has two meanings. The first being that there will always be a fully-connected simplex that persists to infinity. There will be a value of ϵ_{fc} that results in a fully-connected simplex where every vertex is connected to every other vertex. For values $\epsilon > \epsilon_{fc}$ the simplex will remain fully connected, and therefore this will continue to infinity. The second occurrence of this is when a feature persists past the value of ϵ used in the construction of the simplicial complex, as only smaller simplicial complexes of up to that value of ϵ are considered.

Each representation has their merits. For barcodes, it is easy to see which features a simplicial complex will have by drawing a vertical line at $x = \epsilon$, as can be seen in Figure 4. The barcode representation shows repeated intervals, as these are new entries. This is not the case with the birth-death diagram, these intervals will displayed as the same point and overlap, therefore, information is lost in the birth-death diagram in this situation. Despite this, birth-death diagrams are less arbitrary and are displayed more compactly.

The space of barcodes forms a metric space; the distance between the barcodes is a measure of similarity of two barcodes.

As manifolds can be represented by their barcodes, this notion of a metric space allows one to compare the similarity of manifolds. Metrics between barcodes are well established and the one used in this report is the p -Wasserstein distance.

Definition 3.1. Given two barcodes B_1 and B_2 . For $p > 0$, the p -Wasserstein distance is given by,

$$\partial_{W_p}(B_1, B_2) = \left(\inf \sum_{Z \in B_1} d_\infty(Z, \phi(Z))^p \right)^{\frac{1}{p}}$$

where ϕ is a matching between B_1 and B_2 and Z is a persistence interval in B_1 [7].

4 Understanding TDA

To help elaborate the point and highlight some features of TDA, two examples will be given in this section, where the points will be sampled from the manifold S^2 (the two-dimensional sphere embedded in \mathbb{R}^3) in two different ways. The first method used is by taking concentric circles in the plane perpendicular to the z -axis. The second method was to construct a *Fibonacci spiral* around the sphere. The formula used for the first method to generate the point cloud (the set of samples from the manifold) for the embedding of the sphere was,

$$f(u, v) = \begin{cases} x = \cos u \sin v \\ y = \sin u \cos v & u \in [0, 2\pi], v \in [0, \pi] \\ z = \cos v \end{cases}$$

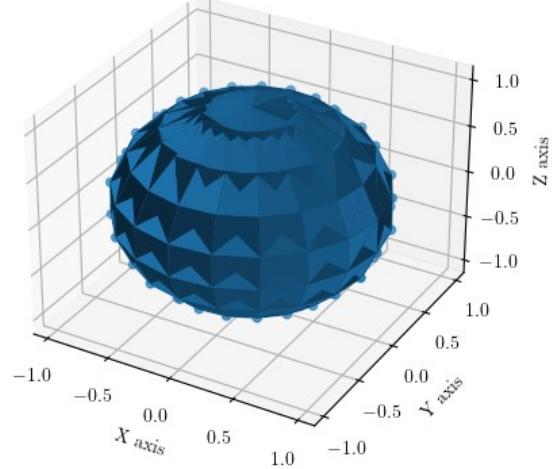


Figure 6: Sphere simplicial complex

When constructing a VR complex, simplices are formed at all points within an open ball B_ϵ ; this means that connections will be formed to points past other points in the same direction. In this example, this is problematic as it results in the shape being over-connected at the poles if one wants the equatorial region to be filled with 2-simplices. This negatively impacts the computation time.

