Acta Numerica (2014), pp. 289–368
doi:10.1017/S0962492914000051

Topological pattern recognition for point cloud data*

Gunnar Carlsson†
Department of Mathematics,
Stanford University,
CA 94305, USA
E-mail: gunnar@math.stanford.edu

In this paper we discuss the adaptation of the methods of homology from algebraic topology to the problem of pattern recognition in point cloud data sets. The method is referred to as *persistent homology*, and has numerous applications to scientific problems. We discuss the definition and computation of homology in the standard setting of simplicial complexes and topological spaces, then show how one can obtain useful signatures, called barcodes, from finite metric spaces, thought of as sampled from a continuous object. We present several different cases where persistent homology is used, to illustrate the different ways in which the method can be applied.

CONTENTS

1. Introduction

Deriving knowledge from large and complex data sets is a fundamental problem in modern science. All aspects of this problem need to be addressed by the mathematical and computational sciences. There are various different aspects to the problem, including devising methods for (a) storing massive amounts of data, (b) efficiently managing it, and (c) developing understanding of the data set. The past decade has seen a great deal of development of powerful computing infrastructure, as well as methodologies for

* Colour online for monochrome figures available at journals.cambridge.org/anu.
† Research supported in part by the National Science Foundation, the National Institutes of Health, and the Air Force Office of Scientific Research.

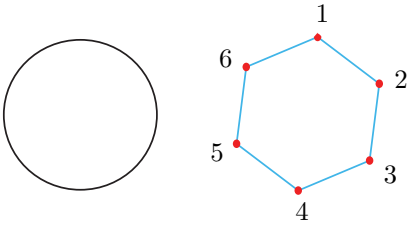


Figure 1.1. A compressed combinatorial representation of a circle.

managing and querying large databases in a distributed fashion. In this paper, we will be discussing one approach to (c) above, that is, to the problem of generating knowledge and understanding about large and complex data sets.

Much of mathematics can be characterized as the construction of methods for organizing infinite sets into understandable representations. Euclidean spaces are organized using the notions of vector spaces and affine spaces, and this allows us to organize the (infinite) underlying sets into understandable objects which can be readily manipulated, and which can be used to construct new objects from old in systematic ways. Similarly, the notion of an algebraic variety allows us to work effectively with the zero sets of sets of polynomials in many variables. The notion of shape is similarly encoded by the notion of a *metric space*, a set equipped with a distance function satisfying three simple axioms. This abstract notion permits us to study not only ordinary notions of shape in two and three dimensions but also higher-dimensional analogues, as well as objects such as the p-adic integers, which are not immediately recognized as being geometric in character. Thus, the notion of a metric serves as a useful organizing principle for mathematical objects. The approach we will describe demonstrates that the notion of metric spaces acts as an organizing principle for finite but large data sets as well.

Topology is one of the branches of mathematics which studies properties of shape. The study of shape particular to topology can be described in terms of three points.

(1) The properties of shape studied by topology are independent of any particular coordinate representation of the shape in question, and instead depend only on the pairwise distances between the points making up the shape.

(2) Topological properties of shape are *deformation invariant*, that is, they do not change if the shape is stretched or compressed. They would of course change if non-continuous transformations were applied, 'tearing' the space.

Figure 1.2. Annulus.

(3) Topology constructs compressed representations of shapes, which retain many interesting and useful qualitative features while ignoring some fine detail.

Topology deals with shape in two distinct ways. The first is by building compressed combinatorial representations of shapes, via processes such as triangulation. Of course, some information about the shape is lost in this discretization, such as fine-scale curvature information, but as in the example in Figure 1.1, the rough overall structure is preserved in passing from the circle to the hexagon. The second method is by attempting to measure shape, or aspects of shape. This is done via *homological signatures*, which essentially count occurrences of patterns within the shape. The adaptation of these signatures to the study of point cloud data is the subject of this paper.

The intuitive idea behind algebraic topology is that one should try to distinguish or perhaps even characterize spaces by the occurrences of patterns within a space. Consider the example of an annulus, in which one could say that a characteristic pattern is the presence of a loop in the space, surrounding the removed disc in the middle. One could say intuitively that the count of loops in an annulus is one, in that there is 'essentially' only one loop in the space, characterized by the fact that it winds around the central removed disc. It is not so easy to make mathematical sense of this observation, for reasons made clear in Figure 1.2.

The presence of essentially one loop is something which is difficult to quantify *a priori*, since in fact there is an uncountable infinity of actual loops which have the same behaviour, that is, they wind around the hole once. In order to resolve this difficulty, and formalize the notion that there is essentially only one loop, we are forced to perform some abstract constructions involving equivalence relations to obtain a sensible way of counting the number of loops. The idea is that one must regard many different loops as equivalent, in order to get a count of the occurrences not of each individual loop, but rather of a whole class of equivalent loops. This step is

what is responsible for much of the abstraction which is introduced into the subject. Once that layer of abstraction has been built, it provides a way to detect the presence of geometric patterns of certain types. The general idea of a pattern is of course a diffuse one, with many different meanings in many different contexts. In the geometric context, we define patterns as maps from a template space, such as a circle, into the space. A large part of the subject concerns the process of reducing the abstract constructions described above to much more concrete mathematical constructions, involving row and column operations on matrices. The goals of the present paper are as follows.

- To introduce the pattern detection signatures which come up in algebraic topology, and simultaneously to develop the matrix methods which make them into computable and usable invariants for various geometric problems, particularly in the domain of *point clouds* or *finite metric spaces*. We hope that the introduction of the relevant matrix algorithms will begin to bridge the gap between topology as practised 'by hand', and the computational world. We will describe the standard methods of homology, which attach a list of non-negative integers (called the Betti numbers) to any topological space, and also the adaptation of homology to a tool for the study of point clouds. This adaptation is called *persistent homology*.

- To introduce the mathematics surrounding the collection of *persistence barcodes* or *persistence diagrams*, which are the values taken by the persistent homology constructions. Unlike the Betti numbers, which are integer-valued, persistent homology takes its values in multisets of intervals on the real line. As such, they have a mix of continuous and discrete structure. The study of these spaces from various points of view, so as to be able to make them maximally useful in various problem domains, is one of the most important research directions within applied topology.

- To describe various examples of applications of persistent homology to various problem domains. There are two distinct directions of application, one being the study of homological invariants of individual data sets, and the other being the use of homological invariants in the study of databases where the data points themselves have geometric structure. In this case, the barcode space can act as the home for a kind of non-linear indexing for such databases.

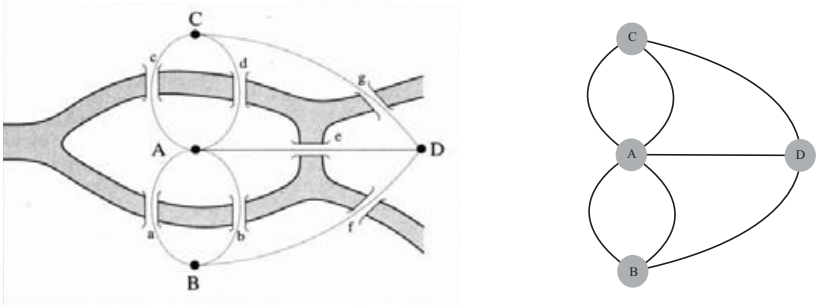


Figure 2.1. Euler's 'Bridges of Königsberg' problem.

2. Topology

2.1. History

Euler's paper of 1741, in which he studies the so-called 'Bridges of Königsberg' problem, is usually cited as the first paper in topology. The question he asked was whether it was possible to traverse all the bridges exactly once and return to one's starting point. Euler answered this question by recognizing that it was a question about paths in an associated network: see Figure 2.1. In fact, the question only depends on certain properties of the paths, independent of the rates at which the paths are traversed. His result concerned the properties of an infinite class of paths, or of a certain type of pattern in the network. Euler also derived *Euler's polyhedral formula*, relating the number of vertices, edges and faces in polyhedra (Euler 1758*a*, 1758*b*). The subject developed in a sporadic fashion over the next century and a half, including work by Vandermonde on knot theory (Vandermonde 1774), the proof of the Gauss–Bonnet theorem (never published by Gauss, but with a special case proved by Bonnet (1848)), the first book on the subject by Listing (1848), and the work of Riemann (1851) identifying the notion of a manifold. In 1895, Poincaré published his seminal paper in which the notions of homology and fundamental group were introduced, with motivation from celestial mechanics. The subject then developed at a greatly accelerated pace throughout the twentieth century. The first paper on persistent homology was published by Robins (1999), and the subject of applying topological methodologies to finite metric spaces has been developing rapidly since that time.

2.2. Equivalence relations

A (binary) relation on a set X is a subset of $X \times X$. We will often denote relations by $\sim$, and write $x \sim x'$ to indicate that (x, x') is in the relation.

Definition 2.1. A relation $\sim$ on a set X is an *equivalence relation* if the following three conditions hold:

(1) $x \sim x$ for all $x \in X$,

(2) $x \sim x'$ if and only if $x' \sim x$,

(3) $x \sim x'$ and $x' \sim x''$ implies $x \sim x''$.

By the *equivalence class* of $x \in X$, denoted by $[x]$, we will mean the set

$$\{x' |\ x \sim x'\}.$$

The sets $[x]$ for all $x \in X$ form a partition of the set X. If $\sim$ is any symmetric binary relation on a set X, then by the *equivalence relation generated by* $\sim$ (or the *transitive closure of* $\sim$) we will mean the equivalence relation $\sim'$ defined by the condition that $x_0 \sim' x_1$ if and only if there is a positive integer n and a sequence of elements $x'_0, x'_1, \ldots, x'_n$ so that $x'_0 = x_0$, $x'_n = x_1$, and $x'_i \sim x'_{i+1}$ for all $0 \leq i \leq n-1$.

For a set X and an equivalence relation $\sim$ on X, we will denote the set of equivalence classes under $\sim$ by $X/\sim$, and refer to it as the *quotient* of X with respect to $\sim$. There are several important special cases of this definition. The first is the quotient of a vector space by a subspace. Let V be a vector space over a field k, and let $W \subseteq V$ be a subspace. We define an equivalence relation $\sim_W$ on V by setting $v \sim_W v'$ if and only if $v - v' \in W$. It is easy to verify that $\sim_W$ is an equivalence relation. We can form the quotient $V/\sim_W$, and one observes that $V/\sim_W$ is itself naturally a vector space over k, with the addition and scalar multiplication rules satisfying $[v] + [v'] = [v + v']$ and $\kappa[v] = [\kappa v]$. In this special case, we will denote $V/\sim_W$ by V/W, and refer to it as the quotient space of V by W. Although the quotient is an apparently abstract concept, it can be described explicitly in a couple of ways.

Proposition 2.2. Suppose that we have a basis B of a vector space V, and $B' \subseteq B$ is a subset. If W is the subspace of V spanned by B', then the quotient V/W has the elements $\{[b] | b \notin b'\}$ as a basis, so the dimension of V/W is $\#(B) - \#(B')$. More generally, if W' is a complement to W in V, so that $W + W' = V$, and $W \cap W' = \{0\}$, then the composite

$$W' \hookrightarrow V \xrightarrow{p} V/W$$

is a bijective linear transformation, so the dimension of V/W is equal to the dimension of W', where p is the map which assigns to $v \in V$ its equivalence class $[v]$ under $\sim_W$.

There is also a matrix interpretation. Let V and W be vector spaces with ordered bases, and let $f : V \to W$ be a linear transformation, with matrix

$A(f)$ associated to the given bases. The *image* of f is a subspace of W, and we write $\theta(f)$ for the quotient space $W/\mathrm{im}(f)$.

Proposition 2.3. Let $g : V \to V$ and $h : W \to W$ be invertible linear transformations. Then $\theta(f)$ is isomorphic to $\theta(hfg)$. It follows that if we have the matrix equation $A(f') = A(h)A(f)A(g)$, then $\theta(f')$ is isomorphic to $\theta(hfg)$.

Proof. This follows from the elementary observation that $w \sim_{\mathrm{im}(f)} w'$ if and only if

$$h(w) \sim_{\mathrm{im}(hf)} h(w').$$ □

Proposition 2.4. Let W be the vector space k^m for some m, and suppose that we are given an $m \times n$-matrix A with entries in a field k. A can be regarded as a linear transformation from $V = k^n$ to W, and the span of the columns in this matrix is the image of the transformation A. Then, if we apply any row or column operation (permuting rows/columns, multiplying a row/column by a non-zero element of k, or adding a multiple of one row/column to another) to obtain a matrix A', then $\theta(A)$ is isomorphic to $\theta(A')$.

Remark 2.5. Note that for any matrix A over a field, one can apply row and column operations to bring it to the form

$$\begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix},$$

where n is the rank of A. In this case, the dimension of the quotient is readily computed using Proposition 2.2.

The second special case is that of the orbit set of a group action. If G is a group, and we have an action of G on a set X, then the action defines an equivalence relation $\sim_G$ on X by $x \sim_G x'$ if there is a $g \in G$ so that $gx = x'$. This is readily seen to be an equivalence relation, and the equivalence classes are called the *orbits* of the action.

Finally, consider the case of a topological space X equipped with an equivalence relation R. Then the quotient set X/R is equipped with a topology by declaring that a set $U \subseteq X/R$ is open if and only if $\pi^{-1}(U)$ is an open set in X, where $\pi : X \to X/R$ is the map which assigns to each $x \in X$ its equivalence class $[x]$.

2.3. Homotopy

The fundamental idea of algebraic topology is that one should develop methods for counting the occurrences of geometric patterns in a topological space in order to distinguish it from other spaces, or to suggest similarities between different spaces. A simple example of this notion is given in Figure 2.2.

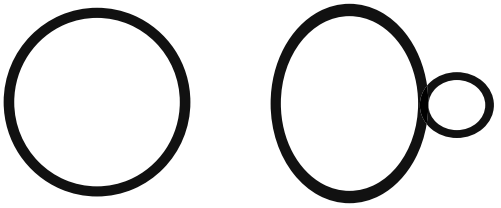


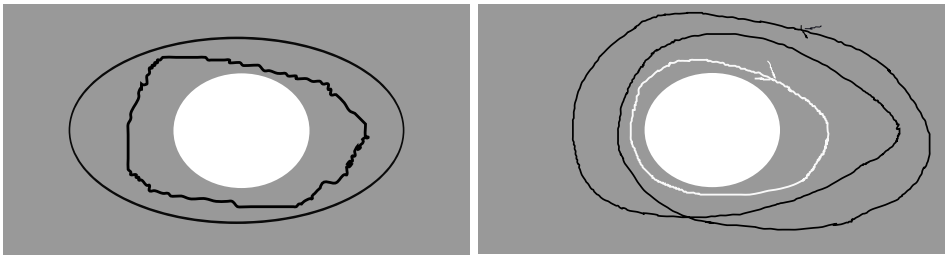
Figure 2.2. A space with a single loop, and a space with two distinct loops.

In examining these two spaces, we see that the left-hand space contains a single loop, while the second one contains two distinct loops. Thus, a count of loops is an interesting quantity to consider, from an intuitive point of view. However, it appears difficult to make precise mathematics out of this intuition. It is reasonably easy to make sense of what one means by a loop in a space X, that is, a continuous map $f : S^1 \to X$. So in this case, the pattern associated to a loop is the circle itself, and an occurrence of the pattern is a continuous map from the circle S^1 to X. However, there are almost always infinitely many loops in a space. For example, any loop can always be reparametrized by precomposing with any self-homeomorphism of the circle. Another difficulty, however, is the situation illustrated by Figure 2.3(a). The interesting feature is the hole in the centre, and both the loops (as well as an infinity of others) capture that feature, in the sense that they 'go around' the hole. This makes for an even larger set of loops, and the idea here is to create a kind of count which captures the feature using the presence of loops around it, rather than producing an infinity of loops. The key insight to be had here is that the idea of counting occurrences of patterns directly is unworkable, but that counting *equivalence classes* of occurrences of patterns under an equivalence relation is workable.

Definition 2.6. Given two maps $f, g : X \to Y$ of topological spaces, we say that f and g are *homotopic*, and write $f \simeq g$ if there is a continuous map $H : X \times [0,1] \to Y$ so that $H(x,0) = f(x)$ and $H(x,1) = g(x)$ for all $x \in X$. The relationship of being homotopic is an *equivalence relation*. When there are fixed basepoints $x_0 \in X$ and $y_0 \in Y$, we speak of *based maps* as maps $f : X \to Y$ for which $f(x_0) = y_0$, and of based homotopies as homotopies $H(x_0, t) = y_0$ for all t. Based homotopy is also an equivalence relation on the set of based maps from X to Y.