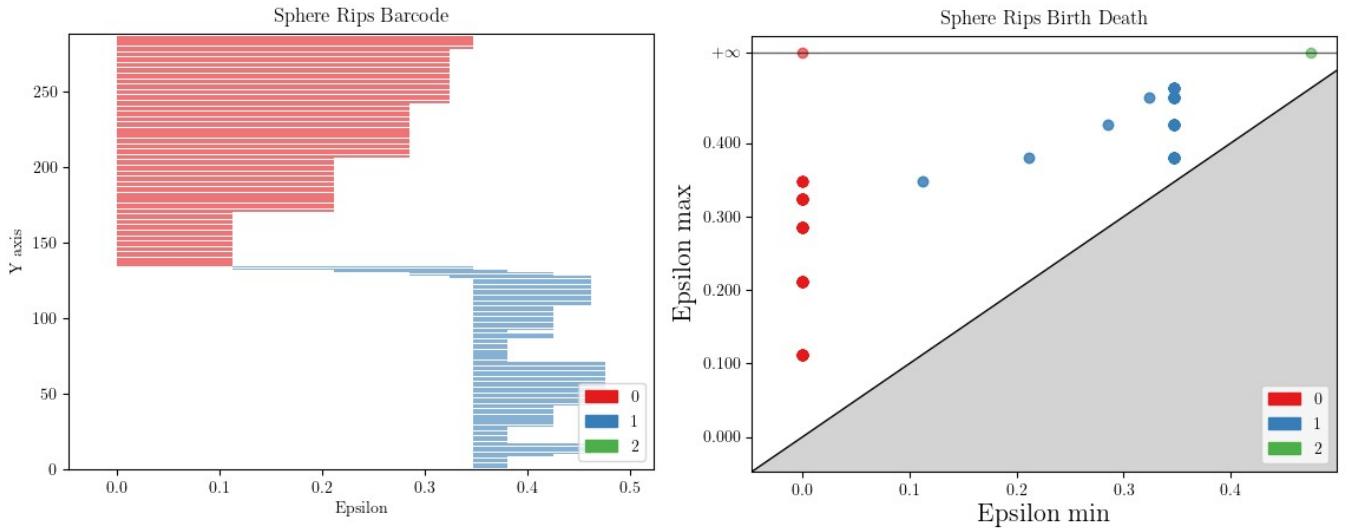


Figure 7: Sphere persistence representations.

The data and values for the given value of ϵ , for this sphere and the Fibonacci sphere is shown in Table 1. It should be noted that the values of the Betti numbers in the table are specific to that value of epsilon. The diagrams contain a greater depth of information as the value of epsilon can be seen to be varied.

The Betti numbers in Table 1 are representative of the simplicial complex shown in Figure 7. These are $\beta_0 = 1$, which says that the simplicial complex is fully connected with no disjoint parts. $\beta_1 = 0$, tells us that there are no holes in

the simplicial complex. $\beta_2 = 1$ says that the simplicial complex enclose a volume, which is the inside of the sphere. It is important to note that the inside of the sphere is not part of the simplicial complex, but the space in which the simplicial complex is embedded.

The second case of sampling from the sphere, was done with *Fibonacci spirals* around the sphere; this gives a much more uniform distribution of points. The formula used to generate the point cloud for the embedding of the sphere was,

$$f(\phi, \theta) = \begin{cases} x = \cos \theta \sin \phi \\ y = \sin \theta \sin \phi & \phi \in [0, \pi], \theta \in [0, n_p \pi(1+\sqrt{5})] \\ z = \cos \phi \end{cases}$$

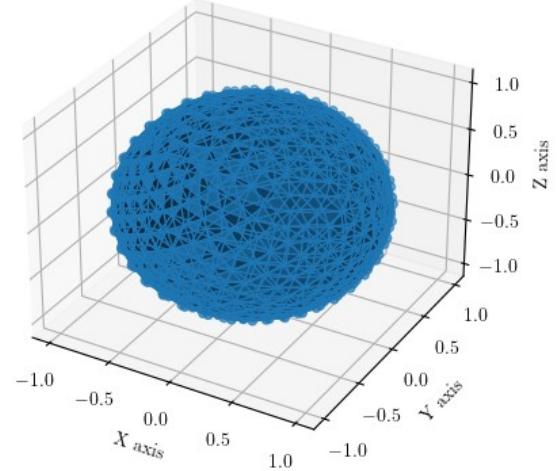


Figure 8: Fibonacci sphere simplicial complex.

where n_p is the number of points being sampled, in this case $n_p = 500$.