Remark 2.7. The fact that one must choose equivalence classes of occurrences of a pattern in order to obtain a workable theory is the fundamental observation in the subject. It is responsible for the power of the method, and on the other hand for the technical complexity of the subject.

The set of homotopy classes of continuous maps from a space X to a space Y is a more discrete invariant, which gives a high-level description of the set



(a) (b)

Figure 2.3. (a) Two equivalent loops. (b) Distinct equivalence classes of loops.

of maps from X to Y. When X is the n-sphere S^n, one can in a natural way impose the structure of a group on the set of equivalence classes of based maps from S^n to Y. The resulting group is denoted by $\pi_n(Y, y_0)$. Applied to the example above, with a single obstacle in the plane, this group π_1 is a single copy of the integers with addition as the operation. The integer assigned to a given loop is the so-called *winding number* of the loop, which counts how many times the loop wraps around the obstacle, with orientation taken into account as a sign.

In Figure 2.3(b) the black loop has winding number $+2$, and the white loop has winding number -1. If there were two obstacles, π_1 would be a free non-commutative group on two generators, and similarly n generators with n obstacles. If we had a three-dimensional region, with a single ball-shaped obstacle, then π_1 would be trivial, but π_2 would be identified with a copy of the integers, the invariant integer given by a higher-dimensional version of the winding number. The groups $\pi_n(Y, y_0)$ are referred to as the *homotopy groups* of a space Y. They serve as a form of pattern recognition for the space, in that they detect occurrences of the pattern corresponding to the n-spheres. The homotopy groups allow us to distinguish between spaces, as follows.

Definition 2.8. Two topological spaces X and Y are said to be *homotopy equivalent* if there are continuous maps $f : X \to Y$ and $g : Y \to X$ such that gf and fg are homotopic to the identity maps on X and Y, respectively. There is a corresponding notion for based maps and based homotopies.

Remark 2.9. We note that this is simply a 'softened' version of the usual notion of isomorphism, where the composites are required to be equal to the corresponding identity maps. Of course, spaces which are actually homeomorphic are always homotopy equivalent.

It is now possible to prove the following.

Proposition 2.10. Suppose that two spaces X and Y are based homotopy equivalent, with base points x_0 and y_0 as base points. Then all their homotopy groups $\pi_n(X, x_0)$ and $\pi_n(Y, y_0)$ are isomorphic.

This result often allows us to conclude that two spaces are not homotopy equivalent, and *a fortiori* not homeomorphic. For example, $\pi_2(S^2, 0)$ is isomorphic to the group of integers, while $\pi_2(\mathbb{R}^2, 0)$ is isomorphic to the trivial group, and we may conclude that they are not homotopy equivalent.

Although they are easy to define and conceptually very attractive, it turns out that homotopy groups of spaces are very difficult to compute. There is another kind of invariant, called *homology*, which, instead of being easy to define and difficult to compute, is difficult to define and easy to compute.

2.4. Homology

Homology was initially defined not for topological spaces directly, but rather for spaces described in a very particular way, namely as *simplicial complexes*. This description is very combinatorial, and it turns out that (a) not every space can be described as a simplicial complex, and (b) spaces can be described as simplicial complexes in many different ways. In the early development of the subject, the apparent dependence of the homology calculation on the simplicial complex structure was a serious problem, and it was the subject of a great deal of research. These problems were eventually resolved by Eilenberg, who showed that there is a way to extend the definition of homology groups to all spaces, and in such a way that the result depends only on the space itself and not on any particular structures as a simplicial complex. Eilenberg's solution was, however, extremely infinite in nature, and is not amenable to direct computation. Calculations of homology for simplicial complexes remain the best method for explicit calculation. Because most spaces of interest are either explicitly simplicial complexes or homotopy equivalent to such complexes, it turns out that simplicial calculation is sufficient for most situations.

Let $S = \{x_0, x_1, \ldots, x_n\}$ denote a subset of a Euclidean space $\mathbb{R}^k$. We say that S is in general position if it is not contained in any affine hyperplane of $\mathbb{R}^k$ of dimension less than n. When S is in general position, we define the *simplex spanned by* S to be the convex hull $\sigma = \sigma(S)$ of S in $\mathbb{R}^k$. The points x_i are called *vertices*, and the simplices $\sigma(T)$ spanned by non-empty subsets of $T \subseteq S$ are called *faces* of σ. By a (finite) *simplicial complex*, we will mean a finite collection $\mathcal{X}$ of simplices in a Euclidean space so that the following conditions hold.

(1) For any simplex σ of $\mathcal{X}$, all faces of σ are also contained in $\mathcal{X}$.

(2) For any two simplices σ and τ of $\mathcal{X}$, the intersection $\sigma \cap \tau$ is a simplex, which is a face of both σ and τ.

We note that any simplicial complex determines a combinatorial object consisting of subsets of the full vertex set of the complex, motivating the following definition.

Definition 2.11. By an abstract simplicial complex X, we will mean a pair $X = (V(X), \Sigma(X))$, where $V(X)$ is a finite set called the *vertices* of X, and where $\Sigma(X)$ is a subset (called the simplices) of the collection of all non-empty subsets of $V(X)$, satisfying the conditions that if $\sigma \in \Sigma(X)$, and $\emptyset \neq \tau \subseteq \sigma$, then $\tau \in \Sigma(X)$. Simplices consisting of exactly two vertices are called *edges*.

We note that a simplicial complex $\mathcal{X}$ determines an abstract simplicial complex whose vertex set $V(\mathcal{X})$ is the set of all vertices of all simplices of $\mathcal{X}$, and where a subset of $V(\mathcal{X})$ is in the collection of simplices $\Sigma(\mathcal{X})$ if and only if the set is the set of vertices of some simplex of $\mathcal{X}$. What is true but less obvious is that the abstract simplicial complex determines the underlying space of the simplicial complex up to homeomorphism. Indeed, given any abstract simplicial complex X, we may associate to it a space $|X|$, the *geometric realization* of X, and every simplicial complex is homeomorphic to the geometric realization of its associated abstract simplicial complex. Further, given two abstract simplicial complexes X and Y, a *map of abstract simplicial complexes* f from X to Y is a map of sets $f_V : V(X) \to V(Y)$ such that, for any simplex $\sigma \in \Sigma(X)$, the subset $f_V(\sigma) \in \Sigma_Y$. The geometric realization construction is *functorial*, in the sense that any map $f : X \to Y$ of abstract simplicial complexes induces a continuous map $|f| : |X| \to |Y|$, so that $|f \circ g| = |f| \circ |g|$ and $|\mathrm{id}_X| = \mathrm{id}_{|X|}$. By a *triangulation* of a space Z, we will mean a homeomorphism from the realization of an abstract simplicial complex with Z. A space can in general be triangulated in many different ways.

It turns out that it is very easy to describe the set of connected components of a simplicial complex in terms of its associated abstract simplicial complex. Let $\mathcal{X}$ be a simplicial complex, and X its associated abstract simplicial complex.

Proposition 2.12. Let R be the equivalence relation on $V(X)$ generated by the binary relation R' on $V(X)$ given by

$$R' = \{(v, v') | \{v, v'\} \text{ is a simplex of } X\}.$$

The connected components of $\mathcal{X}$ are in bijective correspondence with the quotient $V(X)/R$.

We now consider the following very simple example of an abstract simplicial complex, denoted by W.

The list of simplices in the corresponding simplicial complex is now

$$\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B\}, \{A, C\}, \{B, C\}, \{D, E\}\}.$$

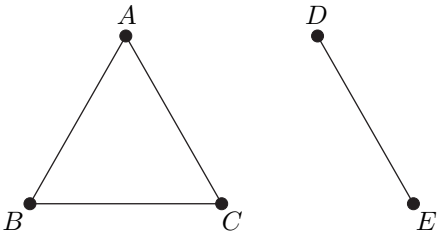


Figure 2.4. The simplicial complex W.

The geometric realization of this complex has two connected components, namely the triangle ABC and the interval DE. Our first goal is to describe the computation of the number of connected components using linear algebraic methods.

Definition 2.13. Let k be a field, and let S be a finite set. Then, by the *free k-vector space on the set S*, we will mean the vector space $V_k(S)$ of k-valued functions on S, with the vector space operations given by pointwise sum and scalar multiplication. $V_k(S)$ has a basis $B_k(S)$ identified with S, consisting of the characteristic functions φ_s defined by $\varphi_s(s') = 1$ if $s' = s$ and $\varphi_s(s') = 0$ if $s \neq s'$. In particular, the dimension of $V_k(S)$ is $\#(S)$. If $f : S \to T$ is a map of sets, then there is an associated linear transformation $V_k(f) : V_k(S) \to V_k(T)$, defined on an element $\varphi : S \to k$ of $V_k(S)$ by the formula

$$V_k(f)(\varphi)(t) = \sum_{\{s|f(s)=t\}} \varphi(s).$$

Note that a basis element φ_s is carried under $V_k(f)$ to a basis element $\varphi_{f(s)}$.

We next suppose that we are given a finite set X, with a binary relation $R \subseteq X \times X$. We define a subspace $V_k(R) \subseteq V_k(X)$ to be the subspace spanned by the set

$$\{\varphi_x - \varphi_{x'} | (x, x') \in R\}.$$

The following is now easy to show.

Proposition 2.14. There is an isomorphism of k-vector spaces

$$V_k(X)/V_k(R) \cong V_k(X/R).$$

Moreover, the composite $V_k(X) \to V_k(X)/V_k(R) \to V_k(X/R)$ is the linear transformation induced by the projection $X \to X/R$.

For simplicity, we will now assume that the field k is the field with two elements $\{0, 1\}$, with $1 + 1 = 0$, so $1 = -1$. We associate to the simplicial

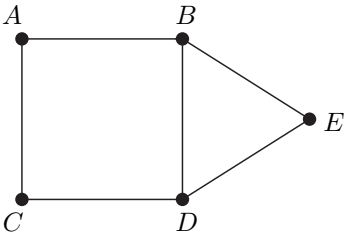


Figure 2.5. The complex X.

complex W shown in Figure 2.4 a matrix denoted by ∂_1, given by

$$\partial_1 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} AB & AC & BC & DE \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The columns are in correspondence with the edges of the complex, and the rows are in correspondence with the vertices. The entries are determined by incidence of a vertex with an edge: it is 1 if the vertex is contained in the edge and 0 if not. Examining the columns, and noting that $v - v' = v + v'$ because of our choice of field, we see that they are exactly the elements

$$\{v - v' | v \text{ and } v' \text{ span an edge in } W\}.$$

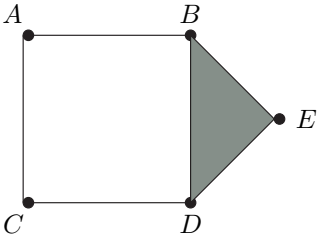
We now have the following consequence of Proposition 2.12.

Proposition 2.15. Let $\pi_0(|W|)$ denote the set of connected components of $|W|$, and let $v_0(W)$ denote the set of vertices of W. Then the vector space $V_k(\pi_0(|W|))$ is isomorphic to the quotient space of $V_k(v_0(W))$ by the column space $\mathrm{col}(\partial_1)$ of ∂_1. Moreover, there is a natural choice of basis for $V_k(\pi_0(W))$ consisting of equivalence classes of vertices.

In particular, the number of connected components is equal to the dimension of $V_k(v_0(W))/\mathrm{col}(\partial_1)$, which by Proposition 2.4 is equal to $\#(v_0(W)) - \mathrm{rank}(\partial_i)$. This linear algebraic interpretation of the number of connected components suggests that we try to interpret the nullity of ∂_1 as well as the rank.

We see easily that the rank of ∂_1 is 3, and consequently (by the rank nullity theorem) that the nullity of ∂_1 is $4 - \mathrm{rank}(\partial_1) = 1$. Inspection shows that a basis for the null space consists of the element

$$\varphi_{AB} + \varphi_{AC} + \varphi_{BC}.$$



$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} AB & AC & BD & BE & CD & DE \end{array} \\ \left[\begin{array}{cccccc} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right] \end{array}$$

Figure 2.6. A complex with a two-simplex added.

If we permit ourselves to think of the sums as unions, this linear combination corresponds to the union of the edges AB, AC, and BC. This union is a cycle of length three in the complex W shown in Figure 2.4. It is a useful exercise to make similar computations for other graphs, in particular n-cycles for numbers > 3, to convince ourselves that the nullity is in each case the number of cycles in the graph, suitably interpreted. To understand the interpretation, consider the following complex.

We can see that there are two obvious cycles, $AB + BD + DC + AC$ and $BE + ED + DB$, represented by the elements of the null space $\zeta_1 = \varphi_{AB} + \varphi_{BD} + \varphi_{CD} + \varphi_{AC}$ and $\zeta_2 = \varphi_{BE} + \varphi_{ED} + \varphi_{DB}$. However, there is another cycle given by

$$AB + BE + ED + DC + CA. \tag{2.1}$$

Note, however, that the sum $\zeta_1 + \zeta_2$ is equal to the element

$$\varphi_{AB} + \varphi_{BE} + \varphi_{ED} + \varphi_{DC} + \varphi_{CA},$$

which is the element of the null space of ∂_1 corresponding to the cycle in (2.1). The cycles actually correspond to the elements of vector space, and can therefore be added and multiplied by scalars. This give an extremely useful way of organizing the cycles. In particular, one can construct a *basis* of the cycles, instead of counting them all individually.

Next, consider the complex in Figure 2.6, with its corresponding ∂_1 matrix. The shading reflects the fact that there is now a two-simplex, namely $\{B, E, D\}$. As in the complex X in Figure 2.5, we have the two loops $ABDC$ and BED. In this case, though, the loop BDE is filled in by a simplex, and the loop $ABEDC$ can be deformed in the space to the loop $ABDC$ by traversing the two-simplex. So, by analogy with the discussion of homotopy in Section 2.3, we should construct our formalism in such a way that the two distinct cycles become equal. More importantly, though, we are interested in constructing vector spaces which in the end should depend only on the underlying space, not on the particular triangulation, that is, on the particular way in which it is described as a simplicial complex.

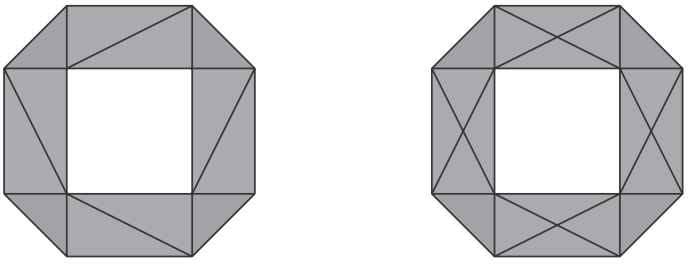


Figure 2.7. Two triangulations of the same space.

In Figure 2.7, we see the same space expressed as a simplicial complex in two different ways. If we construct ∂_1 as above, we find that the null space of ∂_1 has dimension 12 in the case of the left-hand complex, and dimension 20 in the case of the right-hand complex. It is therefore clear that to have a meaningful measure of the number of loops, independent of the particular simplicial complex structure, we will need to modify or augment the linear algebra we have so far introduced.

To motivate the construction, consider the complex X in Figure 2.6. In addition to the vertices and edges, we now have a two-simplex, whose boundary is the cycle *BED*. The corresponding linear combination

$$\varphi_{\{B,E\}} + \varphi_{\{E,D\}} + \varphi_{\{B,D\}}$$

in $V_k(\sigma_1(Z))$ is the difference between the linear combinations

$$\varphi_{\{A,B\}} + \varphi_{\{B,E\}} + \varphi_{\{E,D\}} + \varphi_{\{D,C\}} + \varphi_{\{C,A\}}$$

and

$$\varphi_{\{A,B\}} + \varphi_{\{B,D\}} + \varphi_{\{D,C\}} + \varphi_{\{C,A\}},$$

again recalling that differences and sums are the same thing in our field k. The key idea is that we form the quotient of the null space of ∂_1 by the element

$$\varphi_{\{B,E\}} + \varphi_{\{E,D\}} + \varphi_{\{B,D\}},$$

which will be called a *boundary* because it is derived from the boundary of the simplex $\{B, E, D\}$. This discussion now leads to the following construction attached to an abstract simplicial complex, which includes all the higher-dimensional simplices.

We let X be a simplicial complex, and as above let $\Sigma_i(X)$ denote the set of i-dimensional simplices, that is, those simplices which as subsets of the vertex set $V(X)$ have cardinality $i+1$. We define matrices ∂_i whose columns are in one-to-one correspondence with $\Sigma_i(X)$, and whose rows are in one-to-one correspondence with the $\Sigma_{i-1}(X)$, by declaring that the entry in the row corresponding to an $(i-1)$-simplex τ and in the column corresponding

to an i-simplex τ' is 1 if $\tau \subseteq \tau'$ as sets of vertices, and it is 0 otherwise. This definition is consistent with the matrices we have constructed in the special cases above. There is now a key observation relating the matrices ∂_i and ∂_{i-1}.

Proposition 2.16. The matrix product $\partial_{i-1} \cdot \partial_i$ is equal to the zero matrix.

Proof. The rows of $\partial_{i-1} \cdot \partial_i$ are in one-to-one correspondence with $\Sigma_{i-2}(X)$, and its columns are in one-to-one correspondence with $\Sigma_i(X)$. It is easy to see that the entry in the row corresponding to an $(i-2)$-simplex τ and in the column corresponding to an i-simplex τ' is equal to the number of elements $\hat{\tau}$ of $\Sigma_{i-1}(X)$ which satisfy $\tau \subseteq \hat{\tau} \subseteq \tau'$. This number is either 0, in the case $\tau \not\subset \tau'$, or 2, in the case $\tau \subset \tau'$. Both numbers are zero in the field k. □

The matrices ∂_i can be regarded as the matrices attached to linear transformations from $V_k(\Sigma_i(X))$ to $V_k(\Sigma_{i-1}(X)$, relative to the standard bases of $V_k(\Sigma_i(X))$. Abusing notation, we will denote the matrices and the transformations by ∂_i. What we now have is a diagram,

$$\cdots \xrightarrow{\partial_{i+2}} V_k(\Sigma_{i+1}(X)) \xrightarrow{\partial_{i+1}} V_k(\Sigma_i(X)) \xrightarrow{\partial_i} V_k(\Sigma_{i-1}(X)) \xrightarrow{\partial_{i-1}} \cdots$$
$$\xrightarrow{\partial_2} V_k(\Sigma_1(X)) \xrightarrow{\partial_1} V_k(\Sigma_0(X)),$$

in which each composite of two consecutive linear transformations is identically zero. This observation now suggests the following definition.

Definition 2.17. By a *chain complex* C_* over a field k, we will mean a choice of k-vector space C_i for every $i \geq 0$, together with linear transformations $\partial_i : C_i \to C_{i-1}$ for all i, so that $\partial_{i-1} \cdot \partial_i \equiv 0$ for all i.

We now extract information as follows. For every i, we define two subpsaces B_i and Z_i of C_i. Z_i is defined as the null space of ∂_i, and B_i is defined as the image of ∂_{i+1}. By Proposition 2.16, it follows that $B_i \subseteq Z_i$, and we define $H_i(C_*)$ to be the quotient space Z_i/B_i. One can check that in the case of the complex in Figure 2.6, H_1 turns out to be a one-dimensional vector space over k, with $\varphi_{AB}+\varphi_{BD}+\varphi_{DC}+\varphi_{CA}$ as an element in Z_1 whose image in Z_1/B_1 is a non-trivial element, therefore a basis for H_1. These vector spaces, applied to the chain complex associated with a simplicial complex, will be called the *homology groups* of the complex.

We now describe the linear algebra which is carried out to compute the homology groups, and in particular their dimension. That is, we want to interpret the computation of the homology groups of a complex in terms of row and column operations. The row and column operations will be multiplication of a single row (column) by a non-zero element of k, adding a multiple of one row (column) to another, and transposing a pair of rows

(columns). We recall Remark 2.5, which asserts that given a matrix A over a field k, one may perform both row and column operations of the type described above to obtain a matrix $\overline{A}$ having the normal form

$$\begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix},$$

where n is the rank of A. $\overline{A}$ is uniquely determined by A.

In order to study homology, we will instead need to study the normal forms of pairs of matrices (A, B) with $A \cdot B = 0$.

Proposition 2.18. Let

$$U \xrightarrow{B} V \xrightarrow{A} W$$

be linear transformations, such that $A \cdot B = 0$. Let $F : U \to U$, $G : V \to V$, and $H : W \to W$ be invertible linear transformations. Then we have $HAG^{-1} \cdot GBF = 0$, and there is an isomorphism of vector spaces

$$N(A)/\mathrm{im}(B) \cong N(HAG^{-1})/\mathrm{im}(GBF).$$

Proof. Entirely analogous to Proposition 2.3. □

The matrix version now uses the following set of admissible operations on such a pair. They will be the following.

(1) An arbitrary row operation on A.

(2) An arbitrary column operation on B.

(3) Perform a column operation on A and a row operation on B simultaneously, with the operations related as follows. If the column operation on A is multiplication of the ith column by a non-zero constant x, then the row operation on B is multiplication of the ith row by x^{-1}. If the column operation on A is the transposition of two columns, then the row operation on B is the transposition of the corresponding rows of B. Finally, if the column operation on A is the addition of x times the ith column to the jth column, then the row operation on B is the subtraction of x times the jth row from the ith row.

Note that if we apply any of these operations to a pair (A, B) to obtain a pair (A', B'), then (A', B') also satisfies $A' \cdot B' = 0$. We now have the following counterpart of Proposition 2.4 above.

Proposition 2.19. Given a pair (A, B), with $A \cdot B = 0$, we can perform operations of the type described above to obtain a pair (A', B'), with

$$\left(\begin{bmatrix} I_n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_m \end{bmatrix} \right).$$

Here, if there are k, l and m columns in the leftmost, middle, and rightmost blocks of columns of A respectively (and consequently k, l and m rows in the top, middle, and bottom blocks of rows of B, respectively), then the dimension of the homology is l. The pair (A', B') is uniquely determined by the pair (A, B).

Proof. We first perform arbitrary row and column operations to A (making sure to apply the corresponding row operations to B whenever a column operation is applied to A), to obtain a pair (A', B') of the form

$$\left(\begin{bmatrix} I_k & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} B'_{11} & B'_{12} & B'_{13} \\ B'_{21} & B'_{22} & B'_{23} \\ B'_{31} & B'_{32} & B'_{33} \end{bmatrix} \right)$$

for some k. Because of the condition $A' \cdot B' = 0$, it is clear that $B'_{11} = B'_{12} = B'_{13} = 0$. We can now perform row and column operations on the matrix B', which is of the form

$$\begin{bmatrix} 0 & 0 & 0 \\ B'_{21} & B'_{22} & B'_{23} \\ B'_{31} & B'_{32} & B'_{33} \end{bmatrix}.$$

We perform only row operations involving the last $l+m$ rows, since the upper k rows are identically zero. Each such row operation has a corresponding column operation on the matrix A' which affects only the rightmost $l + m$ columns, and therefore has no effect. Performing these operations is equivalent to performing arbitrary row and column operations to the matrix

$$\begin{bmatrix} B'_{21} & B'_{22} & B'_{23} \\ B'_{31} & B'_{32} & B'_{33} \end{bmatrix},$$

which we denote by $\hat{B}'$. We can therefore perform operations (which have no effect on A') so as to put $\hat{B}'$ into the form

$$\begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix},$$

and then transpositions of rows and columns so as to put it into the form

$$\begin{bmatrix} 0 & 0 \\ 0 & I_m \end{bmatrix}.$$

That this representation is unique is clear from the fact that k and m are the ranks of the matrices A and B, respectively, and this completes the proof. □

2.5. Functoriality

We have defined vector spaces associated to simplicial complexes, and in Section 2.3, groups $\pi_n(X, x_0)$ associated to a space X with base point x_0,

and sets $\pi_0(X)$. The most important property of these constructions is that they are able to reflect the behaviour of maps between simplicial complexes and continuous maps between topological spaces. This property is referred to as *functoriality*. In the case of $\pi_0(X)$, this is simply the statement that given a continuous map $f : X \to Y$, there is an associated map of sets $\pi_0(f) : \pi_0(X) \to \pi_0(Y)$, since any two points which are connected by a path in X (and are therefore in the same path component) map under f to points in Y which are connected by a path. The path in Y is simply the image of the path in X. Informally, this is just the statement that 'components map to components'. More generally, given a continuous map $f : X \to Y$, with $f(x_0) = y_0$, there is an induced homomorphism of groups $\pi_n(f) : \pi_n(X, x_0) \to \pi_n(Y, y_0)$, which is defined by carrying the equivalence class of $\varphi : S^n \to Y$ to the equivalence class of $f \cdot \varphi : S^n \to Y$. It is easily shown that this is well defined. Similarly, for any map of abstract simplicial complexes $f : X \to Y$, there is an induced linear transformation $H_n(f) : H_n(X) \to H_n(Y)$. It is obtained by showing the following.

(1) There are linear transformations $V_k(\Sigma_n(f)) : V_k(\Sigma_n(X)) \to V_k(\Sigma_n(Y))$ which carry the basis elements φ_τ to the basis elements $\varphi_{f(\tau)}$.

(2) The homomorphisms ∂_i respect the maps $V_k(\Sigma_n(f))$ in the sense that the diagrams

$$\begin{array}{ccc} V_k(\Sigma_n(X)) & \xrightarrow{V_k(\Sigma_n(f))} & V_k(\Sigma_n(Y)) \\ \downarrow{\scriptstyle \partial_n} & & \downarrow{\scriptstyle \partial_n} \\ V_k(\Sigma_{n-1}(X)) & \xrightarrow{V_k(\Sigma_n(f))} & V_k(\Sigma_{n-1}(Y)) \end{array}$$

commute.

(3) It follows from (2) above that $V_n(\Sigma_n(f))$ carries $Z_n(X)$ into $Z_n(Y)$ and $B_n(X)$ into $B_n(Y)$.

(4) It follows from (3) above that there is an induced homomorphism

$$H_n(X) = Z_n(X)/B_n(X) \xrightarrow{H_n(f)} Z_n(Y)/B_n(Y) = H_n(Y).$$

Remark 2.20. Functoriality of algebraic invariants in topology is one of its fundamental tools. We will see below that it is what permits us to define sensible homological invariants of finite metric spaces or point clouds.

2.6. Extending from complexes to spaces

As we have defined it, homology applies to simplicial complexes. This means that when one is given a topological space X, say defined as the set of solutions of a set of equations, there is not yet a natural definition of the homology groups of X, without constructing a triangulation of X. Even if

one can construct a triangulation, it is not clear why a different triangulation would not give a different answer. The problem of extending homology spaces without a given triangulation was studied intensively in the early 1900s, and was resolved by Eilenberg (1944). He defined homology groups $H_n(X)$ (called the *singular homology groups*) for any space X by constructing a chain complex of infinite-dimensional vector spaces (a basis of $C_n(X)$ is given by the set of all continuous maps $\Delta^n \to X$) whose homology agrees with the simplicially constructed homology for any triangulation of X. The homology groups constructed with these complexes are functorial for any continuous map $f : X \to Y$, by which we mean that there is a linear transformation $H_n(f) : H_n(X) \to H_n(Y)$ associated to f. An additional important property is *homotopy invariance*, which asserts that if two maps $f, g : X \to Y$ are homotopic (see Definition 2.6), then the linear transformations $H_n(f)$ and $H_n(g)$ are equal. This is a powerful property, and it allows direct calculations of homology in some cases. We say that a topological space X is *contractible* if id_X is homotopic to a constant map.

Example 2.21. $\mathbb{R}^n$ is contractible, because we have the explicit homotopy $H(v, t) = (1 - t)v$ from $\mathrm{id}_{\mathbb{R}^n}$ to the constant map with value 0.

Proposition 2.22. If X is a contractible space, then $H_n(X) = \{0\}$ for all $n > 0$.

Proof. By the homotopy property and functoriality, the identity transformation on $H_n(X)$ is equal to the composite $H_n(X) \to H_n(x_0) \to H_n(X)$, where $x_0 \in X$ is a point such that f is homotopic to the constant map with value x_0. But it is easy to check directly that $H_n(x_0) = \{0\}$, and it now follows that every element in $H_n(X)$ is 0. □

Remark 2.23. The singular homology groups are of course impossible to compute directly, since in particular they involve linear algebra on vector spaces of uncountable dimension. This means that when we wish to compute homology, we must either construct a triangulation or use other computational techniques, called *excision* and *long exact sequences*, which have been developed for this purpose. The important point in the definition is that it gives well-defined groups for any space, which behave functorially for continuous maps.

2.7. Making homology more sensitive

It is useful to ask how sensitive a measure homology is of the shape of a simplicial complex, but considering a simple shape recognition task, namely the recognition of printed letters. We begin with the first three letters of the alphabet, and find that H_1 succeeds in distinguishing between them. However, after this initial success, we see that every other letter has the same first Betti number as one of these three: see Figure 2.8.

(a) $\beta_1 = 1$ (b) $\beta_1 = 2$ (c) $\beta_1 = 0$

$\beta_1 = 0$	{C, E, F, G, H, I, J, K, L, M, N, S, T, U, V, W, X, Y, Z}
$\beta_1 = 1$	{A, D, O, P, Q, R}
$\beta_1 = 2$	{B}

Figure 2.8. Discrimination of letters by first Betti number.

Homology can be refined to discriminate more finely between the letters. To understand how this works, we digress a bit to discuss how an analogous problem in manifold topology is approached. In Section 2.6, we saw that the homology groups $H_i(\mathbb{R}^n)$ vanish for all $i > 0$. What this means is that homology is unable to distinguish between $\mathbb{R}^m$ and $\mathbb{R}^n$ when $m \neq n$. From the point of view of a topologist who is interested in distinguishing different manifolds from each other, this means that homology is in some ways a relatively weak invariant. This failure can be addressed by computing homology on 'auxiliary' or 'derived' spaces, constructed using various geometric constructions.

(1) *Removing a point.* While the homology groups of $\mathbb{R}^n$ vanish, the homology groups of $\mathbb{R}^n - \{0\}$ do not. To see this, we observe that we have the inclusion $i : S^{n-1} \hookrightarrow \mathbb{R}^n - \{0\}$ as well as the map $r : \mathbb{R}^n - \{0\} \to S^{n-1}$ defined by $r(v) = f/\|v\|$; $r \cdot i$ is equal to the identity map for S^{n-1}, and the other composite $i \cdot r$ is homotopic to the identity map of $\mathbb{R}^n - \{0\}$ via the straight-line homotopy $H(v, t) = (1 - t)i \cdot r + t\,\mathrm{id}_{\mathbb{R}^n - \{0\}}$. The existence of this homotopy shows that the map $H_i(r)$ is an isomorphism. It is injective because, given any $0 \neq x \in H_i(\mathbb{R}^n - \{0\})$, we have $x = H_i(i \cdot r)(x) = H_i(i) \cdot H_i(r)(x)$, so $H_i(r)(x) \neq 0$. It is surjective because for any $y \in H_i(S^{n-1})$, $y = H_i(r \cdot i)(y) = H_i(r)H_i(i)(y)$, which exhibits y as the image of $H_i(i)(y)$. Because it is easy to show that $\mathbb{R}^n - \{v\}$ is homeomorphic to $\mathbb{R}^n - \{0\}$, one can detect the difference between $\mathbb{R}^n$ and $\mathbb{R}^m$, with $m \neq n$, by recognizing that the homology of the results of removing a single point from the two spaces are different, since homology detects the difference between spheres of different dimensions.

(2) *One-point compactification.* For any space X, one may construct its *one-point compactification* by adjoining a single point ∞ to X, and declaring that neighbourhoods of ∞ are exactly the complements of

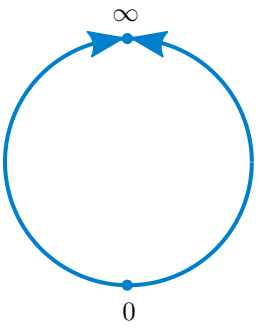


Figure 2.9. One-point compactification.

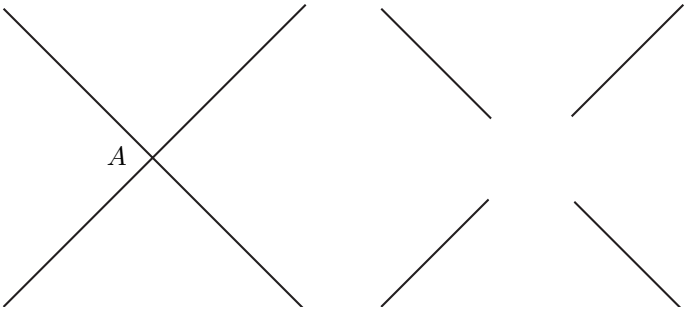


Figure 2.10. Removing singular points.

compact subsets of X together with ∞. Figure 2.9 shows the one-point compactification of the real line. Note that although the real line is contractible and therefore has vanishing homology, its one-point compactification is homeomorphic to the circle, and has non-vanishing H_1.