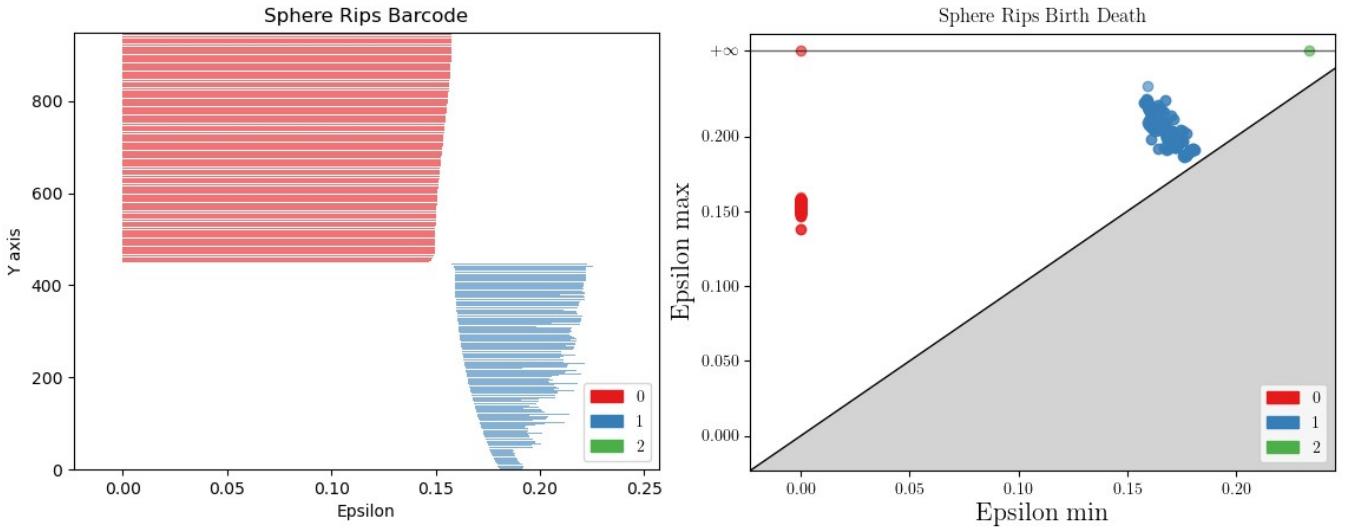


Figure 9: Fibonacci sphere persistence representations.

From Table 1 it can be seen that the first scenario has fewer vertices, but still results in more simplices after constructing the simplicial complex. This is due to the poor distribution of points and the larger epsilon. The example goes to show that data cleansing steps and the correct model should be used to output meaningful and efficient results from TDA.

Description	Sphere	Fibonacci sphere
ϵ	0.5	0.25
Number of vertices	200	500
Number of simplices	112094	4202
Betti Numbers	[1, 0, 1]	[1, 0, 1]
Dimension	3	3

Table 1: Topological properties for each sphere.

When using the Wasserstein metric between the two data samples, a relatively small value of $\partial_{W_p} = 2.673$ is obtained; this shows that the manifolds are relatively similar. This result is to be expected as the two spheres are sampled from the same manifold. When qualitatively analysing the persistence data, this backs up the case as the barcodes and birth-death diagrams, whilst not the same, do exhibit very similar features.

5 Application

In this case, a four-dimensional surface plot will be considered. This example is defined by linear algebraic equations, where the variables form the axes of the higher-dimensional plots.

The system under analysis is a simple 3DOF system, with springs between masses and ground. A representation of a general case can be seen in Figure 10. In this modelling scenario, it is assumed that the stiffness of the second spring, k_2 is a function of the coefficient of thermal expansion, α_2 , the temperature of the spring, T_2 , and the damage present in the spring, D_2 . It is the aim of this analysis to understand the topology of the manifold constructed when the natural frequency ω_i is calculated as a function of the variable parameters α_2, T_2, D_2 and all other values are constant.

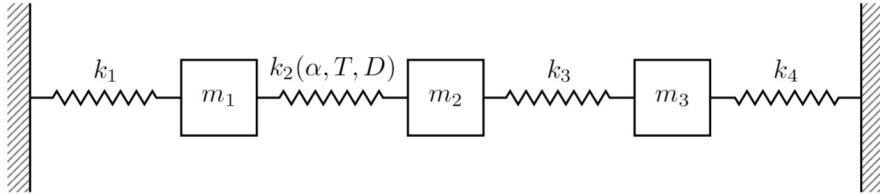


Figure 10: General mass-spring-damper system with variable k_2 .

A simple relationship is assumed, where an increase in the temperature of a spring causes a reduction of the stiffness, by the relation $k_2(1 - \alpha_2 T_2)$. It is also assumed that the damage coefficient also reduces the stiffness of the spring by a factor $k_2(1 - D_2)$.