(3) *Removing singular points.* Consider the space given by the crossing of two lines. This space is also contractible, as one can easily see by retracting each line segment onto the crossing point A. On the other hand, we recognize that its shape has features which distinguish it from an interval or a circle, and might want to detect that homologically. If we remove the singular point A, we will find that the space remaining breaks up into four distinct components, which can be detected by H_0.

To think through how we might apply these ideas to the problem of distinguishing between letters, let us define an *end* of a space X to be a point $x \in X$, so that there is a neighbourhood N of x which is homeomorphic to $[0, 1)$, and so that the homeomorphism carries x to 0.

In this case, the auxiliary or derived space is the set of ends of the space, $\mathfrak{e}(X)$, and we can compute its zero-dimensional homology H_0, to get the Betti number β_0. We now obtain the partition of the set of letters shown in Figure 2.11.

(a) two ends (b) no ends (c) four ends

$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 0$	none
$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 1$	none
$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 2$	{C, G, I, J, L, M, N, S, U, V, W, Z}
$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 3$	{E, F, T, Y}
$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 4$	{H, K, X}
$\beta_1 = 1,\ \beta_0(\mathfrak{e}(X)) = 0$	{D, O}
$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 1$	{P, Q}
$\beta_1 = 0,\ \beta_0(\mathfrak{e}(X)) = 2$	{A, R}
$\beta_1 = 2,\ \beta_0(\mathfrak{e}(X) = 0$	{B}

Figure 2.11. Improved discrimination of letters by using ends as well as first Betti number.

We have obtained improved discrimination in this way. Removing singular points and computing β_0 produces further resolution on the letters. For further discussion of this kind of analysis, see Carlsson, Zomorodian, Collins and Guibas (2005). If one uses homology on suitably constructed auxiliary spaces, one can obtain classification criteria for many interesting problems in shape discrimination.

3. Shape of data

3.1. Motivation

The main goal of this paper is to extend the homological methods described in Section 2.4 to obtain similar methods applicable to finite metric spaces. The methods that we describe have been developed in Robins (1999), Edelsbrunner, Letscher and Zomorodian (2002), and Zomorodian and Carlsson (2005). There is of course a naive extension which regards a finite metric space as a topological space in the usual way, but this method endows the points with the discrete topology. In other words, it contains no information about shape. What we would like to do is to produce a tool which is capable, for example, of recognizing that the metric space represented by Figure 3.1 contains the pattern of a loop. The set of points, as a topological space, is discrete, but we are able visually to infer that it appears to be sampled from some kind of circular geometry. What will be shown in this

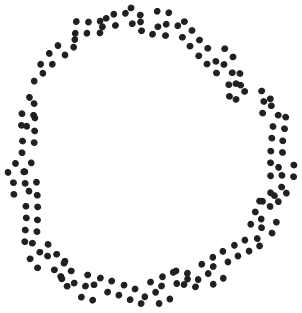


Figure 3.1. A 'statistical circle'.

section is that it is possible to develop just such a tool. To motivate this construction, we will look at a commonly used statistical methodology.

3.2. Single linkage clustering

The very simplest aspect of the shape of a geometric object is its number of connected components. Statisticians have thought a great deal about what the counterpart to connected components should be for point cloud data, under the heading of *clustering*: see Hartigan (1975) and Kogan (2007). One scheme for clustering proceeds as follows. We suppose we are given a finite metric space with points $X = \{x_1, \ldots, x_n\}$, and given pairwise distances. For every non-negative threshold R, we may form the relation $\sim_R$ on the set X by the criterion

$$x \sim_R x' \text{ if and only if } d(x, x') \leq R.$$

We let $\simeq_R$ denote the equivalence relation generated by $\sim_R$. The set of equivalence classes under $\simeq_R$ now gives a partition of X, which can be thought of as a candidate for the connected components in X. So for each threshold R, we obtain a partition of X. One can now ask which choice of R is the 'right' one. This is an ill-defined question, although there are interesting heuristics. Another approach is to observe that there is compatibility across changes in R, in that if $R \leq R'$, then the partition associated to R' is coarser than the partition associated to R, as is indicated in Figure 3.2. The diagram indicates the change in clustering as the threshold is altered, and shows the increasing coarseness as R increases. What was recognized by statisticians is that there is a single profile, called a *dendrogram*, which encodes the clusterings at all the thresholds simultaneously. Figure 3.3 shows a dendrogram which is associated to the situation given above. The result is a tree (a simplicial complex with no loops) T together with a reference map from T to the non-negative real line. This tree can be viewed all at once. The clustering at a given threshold R is given by drawing a horizontal

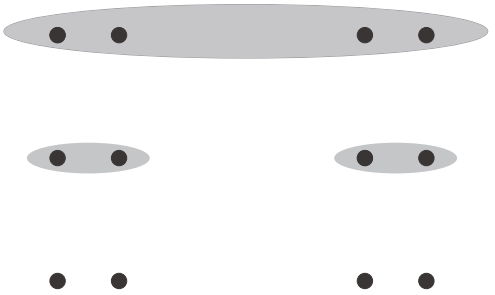


Figure 3.2. Single linkage hierarchical clustering.

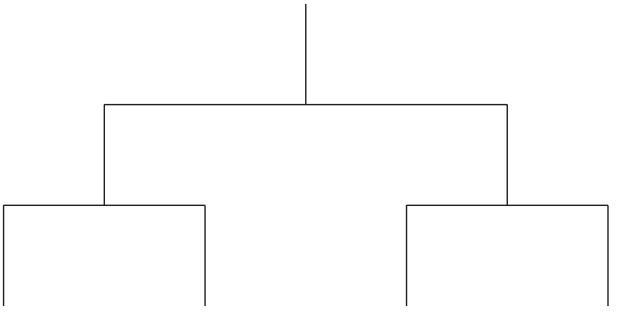


Figure 3.3. Dendrogram.

line at level R across the tree, and the clusters correspond to the points of intersection. In Figure 3.3, the reference map is the height function above the x-axis.

There is another way to interpret the dendrogram, less visually but formally identical. For each threshold R, we will let X_R denote the set of equivalence classes for the equivalence relation $\simeq_R$. Because the partition only coarsens as R increases, there is a map of sets $X_R \to X_{R'}$ whenever $R \leq R'$, which assigns to each cluster at level R the (unique) cluster at level R' in which it is included. This construction is sufficiently useful that we will give it a name.

Definition 3.1. By a *persistent set*, we will mean a family of sets $\{X_R\}_{R\in\mathbb{R}}$ together with set maps

$$\varphi_R^{R'} : X_R \to X_{R'} \quad \text{for all } R \leq R',$$

so that

$$\varphi_{R'}^{R''}\varphi_R^{R'} = \varphi_R^{R''} \quad \text{for all } R \leq R' \leq R''.$$

More generally, for any kind of objects, such as simplicial complexes, vector spaces or topological spaces, we may speak of a persistent object as a family of such objects parametrized by $\mathbb{R}$, together with maps (maps of simplicial

complexes, linear transformations, continuous maps, *etc.*) from the object parametrized by r to the one parametrized by r' whenever $r \leq r'$, with the same compatibilities mentioned above.

For each persistent set there is an associated dendrogram and *vice versa*. There is a reformulation of the dendrogram (and hence a persistent set) associated to a finite metric space using topological notions which will greatly clarify the development of methods of defining higher-dimensional homology for finite metric spaces.

Definition 3.2. Given any finite metric space X and non-negative real number R, we construct an abstract simplicial complex $\mathrm{VR}(X, R)$, called the *Vietoris–Rips complex of* X, by letting its vertex set be the underlying set of X, and declaring that any subset $\{x_0, \ldots, x_n\}$ of X is a simplex of $\mathrm{VR}(X, R)$ if and only if

$$d(x_i, x_j) \leq R \text{ for all } i, j \in \{0, \ldots, n\}.$$

We note that whenever $R \leq R'$ there is an inclusion $\mathrm{VR}(X, R) \hookrightarrow \mathrm{VR}(X, R')$, because the vertex sets of the two abstract complexes are the same, and that any simplex of $\mathrm{VR}(X, R)$ is also a simplex of $V(X, R')$. It follows that we obtain maps

$$|\mathrm{VR}(X, R)| \to |\mathrm{VR}(X, R')|$$

as well. In short, the family of Vietoris–Rips complexes $\{\mathrm{VR}(X, R)\}_{R \in \mathbb{R}}$ forms a persistent simplicial complex.

The point of this definition is that given any finite metric space X, we may define the set $\pi_0(|\mathrm{VR}(X, R)|)$ of connected components of $|\mathrm{VR}(X, R)|$, and due to the functoriality of the construction π_0, we obtain a persistent set $\{\pi_0(|\mathrm{VR}(X, R)|)\}_R$. This persistent set can easily be seen to be identical to the persistent set obtained above. The point is that it is induced by a map of topological spaces, which will point the way to defining homological shape invariants of finite metric spaces.

3.3. Persistence

The value of the Vietoris–Rips construction is that for each threshold we are able to construct a simplicial complex and therefore a topological space, rather than just a partitioning or clustering of the finite metric space. In the example shown in Figure 3.4 the underlying metric space consists of six points, and is pictured on the left. The lower bar is the threshold parameter R, and on the right we have a picture of the Vietoris–Rips complex associated to this metric space and this threshold. The metric space looks as if it might be sampled from a loop, and we note that the Vietoris–Rips complex contains a loop. This suggests that we should use Vietoris–Rips

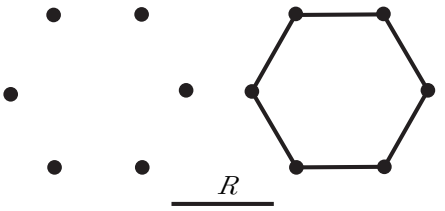


Figure 3.4. Metric space and associated Vietoris–Rips complex.

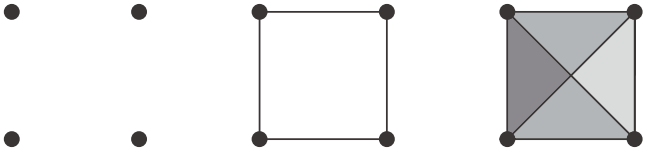


Figure 3.5. Increasing family of Vietoris–Rips complexes.

complexes as representations of the shape of finite metric spaces. Of course, the Vietoris–Rips complex is dependent on the choice of threshold, with the Vietoris–Rips complex consisting of six discrete points when R is smaller, and it becomes a full simplex (which does not admit a non-trivial loop) when R is larger than the diameter of the original metric space. This suggests that we would need to find the 'right' value of R to capture the shape. However, there is no obvious heuristic for making this choice. The situation is entirely analogous to the situation arising in single linkage clustering, described in the previous section. There, it was possible to construct a profile, the dendrogram, which provided a simple representation of the behaviour of the clustering for all values of the threshold parameter R at once. We will now show that there is a similar profile which allows us to study higher-dimensional homology of the Vietoris–Rips complexes at all threshold values at once. In the example in Figure 3.5 we show the Vietoris–Rips complexes for increasing values of R, and we observe that we have the inclusions of complexes for smaller values of R into the complexes for larger values. We now recall the functoriality of homology from Section 2.5, and fix a non-negative integer i. By applying H_i to the family $\{\mathrm{VR}(X,R)\}_{R\in\mathbb{R}}$, we obtain a family $\{H_i(\mathrm{VR}(X,R))\}_{R\in\mathbb{R}}$ of vector spaces parametrized by the real line, and the functoriality of H_i allows us to give this family the structure of a *persistent vector space*. This persistent vector space contains within it the information about the homology of all the Vietoris–Rips complexes for every threshold parameter R, as well as the behaviour of the linear transformations induced by the inclusion maps from one Vietoris–Rips complex to another. The question for us now is whether or not there is a simple visual or conceptual representation of the persistent vector space, like the

dendrogram for the case of the persistent sets in single linkage clustering. There is such a representation, and in order to present it we will need to introduce a bit of algebra.

3.4. The algebra of persistence vector spaces

We first define persistence vector spaces.

Definition 3.3. Let k be any field. Then, by a *persistence vector space* over k, we will mean a family of k-vector spaces $\{V_r\}_{r\in\mathbb{R}}$, together with linear transformations $L_V(r,r') : V_r \to V_{r'}$ whenever $r \leq r'$, so that $L_V(r',r'') \cdot L_V(r,r') = L_V(r,r'')$ for all $r \leq r' \leq r''$. A *linear transformation* f of persistence vector spaces over k from $\{V_r\}$ to $\{W_r\}$ is a family of linear transformations $f_r : V_r \to W_r$, so that for all $r \leq r'$, all the diagrams

$$\begin{array}{ccc} V_r & \xrightarrow{L_V(r,r')} & V_{r'} \\ {\scriptstyle f_r}\downarrow & & \downarrow{\scriptstyle f_{r'}} \\ W_r & \xrightarrow{L_W(r,r')} & W_{r'} \end{array}$$

commute in the sense that

$$f_{r'} \circ L_V(r,r') = L_W(r,r') \circ f_r.$$

A linear transformation is an *isomorphism* if it admits a two-sided inverse. A *sub-persistence vector space* of $\{V_r\}$ is a choice of k-subspaces $U_r \subseteq V_r$, for all $r \in [0,+\infty)$, so that $L_V(r,r')(U_r) \subseteq U_{r'}$ for all $r \leq r'$. If $f : \{V_r\} \to \{W_r\}$ is a linear transformation, then the *image* of f, denoted by $\mathrm{im}(f)$, is the sub-persistence vector space $\{\mathrm{im}(f_r)\}$.

Remark 3.4. In many constructions the variable will be restricted to $[0,+\infty)$. This will be clear from the context and should not cause confusion.

The notion of a quotient space also extends to persistence vector spaces. If $\{U_r\} \subseteq \{V_r\}$ is a sub-persistence vector space, then we can form the persistence vector space $\{V_r/U_r\}$, where $L_{V/U}(r,r')$ is the linear transformation from V_r/U_r to $V_{r'}/U_{r'}$ given by sending the equivalence class $[v]$ to the equivalence class $[L_V(r,r')(v)]$ for any $v \in V_r$.

We will also want to extend the notion of the free vector space on a set. Let X be any set, equipped with a function $\rho : X \to [0,+\infty)$. We will refer to such a pair (X,ρ) as an $\mathbb{R}_+$-filtered set. Then, by the *free persistence vector space* on the pair (X,ρ), we will mean the persistence vector space $\{W_r\}$, with $W_r \subseteq V_k(X)$ equal to the k-linear span of the set $X[r] \subseteq X$ defined by $X[r] = \{x \in X | \rho(x) \leq r\}$. Note that $X[r] \subseteq X[r']$ when $r \leq r'$, so there is an inclusion $W_r \subseteq W_{r'}$. The following is a simple observation.

Proposition 3.5. A linear combination $\sum_x a_x x \in V_k(X)$ lies in W_r if and only if $a_x = 0$ for all x with $\rho(x) > r$.

We will write $\{V_k(X,\rho)_r\}$ for this persistence vector space. We say a persistence vector space is *free* if it is isomorphic to one of the form $V_k(X,\rho)$ for some (X,ρ), and we say it is *finitely generated* if X can be taken to be finite.

Definition 3.6. A persistence vector space is *finitely presented* if it is isomorphic to a persistence vector space of the form $\{W_r\}/\operatorname{im}(f)$ for some linear transformation $f : \{V_r\} \to \{W_r\}$ between finitely generated free persistence vector spaces $\{V_r\}$ and $\{W_r\}$.

The choice of a basis for vector spaces V and W allows us to represent linear transformations from V to W by matrices. We will now show that there is a similar representation for linear transformations between free persistence vector spaces. For any pair (X,Y) of finite sets and field k, an (X,Y)-*matrix* is an array $[a_{xy}]$ of elements of a_{xy} of k. We write $r(x)$ for the row corresponding to $x \in X$, and $c(y)$ for the column corresponding to y. For any finitely generated free persistence vector space $\{V_r\} = \{V_k(X,\rho)_r\}$, we observe that $V_k(X,\rho)_r = V_k(X)$ for r sufficiently large, since X is finite. Therefore, for any linear transformation $f : \{V_k(Y,\sigma)_r\} \to \{V_k(X,\rho)_r\}$ of finitely generated free persistence vector spaces, f gives a linear transformation $f_\infty : V_k(Y) \to V_k(X)$ between finite-dimensional vector spaces over k, and using the bases $\{\varphi_x\}_{x\in X}$ of $V_k(X)$ and $\{\varphi_y\}_{y\in Y}$ of $V_k(Y)$ determines an (X,Y)-matrix $A(f) = [a_{xy}]$ with entries in k. Note that in order to obtain the usual notion of a matrix as a rectangular array, we would need to impose total orderings on X and Y, but the matrix manipulations do not require this.