After undertaking the appropriate free body analysis, with the assumptions that there is no damping and the system has no external forcing. The following equation is derived,

$$\mathbf{M}\ddot{\mathbf{X}} + \mathbf{K}\mathbf{X} = 0$$

where,

$$\mathbf{M} = \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} k_1 + k_2(1 - \alpha_2 T)(1 - D) & -k_2(1 - \alpha_2 T)(1 - D) & 0 \\ -k_2(1 - \alpha_2 T)(1 - D) & k_2(1 - \alpha_2 T)(1 - D) + k_3 & -k_3 \\ 0 & -k_3 & k_3 + k_4 \end{bmatrix}$$

Assuming a harmonic response, a solution can be obtained by determining the eigenvalues of the equation $|\mathbf{-M}\omega_i^2 + \mathbf{K}| = 0$. The result is a matrix equation that determines the response of the system, where ω_i is a function of the parameters $\omega_i(\alpha_2, D_2, T)$. This is a 3DOF system, therefore, it is possible to obtain three eigenvalues (ω_i^2) and eigenvectors (mode shapes), X_i , where the eigenvalues and eigenvectors satisfy the equation,

$$\mathbf{M}^{-1}\mathbf{K}X = \omega_i^2 X, \quad i \in \{1, 2, 3\}$$

5.1 TDA of Mass-Spring-Damper Model

This example is embedded in \mathbb{R}^4 ; for this reason the topological features shown in the persistence barcodes will have to be studied to determine the topology of the data, rather than visualisation. In this case, analysis is being undertaken to determine the topology of the manifold from which the data are sampled. This was not the case with the sphere, as the topology was known prior to analysis. Despite this, it is still possible to determine some features of the topology of the data. It is known that the manifold is connected, as it is described by continuous functions. Therefore, the condition that $\beta_0 = 1$ must be satisfied.

It is also expected that there is a self-intersection in the manifold. This can be inferred from the model, where an increase in temperature would reduce the spring to become less stiff. This is also the case for the damage parameter. This must mean that there is a value of αT that will give the same reduction in k , that a value of D will. Since all the variables are contained within \mathbf{K} , this means that the natural frequency must be the same value at these two points, and thus there is a self intersection in the manifold. It is actually the case that this is a family of values, and not just a single point. The family of values is parameterised by,

$$(1 - \alpha_2 T)(1 - D_2) = (1 - \alpha'_2 T')(1 - D'_2)$$

When creating the data, the inputs were $T \in [250, 500]$ with seven divisions, $\alpha_2 \in [0, 0.005]$ with six divisions and $D \in [0, 1]$ with six divisions; meaning that there are 252 data points. The values for T , α and D can be thought of as coordinates of the form (T, α, D) . Each point is then input into the equations above to give a value ω_i . Each point can then be embedded in 4D in the form (T, α, D, ω_i) . These points will then form the vertices of the simplicial complex, which can be analysed using TDA. When generating data this way, the number of points generated is proportional to the power of the dimension, so calculations can get cumbersome for high dimensions or high sample sizes. After the points had been generated, each point was scaled down by the largest value in that dimension, meaning that all the data are inside $[0, 1]^4$.

The values used for the calculations in this first example were:

$$m_1 = m_2 = m_3 = 10\text{kg}, k_1 = k_2 = k_3 = k_4 = 10000\text{Nm}^{-1}$$

Description	ω_1	ω_2	ω_3
Epsilon	$\epsilon_1 = 0.77$	$\epsilon_2 = 0.33$	$\epsilon_3 = 0.31$
Num of vert	252	252	252
Num of Simp	188087202	160639	93917
Betti Numbers	[1, 0, 0, 0]	[1, 0, 0, 0]	[1, 0, 0, 0]
Dimension	4	4	4

Table 2: $\omega_{1,2,3}$ simplicial complexes data.

The Betti numbers in Table 2 are relatively uninteresting, but these Betti numbers are for only for the values of epsilon listed. The persistence diagrams in Figures 11, 12 and 13 give more insight into the features for values of epsilon less than the ones listed in Table 2.

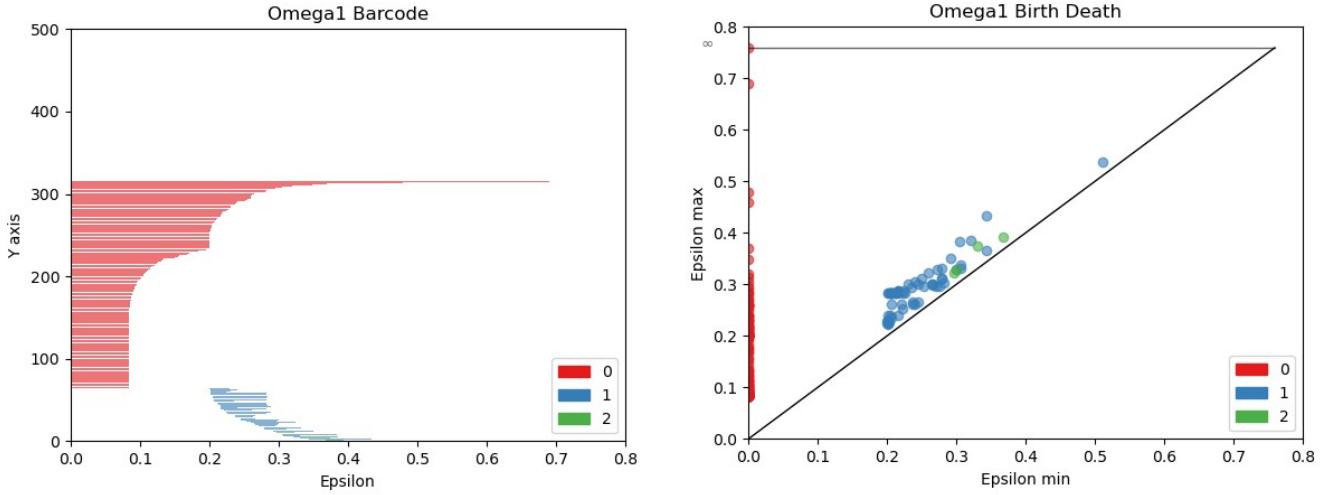


Figure 11: ω_1 persistence representations.

In the case of ω_1 , the manifold appears to have a sudden and sharp change with the parameters; this can be seen in the birth-death diagram in Figure 11, as the second to last red point $\approx (0, 0.69)$ is the instance at which the simplicial complex is fully connected. It appears that the data are roughly divided into two clusters either side of this sudden change. A high value of epsilon is required to span this sudden change. As a result of a high value of epsilon, the clusters become highly connected within. This can be seen in the relatively large number of simplices in Table 2 when compared to the other natural frequencies.

Another method of spanning the gap between the sudden change, instead of increasing ϵ_1 , is to increase the resolution of the data by taking more divisions for each parameter. This method also has an issue. More points are added to the sudden change, but there will also be more points in each cluster. Meaning there are more points to be connected; this massively compromises the computation time.

It can also be seen on the barcode in Figure 11 that there are numerous instances of 1D holes that persist for about 0.1 units in epsilon. As features only persist for small ranges of epsilon, this is likely caused by the lack of resolution in the data. In the case of ω_1 , there are also instances of cavities forming. These features also do not persist for long, and are likely artefacts due to the lack of resolution in the data.

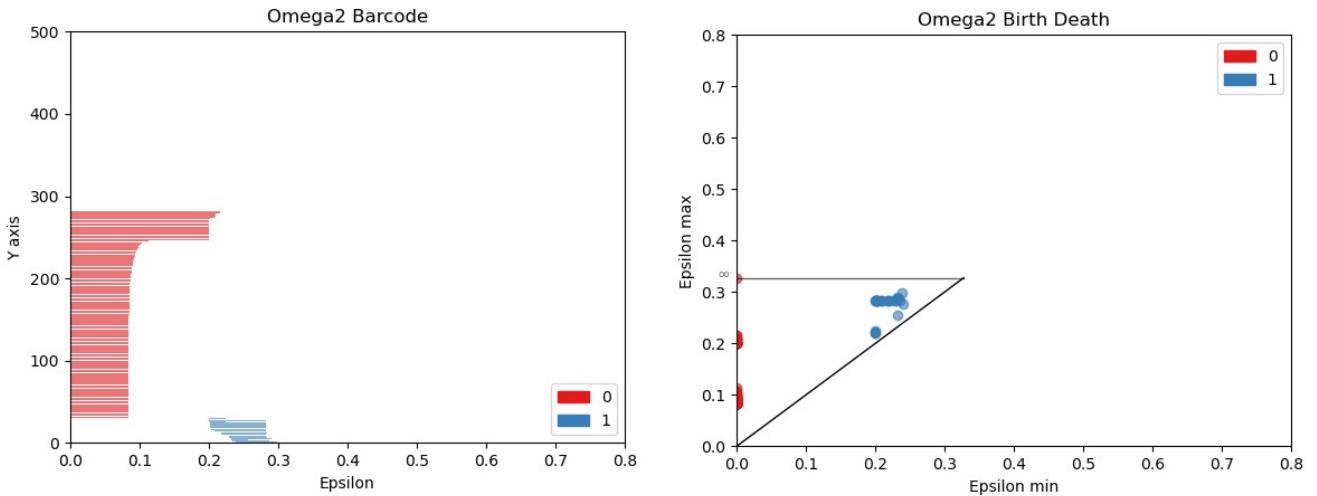


Figure 12: ω_2 persistence representations.