Proposition 3.7. The (X,Y)-matrix $A(f)$ has the property that $a_{xy} = 0$ whenever $\rho(x) > \sigma(y)$. Any (X,Y)-matrix A satisfying these conditions uniquely determines a linear transformation of persistence vector spaces

$$f_A : \{V_k(Y,\sigma)_r\} \to \{V_k(X,\rho)_r\}$$

and the correspondences $f \to A(f)$ and $A \to f_A$ are inverses to each other.

Proof. The basis vector y lies in $V_k(Y,\sigma)_{\sigma(y)}$. On the other hand,

$$f(\varphi_y) = \sum_{x\in X} a_{xy}\varphi_x.$$

By Proposition 3.5, on the other hand, $\sum_{x\in X} a_{xy}\varphi_x$ only lies in $V_k(X,\rho)_{\sigma(y)}$ if all coefficients a_{xy}, for $\rho(x) > \sigma(y)$, are zero. □

When we are given a pair of $\mathbb{R}_+$-filtered finite sets (X, ρ) and (Y, σ), we will call an (X, Y)-matrix satisfying the conditions of Proposition 3.7 (ρ, σ)-*adapted*.

Suppose now that we are given (X, ρ) and (Y, σ), with ρ and σ both $[0, +\infty)$-valued functions on X and Y, respectively. Then any matrix $A = [a_{xy}]$ satisfying the conditions of Proposition 3.7 determines a persistence vector space via the correspondence

$$A \xrightarrow{\theta} V_k(X, \rho)/\mathrm{im}(f_A).$$

We have the following facts about this construction.

Proposition 3.8. For any A as described above, $\theta(A)$ is a finitely presented persistence vector space. Moreover, any finitely presented persistence vector space is isomorphic to one of the form $\theta(A)$ for some such matrix A.

Proof. Immediate from the correspondence between matrices and linear transformations given in Proposition 3.7. □

Proposition 3.9. Let (X, ρ) be an $\mathbb{R}_+$-filtered set. Then, under the matrix/linear transformation correspondence, the automorphisms of $V_k(X, \rho)$ are identified with the group of all invertible (ρ, ρ)-adapted (X, X)-matrices.

We now have the following sufficient criterion for $\theta(A)$ to be equal to $\theta(A')$, entirely analogous to Proposition 2.4.

Proposition 3.10. Let (X, ρ) and (Y, σ) be $\mathbb{R}_+$ filtered sets, and let A be a (ρ, σ)-adapted (X, Y)-matrix. Let B and C be (ρ, ρ)-adapted (respectively (σ, σ)-adapted) (X, X)-matrices (respectively (Y, Y)-matrices). Then BAC is also (ρ, σ)-adapted, and the persistence vector space $\theta(A)$ is isomorphic to $\theta(BAC)$.

Remark 3.11. For any $r \in F$, where F is a field, the elementary matrix $e(i, j, r)$ is given by $e_{tt}(i, j, r) = 1$ for all t, $e_{ij}(i, j, r) = r$, and $e_{uv}(i, j, t) = 0$ whenever $(u, v) \neq (i, j)$ and $u \neq v$. Left multiplication by $e(i, j, r)$ has the effect of adding r times the jth row to the ith row, and right multiplication by $e(i, j, r)$ has the effect of adding r times the ith column to the jth column. This observation now suggests that given two $\mathbb{R}_+$ filtered sets (X, ρ) and (Y, σ), and a (ρ, σ)-adapted matrix A, we define an *adapted* row operation to be an operation which adds a multiple of $r(x)$ to $r(x')$, where $\rho(x) \geq \rho(x')$. Similarly, we define an adapted column operation to be an operation which adds a multiple of $c(y)$ to $c(y')$, where $\sigma(y) \leq \sigma(y')$.

We will use this result to classify up to isomorphism all finitely presented persistence vector spaces. We begin by defining a persistence vector space $P(a, b)$ for every pair (a, b), where $a \in \mathbb{R}_+$, $b \in \mathbb{R}_+ \cup \{+\infty\}$, and $a < b$, with the obvious interpretation when $b = +\infty$. $P(a, b)$ is defined by $P(a, b)_r = k$

for $r \in [a,b)$, $P(a,b) = \{0\}$ when $r \notin [a,b)$, and where $L(r,r') = \mathrm{id}_k$ whenever $r, r' \in [a,b)$. This definition can be interpreted in the obvious way when $b = +\infty$. We note that $P(a,b)$ is finitely presented. For, in the case where b is finite, let (X,ρ) and (Y,σ) denote $\mathbb{R}_+$-filtered sets (X,ρ) and (Y,σ), with the underlying sets consisting of single elements x and y, and with $\rho(x) = a$ and $\sigma(y) = b$. Then the (1×1) (X,Y)-matrix matrix $[1]$ is (ρ,σ)-adapted since $a \leq b$, and it is clear that $P(a,b)$ is isomorphic to $\theta([1])$. When $b = +\infty$, $P(a,b)$ is isomorphic to the persistence vector space $V_k(X,\rho)$, and can therefore be written as $\theta(0)$, where 0 denotes the zero linear transformation from the persistence vector space $\{0\}$.

Proposition 3.12. Every finitely presented persistence vector space over k is isomorphic to a finite direct sum of the form

$$P(a_1,b_1) \oplus P(a_2,b_2) \oplus \cdots \oplus P(a_n,b_n) \tag{3.1}$$

for some choices $a_i \in [0,+\infty)$, $b_i \in [0,+\infty]$, and $a_i < b_i$ for all i.

Proof. It is clear that a (ρ,σ)-adapted (X,Y)-matrix A which has the property that every row and column has at most one non-zero element, which is equal to 1, has the property that $\theta(A)$ is of the form described in the proposition. For if we let $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ be all the pairs (x_i,y_i) so that $a_{x_i,y_i} = 1$, then there is a decomposition

$$\theta(A) \cong \bigoplus_i P(\rho(x_i),\sigma(y_i)) \oplus \bigoplus_{x \in X - \{x_1,\ldots,x_n\}} P(\rho(x),+\infty).$$

So, it suffices to construct matrices B and C, which are (ρ,ρ)-adapted (respectively (σ,σ)-adapted) (X,X)-matrices (respectively (Y,Y)-matrices), so that BAC has the property that every row and column has at most one non-zero element, and that element is 1. To see that we can do this, we adapt the row and column operation approach to this setting. The (ρ,σ)-adapted row and column operations consist of all possible multiplications of a row or column by a non-zero element of k, all possible additions of a multiple of $r(x)$ to $r(x')$ when $\rho(x) \geq \rho(x')$, and all possible additions of a multiple of $c(y)$ to $c(y')$ when $\sigma(y) \leq \sigma(y')$. We claim that by performing (ρ,σ)-adapted row and column operations we can arrive at a matrix with at most one non-zero entry in each row and column. To see this, first find a y which maximizes $\sigma(y)$ over the set of all y with $c(y) \neq 0$. Next, find an x which minimizes $\rho(x)$ over the set of all x for which the entry $a_{xy} \neq 0$. Because of the way x is chosen, we are free to add multiples of $r(x)$ to all the other rows so as to 'zero out' $c(y)$ except in the xy-entry. Because of the way y is chosen, we can add multiples of $c(y)$ to zero out $r(x)$ except in the xy-slot, without affecting $c(y)$. The result is a matrix in which the unique non-zero element in both $r(x)$ and $c(y)$ is a_{xy}. By multiplying $r(x)$ by $1/a_{xy}$, we can make the xy-entry in the transformed matrix $= 1$. By deleting $r(x)$

and $c(y)$, we obtain a $(X - \{x\}, Y - \{y\})$-matrix which is (ρ', σ')-adapted, where ρ' and σ' are the restrictions of ρ and σ to $X - \{x\}$ and $Y - \{y\}$, respectively. We can now apply the process inductively to this matrix. Each of the row and column operations required can be interpreted as row and columns on the original matrix, and will have no effect on $r(x)$ or $c(y)$. The result is that by iterating this procedure, we will eventually arrive at a matrix with only zero entries, and it is clear that the transformed matrix has at most one non-zero element in each row and column. The result now follows by Proposition 3.10. □

We will also establish that any two decompositions of the form (3.1) above for a given persistence vector space are essentially unique.

Proposition 3.13. Suppose that $\{V_r\}$ is a finitely presented persistence vector space over k, and that we have two decompositions

$$\{V_r\} \cong \bigoplus_{i \in I} P(a_i, b_i) \text{ and } \{V_r\} \cong \bigoplus_{j \in J} P(c_j, d_j),$$

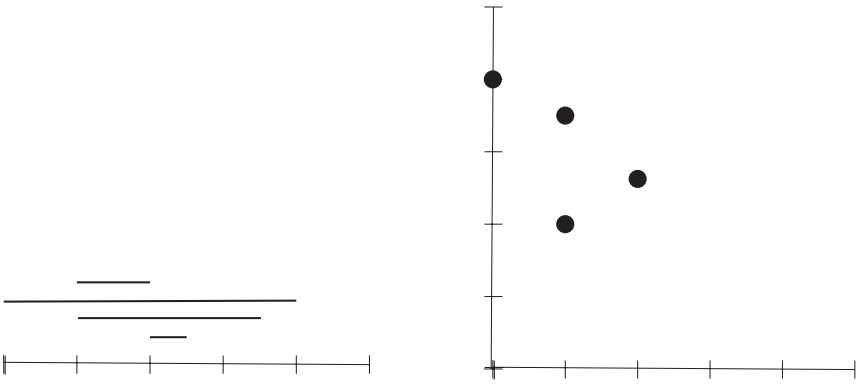
where I and J are finite sets. Then $\#(I) = \#(J)$, and the set of pairs (a_i, b_i) occurring, with multiplicities, is identical to the set of pairs (c_j, d_j) occurring.

Proof. We let $a_{\min}$ and $c_{\min}$ denote the smallest value of a_i and c_j, respectively; $a_{\min}$ can be characterized intrinsically as $\min\{r | V_r \neq 0\}$, and it follows that $a_{\min} = c_{\min}$. Next, let $b_{\min}$ denote $\min\{b_i | a_i = a_{\min}\}$, and make the corresponding definition for $d_{\min}$; $b_{\min}$ is also defined intrinsically as $\min\{r' | N(L(r, r')) \neq 0\}$, where N denotes null space, so $b_{\min} = d_{\min}$ as well. This means that $P(a_{\min}, b_{\min})$ appears in both decompositions. For each decomposition, we consider the sum of all the occurrences of the summand $P(a_{\min}, b_{\min})$. They are both sub-persistence vector spaces of $\{V_r\}$, and can in fact be characterized intrinsically as the sub-persistence vector space $\{W_r\}$, where W_r is the null space of the linear transformation

$$\mathrm{im}(L(a_{\min}, r)) \xrightarrow{L(r, b_{\min}) | \mathrm{im}(L(a_{\min}, r))} V_{b_{\min}}.$$

It now follows that the number of summands of the form $P(a_{\min}, b_{\min})$ in the two decompositions are the same, and further that they correspond isomorphically under the decompositions. Let I' denote the subset of I obtained by removing all i so that $a_i = a_{\min}$ and $b_i = b_{\min}$, and define J' correspondingly. We can now form the quotient of $\{V_r\}$ by $\{W_r\}$, and observe that we obtain identifications

$$\{V_r\}/\{W_r\} \cong \bigoplus_{i \in I'} P(a_i, b_i) \text{ and } \{V_r\}/\{W_r\} \cong \bigoplus_{j \in J'} P(c_j, d_j).$$



(a) barcode (b) persistence diagram

Figure 3.6. Two methods for representing persistent vector spaces.

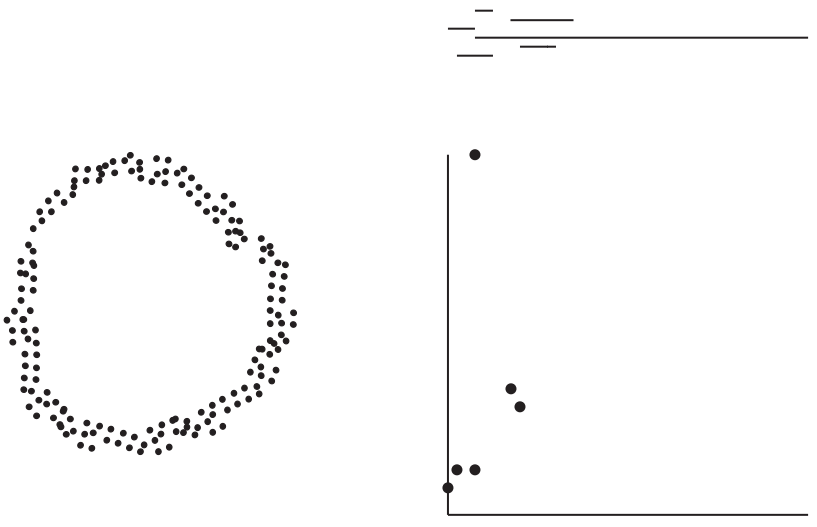
By an induction on the number of summands in the decompositions, we obtain the result. □

We observe that there is an algorithm analogous to the one constructed in Proposition 2.19 for computing the homology, in this case using adapted row and column operations in place of arbitrary operations. This algorithm then produces a presentation for persistent homology.

The isomorphism classes of finitely presented persistence vector spaces are in one-to-one correspondence with finite subsets (with multiplicity) of the set $\{(a,b)|a \in [0,+\infty), b \in [0,+\infty], \text{ and } a < b\}$. Such sets can be represented visually in two distinct ways, one as families of intervals on the non-negative real lines, and the other as a collection of points in the subset $\{(x,y)|x \geq 0 \text{ and } y > x\}$ of the first quadrant in the (x,y)-plane. In the second case, one must place points with $b = +\infty$ above the whole diagram in a horizontal line indicating infinity. The first representation is called a *barcode*, and the second a *persistence diagram*. We will use and refer to these representations interchangeably.

We now have a solution to the problem posed in Section 3.1. We may associate to any finite metric space a persistence barcode or persistence diagram. What has now happened is that the Betti numbers have been replaced by the barcodes. The way to reconcile these two notions is that, roughly speaking, the persistence barcodes often consist of some 'short' intervals and some 'long' intervals. The short intervals are typically considered noisy, and the long ones are considered to correspond to larger-scale geometric features, which one would expect to have a correspondence with the features of a space from which the metric space is sampled.

Figure 3.7 shows the persistence barcode and persistence diagram for one-dimensional homology associated to the sampled version of a circle.



(a) 'statistical circle' (b) representations

Figure 3.7. Barcode and persistence diagram for a sampled circle.

The barcode reflects the fact that the first Betti number is equal to 1 by the fact that it possesses a single long interval and multiple shorter ones.

Remark 3.14. Not all barcodes display this kind of dichotomous behaviour between short and long. This reflects the fact that the metric space might not be representing a simple topological object at a single scale, but rather a multiscale object of interest in its own right. In addition, we will see in the next section that we will devise a number of different methods for generating barcodes, which will reflect more subtle aspects of the shape of the data set.

Remark 3.15. It is clear from the descriptions given above that the complexity of the persistent homology calculations is the same as that for Gaussian elimination, that is, n^3 for an $n \times n$-matrix. What this means is that direct calculation will be extremely expensive. There are several approaches to mitigating this problem.

(1) The α and witness complexes, described in Section 3.6 below, allow us to compute with much smaller complexes, based either on an embedding in a low-dimensional Euclidean space in the case of the α complex, or on a chosen set of landmarks in the case of the witness complex. Both are very practical options, and the algorithms described above apply to that situation as well.

(2) There are also methods for simplifying and drastically reducing the size of the Vietoris–Rips complex, described in Zomorodian (2010).

(3) When we are given a space X with a finite covering by sets $\{U_\alpha\}_{\alpha \in A}$, there is a construction known as the *Mayer–Vietoris blowup*, and a corresponding computational device known as the *Mayer–Vietoris spectral sequence*, which permits the parallelization of the homology computation into calculations of much smaller complexes. See Hatcher (2002) for a discussion of the case of a covering by two sets (the Mayer–Vietoris long exact sequence) and Segal (1968, § 4), for the general case. It proceeds by performing the individual calculations, and then provides a reconstruction step. This procedure has been adapted to the persistent homology situation in Lipsky, Skraba and Vejdemo-Johansson (2011).