The persistence representations shown in Figures 12 and 13 are fairly similar, both having nearly the same features present.

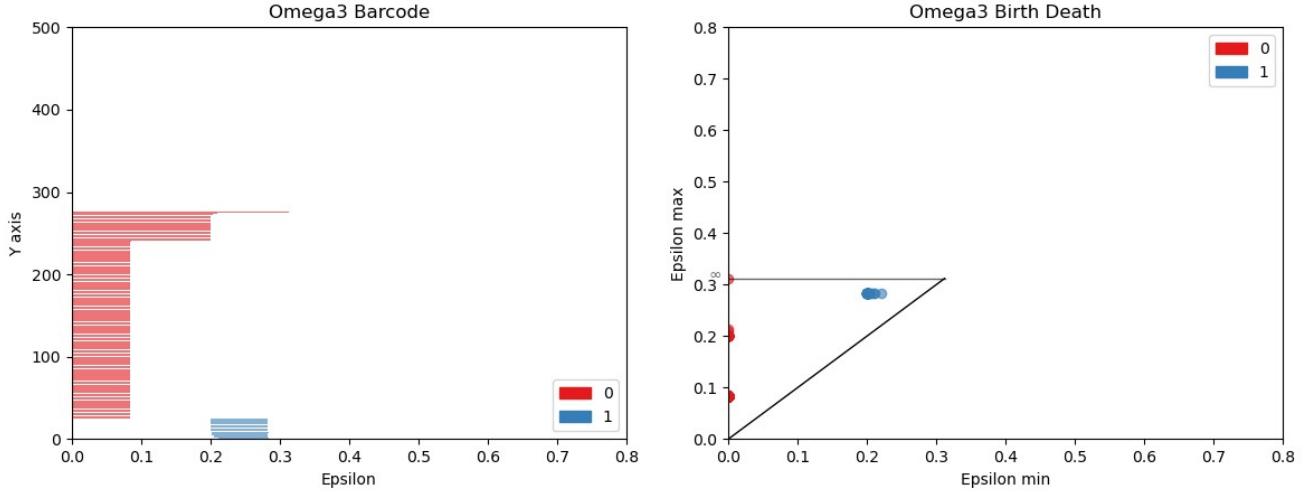


Figure 13: ω_3 persistence representations.

Both ω_2, ω_3 achieved the criteria of $\beta_0 = 1$ at a much lower value of ϵ when compared with ω_1 , which indicates there is no sudden abrupt change in the manifold. There are also 1D holes present, again this is likely to be an artefact of the lack of resolution in the data.

There is also a general trend of the persistence intervals becoming more square in the β_0 section of the barcode, going from ω_1 to ω_3 . This says that the distance between the connected components is becoming more uniform.

5.1.1 Values at different stiffness

Following the previously-outlined procedure, a new data set was collected for the values,

$$m_1 = m_2 = m_3 = 10\text{kg}, k_2 = 5000\text{Nm}^{-1}, k_1 = k_3 = k_4 = 10000\text{Nm}^{-1}$$

For continuity, the same number of divisions was used along each of the inputs, and the same process of selecting the lowest value of epsilon that gave a fully-connected simplicial complex. Following these conditions, the results in Table 3 were obtained.

Description	ω'_1	ω'_2	ω'_3
Epsilon	$\epsilon'_1 = 0.38$	$\epsilon'_2 = 0.32$	$\epsilon'_3 = 0.30$
Num of vert	252	252	252
Num of Simp	339447	87571	97341
Betti Numbers	[1, 0, 0, 0]	[1, 0, 0, 0]	[1, 0, 0, 0]
Dimension	4	4	4

Table 3: $\omega'_{1,2,3}$, $k_2 = 5000\text{Nm}^{-1}$ simplicial complexes data.

Immediately it is clear that the ϵ'_1 is much smaller than ϵ_1 ; this means that the computation time was much less when calculating the simplicial complex for $k_2 = 5000\text{Nm}^{-1}$ than $k_2 = 10000\text{Nm}^{-1}$.

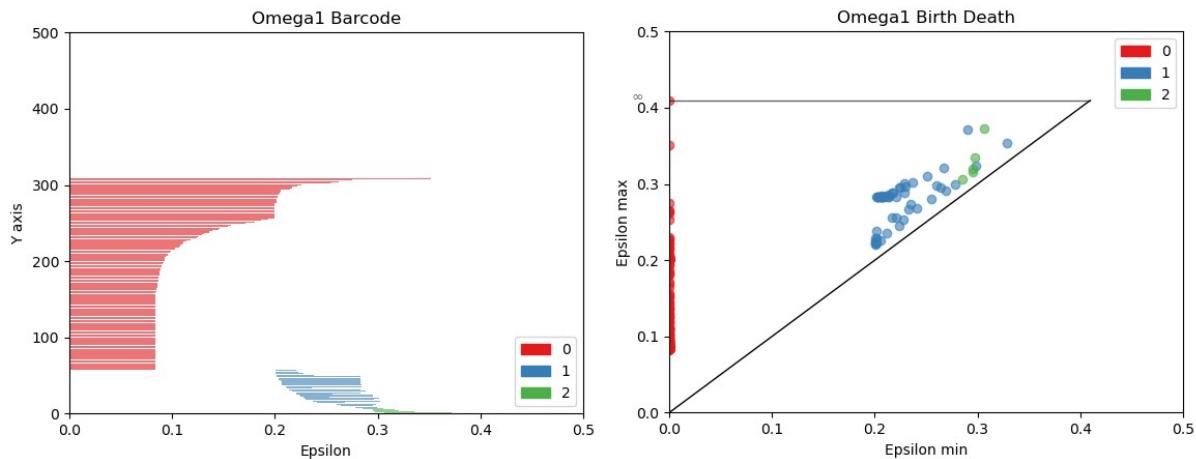


Figure 14: ω'_1 persistence representations.

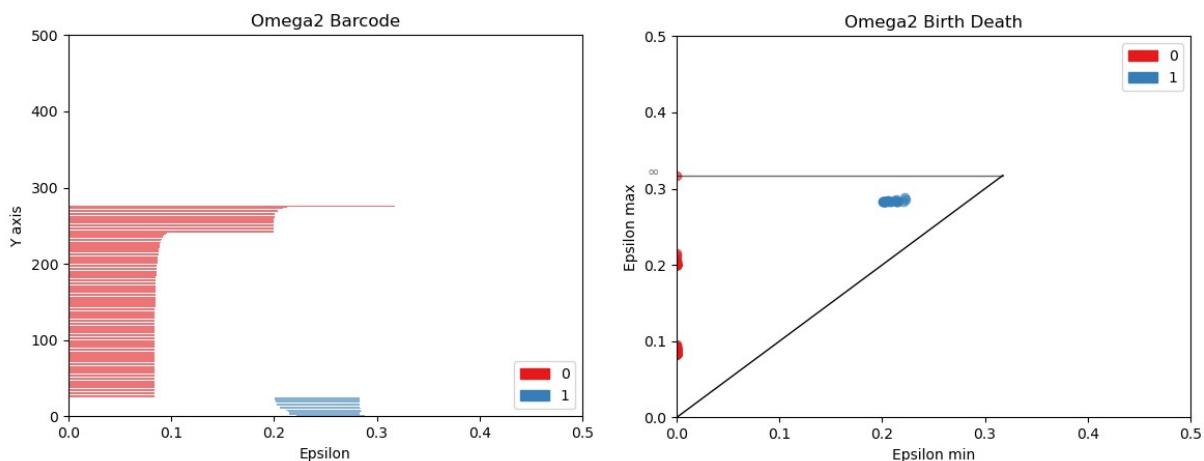


Figure 15: ω'_2 persistence representations.