Remark 3.16. There are a number of software packages which compute homology and persistent homology, for example:

- CHOMP (http://chomp.rutgers.edu/),
- Javaplex (https://code. google.com/p/javaplex/), and
- Dionysus (http://www.mrzv.org/software/dionysus/).

Remark 3.17. There are a number of theorems which produce theoretical guarantees for the computation of homology via various complexes. The so-called *nerve theorem*, which follows directly from the construction in Segal (1968, § 4), gives sufficient conditions for a much smaller construction based on a covering of the space to compute the homology accurately. Niyogi, Smale and Weinberger (2008) give conditions which show that with high confidence, a construction based on ϵ balls around a finite sample from a submanifold of Euclidean space computes homology of the submanifold accurately.

3.5. Making persistent homology more sensitive: functional persistence

Applying persistent homology naively to many data sets will often produce trivial barcodes, with no long intervals in their barcodes. The reason for this is that data sets often have a central core, to which everything is connected. In the data set shown in Figure 3.8, we see that there is a central core and apparently three 'flares' emanating from it.

Roughly, this is a 'T' or 'Y' shape, and we would not expect to capture that aspect of its shape with homological methods, since these spaces are contractible and therefore have vanishing homology. Similarly, if we are looking at a data set lying along a plane in a high-dimensional space, we would not expect to be able to detect that fact with homology, since a plane is also contractible. In Section 2.7, we were able to adapt topological methods to make them more sensitive by studying auxiliary spaces such as spaces of ends, one-point compactifications, and the results of removing points of various types. In this section, we will see how to adapt persistent homology similarly, so as to be able to capture phenomena of this type.

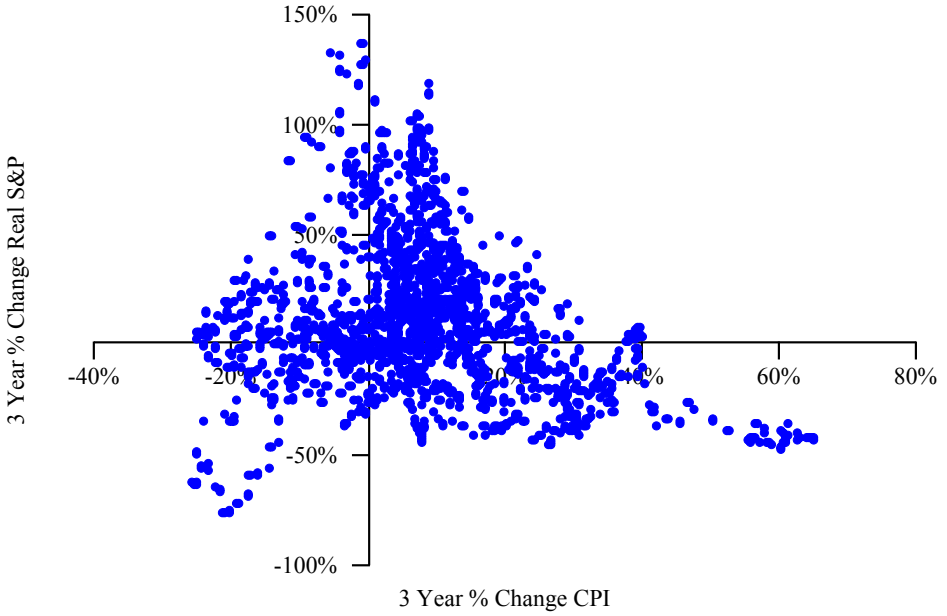


Figure 3.8. Central core and three 'flares'. From http://macromarketmusings.blogspot.com/2007_09_01_archive.html

When first introducing persistent homology, each finite metric space was associated to an increasing family of Vietoris–Rips complexes $\{\mathrm{VR}(X,r)\}_r$, which were used to compute persistent homology. There is another method of constructing increasing families of simplicial complexes.

Definition 3.18 (functional persistence). We let X be a finite metric space, and let $f : X \to \mathbb{R}$ be a non-negative real-valued function. Let us also select a positive real number ρ. Then by the *f-filtered simplicial complex with scale ρ*, we will mean the increasing family of simplicial complexes

$$\{\mathrm{VR}(f^{-1}([0,R]),\rho)\}_R.$$

This construction will have persistence barcodes of its own, which reflect topological properties of the sublevel sets of f. We will call this method of producing persistence diagrams *functional persistence*.

Remark 3.19. Of course, it is not clear how to choose R in general. Often inspection of a data set can suggest the right scale, so that one can obtain useful information. In general, though, it would be much better to be able to construct two-dimensional profiles which encode both the function values and the scale at the same time.

One very interesting family of functions to study in this way is the class of functions measuring the degree of centrality, or *data depth*, of a data point.

(a) (b) (c)

Figure 3.9. Detection of ends by zero-dimensional functional persistence barcode.

For example, consider the family of functions e_p given by

$$e_p(x) = \sum_{x' \in X} d(x, x')^p.$$

One can also define the $p = \infty$ version by setting $e_\infty(x) = \max_{x' \in X} d(x, x')$. Clearly points which are informally closer to the centre of X will have lower values of each of these e_p, and points on the 'periphery' will have larger values. We will use these function values in order to study persistent analogues of the 'ends' which were used for refining ordinary homology in Section 2.7, but we will want to introduce the larger values before the smaller values of the function. For this reason, we will replace e_p by the function $\hat{e}_p$ given by

$$\hat{e}_p(x) = \frac{e_p^{\max} - e_p(x)}{e_p^{\max} - e_p^{\min}},$$

where $e_p^{\max}$ and $e_p^{\min}$ are the maximum and minimum values taken by the function e_p on X. The function $\hat{e}_p$ takes values on $[0, 1]$, and attains both 0 and 1. We will then consider the increasing family of complexes

$$\{\mathrm{VR}(\hat{e}_p^{-1}([0, R]), \rho)\}_R$$

and their persistence barcodes for a couple of examples.

Example 3.20. The data set in Figure 3.9 is shaded by values of $\hat{e}_p$, with light shading corresponding to low values of $\hat{e}_p$, and dark shading to high values. For a small value of R, the sublevel set of $\hat{e}_p$ would look as in image (b). and so the zero-dimensional barcode would be of the form in (c).

Example 3.21. The data set in Figure 3.10 is also shaded by values of $\hat{e}_p$. A sublevel set of $\hat{e}_p$ for a small value of R would be as in (b). Since this point cloud is roughly circular, the one-dimensional barcode would look like (c).

(a) (b) (c)

Figure 3.10. Detection of dimensionality by one-dimensional functional persistence barcode.

In both cases, the behaviour of the set of points 'far from the centre' measures an interesting aspect of the shape of a data set, in the one case a set of clusters and in the other a circle.

3.6. Other constructions of complexes

In Section 3.3, we constructed the Vietoris–Rips complex associated to any metric space. While useful, this construction is often too large to compute with effectively. There are two constructions of complexes that get around this problem, which are often used in place of the Vietoris–Rips complex, which we describe in this section.

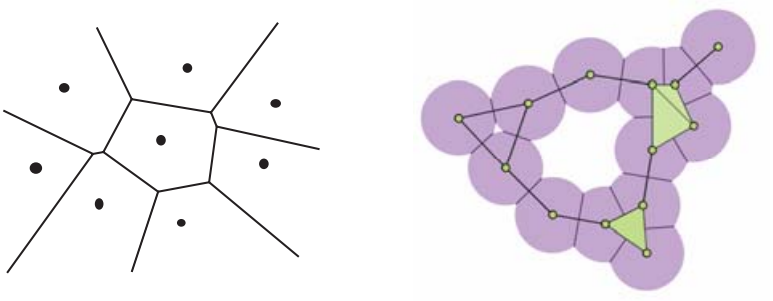
The α complex

This construction is performed on a metric space X which is a subspace of a metric space Y. Typically Y is a Euclidean space $\mathbb{R}^N$, and most often N is small, that is, $N = 2, 3,$ or 4. For any point $x \in X$, we define the *Voronoi cell* of x, denoted by $V(x)$, by

$$V(x) = \{y \in Y | d(x, y) \leq d(x', y) \text{ for all } x' \in X\}.$$

The collection of all Voronoi cells for a finite subset of Euclidean space is called its *Voronoi diagram*. A picture of part of a Voronoi diagram in $\mathbb{R}^2$ is shown in Figure 3.11(a). For each $x \in X$, we also denote by $B_\epsilon(x)$ the set $\{y \in Y | d(x, y) \leq \epsilon\}$. By the α *cell* of $x \in V(x)$ with scale parameter ϵ, we will mean the set $A_\epsilon(x) = B_\epsilon(x) \cap V(x)$. The α *complex with scale parameter* ϵ of a subset $x \in X$, denoted by $\alpha_\epsilon(X)$, will be the abstract simplicial complex with vertex set X, and where the set $\{x_0, \ldots, x_k\}$ spans a k-simplex if and only if

$$\bigcap_{i=0}^{k} A_\epsilon(x_i) \neq \emptyset.$$



(a) (b)

Figure 3.11. (a) Voronoi decomposition, and (b) α-complex.

An example might look as in Figure 3.11(b).

We observe that it follows immediately from the definition that if $\epsilon \leq \epsilon'$, then there is an inclusion $\alpha_\epsilon(X) \hookrightarrow \alpha_{\epsilon'}(X)$, and therefore that for any j, we obtain a persistence vector space

$$\{H_j(\alpha_\epsilon(X))\}_\epsilon.$$

This construction is typically *much* smaller than the Vietoris–Rips complex, even though it has the same vertex set, namely X. The number of simplices included will in general be much smaller. Generically, in fact, all simplices will be of dimension $\leq N$ if the set is embedded in $\mathbb{R}^N$. The α complex is typically computed using algorithms for computing the Voronoi diagram for X as a subset of $\mathbb{R}^N$. These algorithms rapidly become intractable as N increases, which means that the complex can generally be applied only when the point set is included in a low-dimensional Euclidean space, or when the embedding in a Euclidean space can be modified to a low-dimensional embedding via dimensionality reduction techniques. The α complex is discussed in detail by Zomorodian (2005).

The witness complex

A second construction that yields smaller complexes is the *witness complex* (de Silva and Carlsson 2004) in its various forms. The idea here is to use a version of the Voronoi diagram on the data set X itself, rather than on a space in which X is embedded. The vertex set of the complex constructed is smaller than X, consisting of a set of *landmark points* within X. What this means is that we may select the size of complex we are willing to work with.

Definition 3.22. Let X be any metric space, and suppose we are given a finite set $\mathcal{L}$ of points in X, called the landmark set, and a parameter $\epsilon > 0$. For every point $x \in X$, we let m_x denote the distance from this point to the set $\mathcal{L}$, that is, the minimum distance from x to any point in the landmark set. Then we define the strong witness complex attached to

this data to be the complex $W^s(X, \mathcal{L}, \epsilon)$ whose vertex set is $\mathcal{L}$, and where a collection $\{l_0, \ldots, l_k\}$ spans a k-simplex if and only if there is a point $x \in X$ (the witness) so that $d(x, l_i) \leq m_x + \epsilon$ for all i. We can also consider the version of this complex in which the one-simplices are identical to those of $W^s(X, \mathcal{L}, \epsilon)$, but where the family $\{l_0, \ldots, l_k\}$ spans a k-simplex if and only if all the pairs (l_i, l_j) are one-simplices. We will denote this complex by W^s_{VR}.

There is a modified version of this construction, which is quite useful, called the *weak witness construction*. Suppose we are given a metric space X, and a set of points $\mathcal{L} \subseteq X$. Let $\Lambda = \{l_0, \ldots, l_k\}$ denote a finite subset of $\mathcal{L}$. We say a point $x \in X$ is a *weak witness* for Λ if $d(x, l) \geq d(x, l_i)$ for all i and all $l \in \Lambda$. Given ϵ, we will also say that x is an *ϵ-weak witness* for Λ if $d(x, l) + \epsilon \geq d(x, l_i)$ for all i and all $l \notin \Lambda$.

Definition 3.23. Let X, $\mathcal{L}$, and ϵ be as above. By the *weak witness complex* $W^w(X, \mathcal{L}, \epsilon)$ for the given data we will mean the complex whose vertex set is $\mathcal{L}$, and for which a family $\Lambda = \{l_0, \ldots, l_k\}$ spans a k-simplex if and only if Λ and all its faces admit ϵ-weak witnesses. This complex clearly also has a version in which a k-simplex is included as a simplex if and only if all its 1-faces are, and this version is denoted by W^w_{VR}.

One verifies directly from the definition that whenever $0 \leq \epsilon \leq \epsilon'$, we then have an inclusion

$$W^s(X, \mathcal{L}, \epsilon) \hookrightarrow W^s(X, \mathcal{L}, \epsilon'),$$

and therefore that after applying homology, we obtain a persistence vector space $\{H_k(W^s(X, \mathcal{L}, \epsilon))\}_\epsilon$. The same applies to the variant constructions W^s_{VR}, W^w, and W^w_{VR}.

Remark 3.24. The point of introducing the complexes W^s_{VR} and W^w_{VR} is that they require much less memory, since all the information is contained in the zero- and one-simplices.

3.7. Zigzag persistence

Persistence objects (sets, vector spaces, *etc.*) are defined by a collection of objects parametrized by the non-negative real line together with morphisms from the objects with parameter value r to objects with parameter value r', whenever $r \leq r'$. If we restrict the persistence object to the lattice of non-negative integers, we can view a persistence object as a diagram

$$X_0 \to X_1 \to X_2 \to \cdots \to X_n \to \cdots$$

of objects X_n and morphisms $X_n \to X_{n+1}$ for all $n \geq 0$. We can informally state that a persistence object restricted to the integers is equivalent to a

diagram of objects having the shape

where the nodes are objects (sets, vector spaces, simplicial complexes, ...), and the arrows indicate a morphism from one object to another. As such, it is an infinite *quiver diagram* (Derksen and Weyman 2005). Any directed graph Γ is called a *quiver*, and a representation of Γ over a field k is an assignment of a k-vector space V_v to each vertex of Γ and a k-linear transformation $L_e : V_v \to V_w$ for every edge e from v to w. We will also refer to a representation of Γ as a *diagram of shape* Γ. One could also consider diagrams with the shape

Precisely, this corresponds to a family of objects parametrized by the non-negative integers, with a morphism from $X_i \to X_{i+1}$ when i is even, and a morphism $X_{i+1} \to X_i$ when i is odd. We will call such a diagram of k-vector spaces, where k is a field, a *zigzag persistence vector space*. Zigzag vector spaces arise in a number of ways.

Example 3.25. Consider a very large finite metric space X, so large that we do not expect to be able to compute its persistent homology using a Vietoris–Rips complex. We might instead try to form many smaller samples $\{S_i\}$ from X, and attempt to understand how consistent these calculations are. We note that given such a family of samples, we may fix a threshold parameter R and construct $\mathrm{VR}(S_i, R)$ for all i. To attempt to assess the consistency of the computations which arise out of these complexes, we can form the unions $S_i \cup S_{i+1}$ and note that we have inclusion maps $S_i \hookrightarrow S_i \cup S_{i+1}$ and $S_{i+1} \hookrightarrow S_i \cup S_{i+1}$. These maps induce maps on the Vietoris–Rips complexes, and by applying homology to the Vietoris–Rips complexes, we obtain k-vector spaces $V_i = H_j(\mathrm{VR}(S_i, R))$ and $V_{i,i+1} = H_j(\mathrm{VR}(S_i \cup S_{i+1}, R)$. The inclusion maps now mean that we obtain a diagram of the form

$$V_0 \to V_{0,1} \leftarrow V_1 \to V_{1,2} \leftarrow V_2 \to V_{2,3} \leftarrow V_3 \to \cdots .$$

On an intuitive level, consistency between the calculations is measured by the existence of classes $x_i \in V_i$ and $x_{i+1} \in V_{i+1}$, so that the images of x_i and x_{i+1} in $V_{i,i+1}$ are the same non-zero class, and more generally for sequences of classes $\{x_i\}$, with $x_i \in V_i$, so that for all i, the images of x_i and x_{i+1} in $V_{i,i+1}$ are equal to the same non-zero element.

Example 3.26. Given a simplicial complex X and a simplicial map from X to the non-negative real line, triangulated, with the vertex set being the

non-negative integers and the edges being the closed intervals $[n, n+1]$, we can form the subcomplexes $f^{-1}([n, n+1])$, as well as $f^{-1}(n)$. We have the diagram

$$f^{-1}(0) \to f^{-1}([0,1]) \leftarrow f^{-1}(1) \to f^{-1}([1,2]) \leftarrow f^{-1}(2) \to \cdots$$

A scheme which could compute homology only of complexes based at the nodes of this diagram, and which could extract the homology of the entire complex from these calculations, would permit the parallelization of homology computations into smaller pieces. This is of course very desirable.

Example 3.27. In the discussion of the witness complex, we selected a set of landmarks $\mathcal{L}$ from a metric space X, and computed certain complexes $W(X, \mathcal{L}, \epsilon)$, for which there were several variants. We suppress the superscripts s and w, as well as the other labels, since the construction works for all the values. In general, it is difficult to assess how accurately the persistent homology of X is captured by the witness complex. One piece of evidence for accuracy would be a method which assesses how consistent different choices of landmarks are. In order to do this, one can construct a bivariate version of the witness complex which takes as input a pair of landmark sets $(\mathcal{L}_1, \mathcal{L}_2)$, denoted by $W(X, (\mathcal{L}_1, \mathcal{L}_2), \epsilon)$, which maps to each of the complexes $W(X, \mathcal{L}_1, \epsilon)$ and $W(X, \mathcal{L}_2, \epsilon)$ in a natural way. Given a collection of landmark sets $\mathcal{L}_i$, we obtain a diagram of witness complexes

$$W(X, \mathcal{L}_1, \epsilon) \leftarrow W(X, (\mathcal{L}_1, \mathcal{L}_2), \epsilon) \to W(X, \mathcal{L}_2, \epsilon) \leftarrow .$$

After applying homology, one can again ask for consistent families as in Example 3.25 above.

It now turns out that there is a classification theorem for zigzag persistence vector spaces over a field k analogous to that for ordinary persistence.

Definition 3.28. A zigzag persistence k-vector space $V_\cdot$ is called *cyclic* if there are integers $m \leq n$ so that $V_i = k$ for $m \leq i \leq n$, $V_i = \{0\}$ if $i < m$ or $i > n$, and the homomorphisms $V_i \to V_{i+1}$ or $V_{i+1} \to V_i$ are all equal to the identity homomorphism on k whenever $m \leq i \leq i+1 \leq n$.

One example is

$$\{0\} \to k \stackrel{\mathrm{id}}{\leftarrow} k \stackrel{\mathrm{id}}{\to} k \leftarrow \{0\} \to \{0\} \leftarrow \cdots .$$

Each cyclic persistence vector space is indecomposable, in that it cannot be expressed as a direct sum of diagrams. Note that the cyclic zigzag persistence vector spaces are parametrized by intervals $[m, n]$ with integer endpoints. We write $V[m, n]_\cdot$ for the cyclic k-vector space which is exactly non-zero for the integers in $[m, n]$. We have the following analogue of Proposition 3.12.

Theorem 3.29. We say that a zigzag persistence k-vector space $V_\cdot$ is *of finite type* if (a) each vector space V_i is finite-dimensional, and (b) $V_i = \{0\}$ for sufficiently large i. Then, for any zigzag persistence k-vector space $V_\cdot$ of finite type, there is an isomorphism

$$V_\cdot \cong \bigoplus_i V[m_i, n_i]. \tag{3.2}$$

for some choices of pairs of integers (m_i, n_i). Furthermore, the decomposition (3.2) is unique up to a reordering of the summands.

The content of this theorem is that there is a barcode description of the isomorphism classes of zigzag persistence vector spaces, just like that for ordinary persistence vector spaces, except that the intervals in the barcodes are constrained to be integers. This theorem was first proved by Gabriel (1972), and is discussed from the point of view of computational issues by Carlsson and de Silva (2010). This result can now be applied to Examples 3.25 and 3.27 above. The presence of long bars in the barcode decomposition says that there are elements which are consistent across all the vector spaces in the interval determined by the bar. For Example 3.26 above, the zigzag construction can be used to give a very efficient and parallelizable method for computing the homology of the entire complex. These ideas are discussed by Carlsson and de Silva (2010) and Carlsson, de Silva and Morozov (2009).

4. Structures on spaces of barcodes

4.1. Metrics on barcode spaces

We have now associated to any finite metric space a family of persistence barcodes, or persistence diagrams. One important property to understand is the degree to which the barcode changes when we have small (in a suitable sense) changes in the data. In order to even formulate the answer to such a question, we will need to define what is meant by small changes in the barcode. In order to do this, we will define the *bottleneck distance* between barcodes. For any pair of intervals $I = [x_1, y_1]$ and $J = [x_2, y_2]$, we will define $\Delta(I, J)$ to be the l^∞-distance between the two, regarded as ordered pairs in $\mathbb{R}^2$, that is, $\max(|x_2 - x_1|, |y_2 - y_1|)$. For a given interval $I = [x, y]$, we also define $\lambda(I)$ to be $(y - x)/2$. $\lambda(I)$ is the l^∞-distance to the closest interval of the form $[z, z]$ to I. Given two families $\mathfrak{I} = \{I_\alpha\}_{\alpha \in A}$ and $\mathfrak{J} = \{J_\beta\}_{\beta \in B}$ of intervals, for finite sets A and B, and any bijection θ from a subset $A' \subseteq A$ to $B' \subseteq B$, we will define the *penalty* of θ, $P(\theta)$, to be

$$P(\theta) = \max\Big(\max_{a \in A'}(\Delta(I_a, J_{\theta(a)})\Big), \max_{a \in A - A'}(\lambda(I_a)), \max_{b \in B - B'}(\lambda(J_b)))$$

and then define the bottleneck distance $d_\infty(\mathfrak{I}, \mathfrak{J})$ to be

$$\min_\theta P(\theta),$$

where the minimum is taken over all possible bijections from subsets of A to subsets of B. This is easily verified to be a distance function on the set of barcodes.

Remark 4.1. d_∞ is actually the $p = \infty$ version of a family of metrics d_p, called the Wasserstein metrics. The metric d_p is defined via the penalty function P_p, given by

$$P_p(\theta) = \sum_{\alpha \in A'} \Delta(I_\alpha, J_{\theta(\alpha)})^p + \sum_{\alpha \in A - A'} \lambda(I_\alpha)^p + \sum_{b \in B - B'} \lambda(J_b)^p,$$

and we set $d_p(\mathfrak{I}, \mathfrak{J}) = (\min_\theta P_p(\theta))^{\frac{1}{p}}$.

We now have a notion of what it means for barcodes to be close. There is also a notion of what it means for two compact metric spaces to be close, given by the *Gromov–Hausdorff* distance, first defined by Burago, Burago and Ivanov (2001). It is defined as follows. Let Z be any metric space, and let X and Y be two compact subsets of Z. The *Hausdorff distance* between X and Y, $d^H(X, Y)$, is defined to be the quantity

$$\max\Big\{\max_{x \in X} \min_{y \in Y} d_Z(x, y), \max_{y \in Y} \min_{x \in X} d_Z(x, y)\Big\}.$$

Given any two compact metric spaces X and Y, we consider the family $\mathcal{I}(X, Y)$ of all simultaneous isometric embeddings of X and Y. An element of $\mathcal{I}(X, Y)$ is a triple (Z, i_X, i_Y), where Z is a metric space and i_X and i_Y are isometric embeddings of X and Y, respectively, into Z. The Gromov–Hausdorff distance of X and Y is now defined to be the infimum over $\mathcal{I}(X, Y)$ of $d^H(\mathrm{im}(i_X), \mathrm{im}(i_Y))$. It is known to give a metric on the collection of all compact metric spaces. It is also known to be computationally very intractable. Chazal *et al.* (2009) have proved the following.

Theorem 4.2. Fix a non-negative integer k, let $\mathcal{F}$ denote the metric space of all finite metric spaces, and let $\mathcal{B}$ denote the set of all persistence barcodes. Let $\beta_k : \mathcal{F} \to \mathcal{B}$ be the function which assigns to each finite metric space its k-dimensional homology barcode. Then β_k is distance non-increasing.

This result is interesting not only because it gives some guarantees on how certain changes in the data affect the result, but it also provides an easily computed lower bound on the Gromov–Hausdorff distance in a great deal of generality.

There are also stability results for functional persistence, or rather, an analogue of functional persistence. In Definition 3.18, we defined functional persistence based on a function defined on the points of a metric space, or

equivalently on the vertices of its Vietoris–Rips complex. For any topological space X and continuous real-valued function $f : X \to \mathbb{R}$, one can also define associated persistence vector spaces $\{H_i(f^{-1}((-\infty, r])\}_r$, which are very close analogues to the simplicial complex construction described above. In fact, a real-valued function on the vertex set of a simplicial complex can be extended in a natural way to a continuous function on the geometric realization of the simplicial complex, using a weighted sum of values of the function on the vertices based on the *barycentric coordinates* of the point. The idea will be that small changes in the function should yield only small changes in the associated persistence barcode. The results we will describe appear in Cohen-Steiner, Edelsbrunner and Harer (2007) and Cohen-Steiner, Edelsbrunner, Harer and Mileyko (2010).

To state the results, we will need some definitions. Let X be a topological space, and let $f : X \to \mathbb{R}$ be a real-valued function on X. For every non-negative integer k, $a \in \mathbb{R}$, and $\epsilon \in (0, +\infty)$, we have the induced map

$$j = j_{k,a,\epsilon} : H_k(f^{-1}(-\infty, a - \epsilon]) \longrightarrow H_k(f^{-1}(-\infty, a + \epsilon]).$$

We say that a is a *homological critical value of* f if there is a k such that $j_{k,a,\epsilon}$ *fails* to be an isomorphism for all sufficiently small ϵ. Further, we say that the function f is *tame* if it has a finite number of homological critical values and the homology groups $H_k(f^{-1}(-\infty, a])$ are finite-dimensional for all $k \in \mathbb{N}$ and $a \in \mathbb{R}$. This condition holds, for example, in the case of Morse functions on closed manifolds (see Milnor (1963)) and for piecewise linear functions on finite simplicial complexes, so the result we will state is quite generally applicable. The following theorem is proved in Cohen-Steiner *et al.* (2007).

Theorem 4.3. Let X be any space which is homeomorphic to a simplicial complex, and suppose $f, g : X \to \mathbb{R}$ are continuous tame functions. Then the persistence vector spaces $\{H_k(f^{-1}((-\infty, r])\}_r$ and $\{H_k(g^{-1}((-\infty, r])\}_r$ are finitely presented, and therefore admit barcode descriptions for each $k \in \mathbb{N}$, which we denote by $\beta_k f$ and $\beta_k g$. Moreover, for any k, we have that

$$d_\infty(\beta_k f, \beta_k g) \leq \|f - g\|_\infty.$$

There is also a stability result for the Wasserstein distances d_p, with p finite, in the presence of a Lipschitz property for the functions, proved in Cohen-Steiner *et al.* (2010).

4.2. Coordinatizing barcode space

Another way to describe infinite sets is via the theory of *algebraic varieties*, that is, as the set of solutions to a set of equations, either over the real numbers, the complex numbers, or some other field. When this is possible, it gives a very compact description of a large or infinite set. The method

also produces a ring of functions on the set, by restricting the polynomial functions to it. Functions are useful in studying any set by machine learning and other methods. We will discuss the possibilities for producing a coordinatized model of the set of all barcodes.

Let us consider first the set $\mathfrak{B}_n$ of all barcodes containing exactly n intervals or 'bars'. Each of the intervals is determined by two coordinates, the left-hand endpoint x and the right-hand endpoint y. If we have n intervals, we have $2n$ coordinates $\{x_1, y_1, \ldots, x_n, y_n\}$. The trouble is that the barcode space does not take into account the ordering of the intervals, so it is not possible to assign a value to the ith coordinate itself. To understand how to get around this problem, we discuss a familiar situation from invariant theory.

We consider $\mathbb{R}^n$, let Σ_n be the group of permutations of the set $\{1, \ldots, n\}$, and consider the ring of polynomial functions $A_n = \mathbb{R}[x_1, \ldots, x_n]$ on $\mathbb{R}^n$. We will describe how to coordinatize the set of orbits (as defined in Section 2.2) $\mathbb{R}^n/\Sigma_n$, that is, the 'set of unordered n-tuples of real numbers', or equivalently the collection of multisets of size n. If f is a polynomial function on $\mathbb{R}^n$, it may be treated as a function on the set of orbits if it has the property that $f(\sigma\vec{v}) = f(\vec{v})$ for all $\sigma \in \Sigma_n$ and $\vec{v} \in \mathbb{R}^n$. The group Σ_n acts on the ring A_n, and an element $f \in A_n$ is a function on the orbit set $\mathbb{R}^n/\Sigma_n$ if and only if it is fixed under all permutations $\sigma \in \Sigma_n$. The set of all fixed functions (denoted by $A_n^{\Sigma_n}$) is a ring in its own right, and it turns out that it has a very simple description.

Proposition 4.4. The ring $A_n^{\Sigma_n}$ is isomorphic to the ring of polynomials $\mathbb{R}[\sigma_1, \sigma_2, \ldots, \sigma_n]$, where σ_i denotes the *ith elementary symmetric function* given by

$$\sum_{s_1, s_2, \ldots, s_i} x_{s_1} x_{s_2} \cdots x_{s_i},$$

where the sum is over all i-tuples of distinct elements of $\{1, \ldots, n\}$.

We now have coordinates that describe the set $\mathbb{R}/\Sigma_n$. The analogous construction for the set $\mathfrak{B}_n$ would be to form the subring of $\mathbb{R}[x_1, y_1, \ldots, x_n, y_n]$ fixed under the action of Σ_n which permutes the x_i among themselves and the y_i among themselves, and obtain a coordinatization in this way. This does give a ring of functions, but it is not a pure polynomial ring but has *relations*, or *syzygies*. A full discussion of this situation is given in Dalbec (1999). To give an idea of what this means, we first observe that the elementary symmetric functions $\sigma_i(\vec{x})$ and $\sigma_i(\vec{y})$ are definitely invariant, and generate a full polynomial subring of the ring of functions. Let us restrict to the case $n = 2$. Then there is another function $\xi = (x_1 y_1 + x_2 y_2)$, which cannot be expressed in terms of the elementary symmetric functions applied

to $\vec{x}$ and $\vec{y}$. It turns out that there is now an algebraic relation

$$\xi^2 - \sigma_1(\vec{x})\sigma_1(\vec{y})\xi + \sigma_1(\vec{x})^2\sigma_2(\vec{y}) + \sigma_2(\vec{x})\sigma_1(\vec{y})^2 - 4\sigma_2(\vec{x})\sigma_2(\vec{y}) = 0,$$

which after consideration shows that there is no algebraic coordinatization by the four-dimensional affine space but rather by a subset of a higher-dimensional space defined by one or more algebraic equations. Nevertheless, one can express $\mathfrak{B}_n$ as the points of an algebraic variety in this way.

What we would really like to do is to coordinatize the collection of all the sets $\mathfrak{B}_n$ as a variety, in an appropriate sense. One could consider the disjoint union $\coprod_n \mathfrak{B}_n$, but we would rather disregard intervals of length zero in a systematic way, since they correspond to features which are born and die at the same time, and hence do not have any actual persistence. This suggests that we define a set $\mathfrak{B}_\infty$ as follows. We let $\sim$ denote the equivalence relation on $\coprod_n \mathfrak{B}_n$ generated by equivalences

$$\{[x_1, y_1], [x_2, y_2], \ldots, [x_n, y_n]\} \sim \{[x_1, y_1], [x_2, y_2], \ldots, [x_{n-1}, y_{n-1}]\}$$

whenever $x_n = y_n$. We then define the infinite barcode set $\mathfrak{B}_\infty$ to be the quotient

$$\coprod_n \mathfrak{B}_n / \sim .$$

The question we now pose is whether this infinite set can also be coordinatized, again in a suitable sense. To see how this can be done, we consider two simpler examples of coordinatization.

Example 4.5. We let $\mathfrak{A}_n$ denote the set $\mathbb{R}^n/\Sigma_n$, and we let $\mathfrak{A}_\infty$ denote

$$\coprod_n \mathfrak{A}_n / \sim,$$

where $\sim$ is the equivalence relation given by declaring that $(x_1, \ldots, x_n) \sim (y_1, \ldots, y_m)$ if and only there are sets $S \in \{1, \ldots, n\}$ and $T \subseteq \{1, \ldots, m\}$ satisfying the following properties.

(1) $x_s = 0$ and $y_t = 0$ for all $s \in S$ and $t \in T$.
(2) We have $n - \#(S) = m - \#(T)$, and call this common number k.
(3) The unordered k-tuples obtained by deleting the elements corresponding to S and T from $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ respectively, are identical.