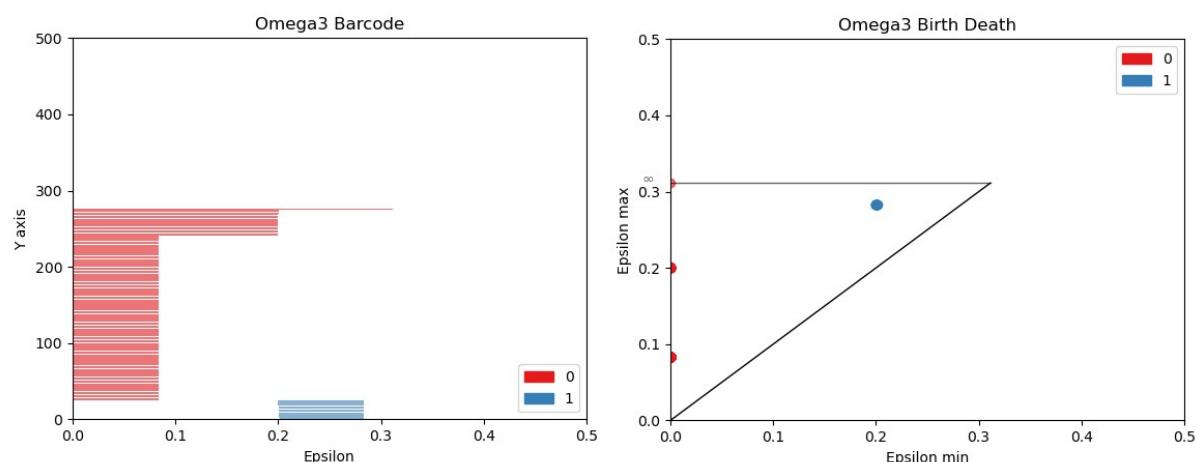


Figure 16: ω'_3 persistence representations.

The barcodes still exhibit a similar structure when compared with their counterparts in Figures 11, 12 and 13. When comparing the barcodes using the Wasserstein distance, the following values were obtained,

$$\partial_{W_p}(B_{\omega_1}, B_{\omega'_1}) = 0.69029, \quad \partial_{W_p}(B_{\omega_2}, B_{\omega'_2}) = 0.077248, \quad \partial_{W_p}(B_{\omega_3}, B_{\omega'_3}) = 0.029308$$

The values may appear relatively small, but the data were scaled by the largest value in that dimension, therefore, the Wasserstein distance is also scaled. There appears to be a general trend, that the higher the natural frequency, the more similar the manifolds are. From this, it can be inferred for this example, that as successive natural frequencies are computed there is a lesser reliance on the stiffness term, k_2 .

The idea of comparing two persistence data sets is useful in engineering. If the manifold structure of data from a machine or structure operating at optimal conditions is known, the shape of the manifold could be used to fine tune or identify differences in the shape of a manifold of similar machine or structure in an unknown state. This could have interesting applications in SHM.

6 Conclusions

The results of this paper show that topological inference is a viable analysis strategy for engineering applications. By opting to use TDA, features in the data can be identified that would not necessarily be found in other analysis techniques. Features of the data such as holes or cavities in the underlying manifold can help to identify trends or anomalies. This inference can then be extended to determine engineering examples such as clustering data by their topological features, this can then be used to identify changes in the operating state of a machine or structure.

An appropriate method for conducting topological data analysis has been outlined, with a special consideration for conducting the analysis in an effective manner. Examples were given that highlighted where TDA can be inefficient or return misleading results.

The engineering-specific example given here is a very fundamental start, but not without its issues. Calculating data at higher dimensions is computationally expensive, and therefore the resolution in the manifold was too low. However, this is not the case with real world data that has been collected experimentally. Another issue arises when data are not evenly distributed, as this results in overly-connected areas.

In summary, a novel data analysis strategy has been discussed, previously unfamiliar to engineering, that can provide new insights into the fundamental structure and shape of the data, even in higher-dimensional analysis where an intuitive understanding of shape is no longer upheld.

Acknowledgements

The authors would like to thank the UK EPSRC for funding via the Established Career Fellowship EP/R003645/1 and the Programme Grant EP/R006768/1.

References

- [1] C R Farrar and K Worden. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*.
- [2] C R Farrar and K Worden. *Structural health monitoring: a machine learning perspective*. John Wiley & Sons, 2012.
- [3] H Edelsbrunner and J Harer. *Computational Topology: An Introduction*. American Mathematical Soc., 2010.
- [4] H Edelsbrunner, D Letscher, and A Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE, 2000.
- [5] J J Rotman. *An Introduction to the Theory of Groups*, volume 148. Springer Science & Business Media, 2012.
- [6] B F Schutz and Director Bernard F Schutz. *Geometrical Methods of Mathematical Physics*. Cambridge university press, 1980.

- [7] R Rabadán and A J Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019.
- [8] C Nash and S Sen. *Topology and Geometry for Physicists*. Elsevier, 1988.
- [9] R W Ghrist. *Elementary Applied Topology*, volume 1. Createspace Seattle, 2014.
- [10] E W Chambers, V De Silva, J Erickson, and R Ghrist. Vietoris-rips complexes of planar point sets. *Discrete & Computational Geometry*, 44(1):75–90, 2010.
- [11] F Chazal and B Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.
- [12] R Ghrist. Homological algebra and data. *The Mathematics of Data*, 25:273, 2018.
- [13] J D Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and Topological Inference*, volume 57. Cambridge University Press, 2018.
- [14] R Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.