There are natural maps from $\mathfrak{A}_n$ to $\mathfrak{A}_\infty$, and the maps are injective on points. We therefore have an increasing system

$$\mathfrak{A}_1 \hookrightarrow \mathfrak{A}_2 \hookrightarrow \mathfrak{A}_3 \hookrightarrow \cdots,$$

which corresponds to a system of ring homomorphisms

$$\mathbb{R}[\sigma_1] \leftarrow \mathbb{R}[\sigma_1, \sigma_2] \leftarrow \mathbb{R}[\sigma_1, \sigma_2, \sigma_3] \leftarrow \cdots,$$

where the homomorphism $\mathbb{R}[\sigma_1, \ldots, \sigma_{n+1}] \to \mathbb{R}[\sigma_1, \ldots, \sigma_n]$ is defined by $\sigma_i \to \sigma_i$ for $1 \leq i \leq n$, and $\sigma_{n+1} \to 0$.

Associated to such a system is its *limit*, which is itself a ring. We will not go into detail about this, but refer to the book by Mac Lane (1998) for background material on limits and their dual construction, a colimit. In this case, the construction produces a ring of functions which can be described as follows. Let $\mathcal{M}$ denote the set of all monomials in the infinite set $\{\sigma_1, \sigma_2, \ldots, \sigma_n, \ldots\}$, and let $\mathcal{M}_n \subseteq \mathcal{M}$ denote the subset of monomials which involve only $\{\sigma_1, \ldots, \sigma_n\}$. Then inverse limit ring is identified with the set of all infinite sums $\sum_{\mu \in \mathcal{M}} r_\mu \mu$ so that the sums $\sum_{\mu \in \mathcal{M}_n} r_\mu \mu$ are all finite. So, for instance, the infinite sum $\sum_n \sigma_n$ is an element of this ring. Elements in this ring certainly define functions on $\mathfrak{A}_\infty$, because the functions σ_N vanish on $\mathfrak{A}_n$ whenever $n \leq N$. The limit ring is, however, a little complicated given the finiteness conditions we are required to impose. It would be simpler to restrict in a natural way to the functions which are polynomial in the elements σ_i. One way to do this is as follows. We may extend the rings $\mathbb{R}[\sigma_1, \ldots, \sigma_n]$ to have complex coefficients, so that they are of the form $\mathbb{C}[\sigma_1, \ldots, \sigma_n]$, and construct the limit ring Λ as in the real case. The group of complex numbers of length one now acts on this ring via the action defined by

$$\zeta \cdot (\sigma_i) = \zeta^i \sigma_i$$

and extended to the limit ring in the obvious way. We say that an element f in Λ is *K-finite* if the span of the entire orbit of f is a finite-dimensional $\mathbb{C}$-vector space. (The notion of K-finiteness is introduced in Knapp.) The intersection of the set of K-finite vectors in Λ with the limit ring over $\mathbb{R}$ can now be identified with the polynomial ring on the variables σ_i.

In the example above, taking the quotient by the equivalence relation $\sim$ did not create complicated rings, as each of the rings $\mathbb{R}[\sigma_1, \sigma_2, \ldots, \sigma_n]$ is a pure polynomial ring in n variables. In general, though, taking quotients by equivalence relations can create rings which are not pure polynomial, and in some cases not even finitely generated as algebras.

Example 4.6. Consider the plane $X = \mathbb{R}^2$, and consider the equivalence relation $\sim$ defined by declaring that $(x, 0) \sim (x', 0)$ for all x, x'. This relation 'collapses' the entire x-axis to a point, while leaving the rest of the set unchanged. The question is whether or not $\mathbb{R}^2/\sim$ can be described as an algebraic variety. As in the case of the ring of invariants of a group action giving a variety structure on an orbit set, in this case we will ask which polynomial functions f on $\mathbb{R}$ have the property that $f(x) = f(x')$ whenever $x \sim x'$. This means that we are asking for the polynomials f in two variables so that $f(x, 0) = f(x', 0)$ for all x, x'. A quick calculation shows that this ring of functions consists of all polynomials in x and y so that the linear term

in x is zero. So, a basis for it is the set of monomials $\{x^i y^j | i > 0 \implies j > 0\}$. This is a ring which is easy to understand, but it is not a finitely generated algebra. It is generated by the elements $\theta_i = x^i y$ together with the element y, and they satisfy the relations $\theta_i^2 = y\theta_{2i}$. We can then obtain a description of $\mathbb{R}^2/\sim$ as the set of points in $\mathbb{R}^\infty = \{(y, \theta_1, \theta_2, \ldots) | \theta_i^2 = y\theta_{2i} \text{ for all } i > 0\}$. Note that in this case infinitely many coordinates will typically be non-zero for a given (x, y).

The image of $\mathfrak{B}_n \to \mathfrak{B}_\infty$, which we denote by $\mathfrak{B}'_n$, can be described as the set obtained from $\mathfrak{B}_n$ by taking the quotient by an equivalence relation $\sim_n$ defined as follows. Given two multisets of intervals (with n intervals) $S = \{[x_1, y_1], [x_2, y_2], \ldots, [x_n, y_n]\}$ and $S' = \{[x'_1, y'_1], [x'_2, y'_2], \ldots, [x'_n, y'_n]\}$, we say that $S \sim_n S'$ if there are subsets $I, I' \subset \{1, \ldots, n\}$, so that $x_i = y_i$ for all $i \in I$, and $x'_{i'} = y'_{i'}$ for all $i' \in I'$, and so that the multisets $S - \{[x_i, x_i] | i \in I\}$ and $T - \{[x'_{i'}, x'_{i'}] | i' \in I'\}$ are identical multisets of intervals. This means that $\mathfrak{B}'_n$ can be expressed as the quotient of $\mathfrak{B}_n$ by an equivalence relation, similar to that described in Example 4.6. It corresponds to a ring $A(\mathfrak{B}'_n)$, which is in general not finitely generated. However, we consider the increasing sequence of sets

$$\mathfrak{B}'_1 \hookrightarrow \mathfrak{B}'_2 \hookrightarrow \mathfrak{B}'_3 \hookrightarrow \cdots$$

and a corresponding system of ring homomorphisms

$$A(\mathfrak{B}'_1) \longleftarrow A(\mathfrak{B}'_2) \longleftarrow A(\mathfrak{B}'_3) \longleftarrow \cdots,$$

as in Example 4.5. This system also has a limit, which we will denote by $A(\mathfrak{B}_\infty)$. There is a construction of K-finite vectors analogous to the one carried out above in the case of the symmetric polynomials, and the ring of K-finite vectors turns out to be isomorphic to a polynomial ring on a set of variables $\{\tau_{ij} | 1 \leq i, 0 \leq j\}$. The variables τ_{ij} correspond to functions on $\mathfrak{B}_\infty$, which can be described as follows. It will suffice to describe τ_{ij} on $\mathfrak{B}_n$, that is, on an unordered n-tuple of intervals. To do this, we first define a function τ'_{ij} on the set of *ordered* n-tuples on intervals by

$$\tau'_{ij}([x_1, y_1], \ldots, [x_n, y_n]) = (y_1 - x_1) \cdots (y_i - x_i) \left(\frac{y_1 + x_1}{2} \right)^j.$$

To obtain τ_{ij} we simply symmetrize by writing

$$\tau_{ij} = \sum_{\sigma \in \Sigma_n} \tau'_{ij} \circ \sigma.$$

So, for example, τ_{10} applied to an unordered n-tuple of intervals is the sum of the lengths of the intervals, τ_{20} is the second elementary symmetric function in the lengths, and τ_{1j} is the sum over all the intervals of the product of the length of the interval with the jth power of its midpoint.

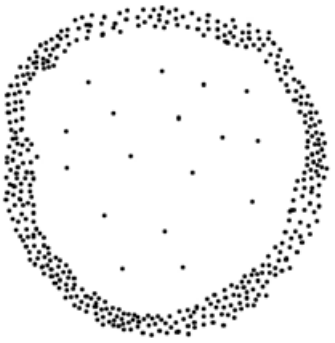


Figure 4.1. A loop with outliers.

4.3. Multidimensional persistence

One of the problems which frequently comes up in persistent homological calculations is that the presence of a few outliers can mask structure. The example in Figure 4.1 is suggestive. Note that the main structure is a sampling from a loop, but there are a number of outliers in the interior of the loop. The homology of the complexes built from this data set will not strongly reflect the loop, because the short connections among the outliers and between the outliers and the points on the actual loop will quickly fill in a disc. In order to remedy this, we would have to find a principled way to remove the outliers in the interior. A measure of density would have this effect, because the outliers by many measures of density would have much lower density than the points on the loop itself. So, if we selected, say, the 90 % densest points by a density measure, we would effectively be choosing the loop itself. However, this choice of threshold is arbitrary, so just as we chose to maintain a profile of homology groups over all threshold values for the scale parameter, we might attempt to create such a profile for values of density.

Another problem with persistent homology has already been made apparent in Section 3.5, where we constructed the increasing family of simplicial complexes $\{\mathrm{VR}(f^{-1}([0,R]),\rho\}_R$ for a non-negative real-valued function f on our data set, and for a *fixed* choice of our scale parameter ρ. These complexes allowed us to study topological features of the data sets which are not captured by the direct application of persistent homology, but using them requires us to make a choice of the scale parameter. There is no universal or natural choice of ρ, so it would be very desirable to be able to track both R and ρ simultaneously.

These considerations motivate the following definition. We let $\mathbb{R}_+ = [0,+\infty)$, and we define a partial order on $\mathbb{R}_+^n$ by declaring that $(r_1,\ldots,r_n) \leq (r_1',\ldots,r_n')$ if and only if $r_i \leq r_1'$ for all $1 \leq i \leq n$.

Definition 4.7. By an *n-persistence vector space* over a field k, we will mean a family of k-vector spaces $\{V_{\vec{r}}\}_{\vec{r}\in\mathbb{R}^n}$ together with linear transformations $L(\vec{r},\vec{r}') : V_{\vec{r}} \to V_{\vec{r}'}$ whenever $\vec{r} \leq \vec{r}'$, so that

$$L(\vec{r}',\vec{r}'') \cdot L(\vec{r},\vec{r}') = L(\vec{r},\vec{r}'')$$

whenever $\vec{r} \leq \vec{r}' \leq \vec{r}''$. There are obvious notions of linear transformations and isomorphisms of n-persistence vector spaces.

We might hope that there is a compact representation of the isomorphism classes of n-persistence vector spaces analogous to the barcode or persistence diagram representations available in 1-persistence, equivalently the ordinary persistence we have already discussed. It turns out, though, that this is not possible, for the following reason. As was observed in Zomorodian and Carlsson (2005), ordinary persistence vector spaces have much in common with the classification of graded modules over the graded ring $k[t]$. Indeed, if we restrict the domain of the scale parameter r in ordinary persistence to the integer lattice $\mathbb{Z}_+ \subseteq \mathbb{R}_+$, then the classification of such restricted persistence vector spaces is identical to the classification of graded $k[t]$-modules. Similarly, it is shown in Carlsson and Zomorodian (2009) that the isomorphism classification of n-persistence vector spaces with the parameter set restricted to $\mathbb{Z}_+^n$ is identical to the classification of n-graded modules over the n-graded ring $k[t_1,\ldots,t_n]$. It is well understood in algebraic geometry that the classification of modules over polynomial rings in more than one variable is fundamentally different from the one-variable case. In particular, for graded one-variable polynomial rings, the parametrization of the set of isomorphism classes is independent of the underlying field k, whereas this does not hold for multigraded polynomial rings in more than one variable. In fact, the classification for more than one variable usually involves *spaces* of structures, rather than discrete sets. What these problems suggest is that we should drop the idea of dealing with a complete classification of the isomorphism classes of n-persistence vector spaces, and instead develop useful invariants which we expect will measure useful and interesting properties of n-persistence vector spaces.

One approach to finding invariants is via the functions defined in Section 4.2. It turns out that some of them can be interpreted in ways which do not depend on obtaining an explicit barcode representation. For any finitely presented persistence vector space $\{V_r\}_r$, we can define two functions attached to V, $\Delta_V : \mathbb{R}_+ \to \mathbb{R}_+$ and $\rho_V : \mathbb{R}_+^2 \to \mathbb{R}_+$. We set

$$\Delta_V(r) = \dim(V_r)$$

and

$$\rho_V(r,r') = \operatorname{rank}(L(r,r')).$$

Example 4.8. We consider the function τ_{10} from Section 4.2, which is given by $\sum_i (y_i - x_i)$. It is easy to check that the equation

$$\tau_{10}(x_1, y_1, \ldots, x_n, y_n) = \int_{\mathbb{R}_+} \Delta_V(r)\, \mathrm{d}r$$

holds for finite barcodes.

Example 4.9. One can also show that

$$\frac{1}{2}(\tau_{10}^2 - 2\tau_{20}) = \frac{1}{2}\sum_i (y_i - x_i)^2 = \int\int_{\mathbb{R}_+^2} \rho_V(r, r')\, \mathrm{d}r\, \mathrm{d}r'.$$

Example 4.10. We also have

$$\tau_{11} = \int_{\mathbb{R}_+} r\Delta_V(r)\, \mathrm{d}r.$$

None of the integrals involve the explicit barcode decomposition, but depend only on the information concerning the dimensions of the spaces V_r and the ranks of the linear transformations $L(r, r')$. The value of this is that Δ_V and ρ_V have direct counterparts in the n-persistence situations, and integrating them (this time over $\mathbb{R}_+^n$ and $\mathbb{R}_+^{2n}$, respectively) yields analogues of at least these invariants in the multidimensional situation. The study of invariants of multidimensional persistence vector spaces is ongoing research.

4.4. Distributions on $\mathfrak{B}_\infty$

We have seen that one can use persistence barcodes to obtain invariants of finite metric spaces which mimic homological invariants for ordinary topological spaces. For finite metric spaces obtained by sampling, one can hope to make an inference on the barcode invariant. For example, if one sees a barcode with a bar which one perceives as 'long', how can one determine if such a long bar could have occurred by chance. More generally, how can one use barcodes to reject a null hypothesis that the sample was obtained from a fixed distribution or family of distributions? In order to make such an inference, one will need to develop a theory of probability measures on $\mathfrak{B}_\infty$, and Mileyko, Mukherjee and Harer (2011) have taken some important first steps in this direction. We summarize their work.

The set $\mathfrak{B}_\infty$ becomes a topological space by equipping it with the quotient topology (see Section 2.2) under the map

$$\coprod_n \mathfrak{B}_n \to \mathfrak{B}_\infty,$$

where each $\mathfrak{B}_n$ is topologized using the quotient topology for $\mathbb{R}^{2n} \to \mathbb{R}^{2n}/\Sigma_n$, and where $\coprod_n \mathfrak{B}_n$ is topologized by declaring that a set in $\coprod_n \mathfrak{B}_n$ is open if and only if its intersection with $\mathfrak{B}_n$ is open for each n. In Section 4.1,

metrics d_p, with the possibility of $p = \infty$, were introduced on $\mathfrak{B}_\infty$. The goal of the work in Mileyko *et al.* (2011) is to study the possibility of defining expectations and variances, in the Fréchet sense (Fréchet 1944, 1948), on the metric spaces $(\mathfrak{B}_\infty, d_p)$. One obstacle to realizing this goal is the fact that the metric space $(\mathfrak{B}_\infty, d_p)$ is not complete. In order to deal with this problem, Mileyko *et al.* (2011) construct completions $\hat{\mathfrak{B}}_p$ of $(\mathfrak{B}, d_p)$. We will first enlarge the set $\mathfrak{B}_\infty$ to include countable multisets of intervals $\{I_a\}_{a\in A}$, where A is a countable set. For a fixed p, the underlying set of $\mathfrak{B}_p$ will be the set of all $\{I_a\}_{a\in A}$ for which

$$\sum_{a\in A} \lambda(I_a) < +\infty. \tag{4.1}$$

It is now clear by inspection that the distance d_p extends naturally to a metric $\hat{d}_p$ on $\hat{\mathfrak{B}}_p$ via the same formula as in Section 4.1: the sum may be infinite, but it is always convergent due to condition (4.1). We now have the following.

Theorem 4.11 (Mileyko *et al.* 2011). The metric space $(\hat{\mathfrak{B}}_p, \hat{d}_p)$ is complete and separable.

The goal of Mileyko *et al.* (2011) is now to construct means and variances on $\hat{\mathfrak{B}}_p$. In general metric spaces X there is no notion of a single point-valued mean, but the mean will actually be a subset of X. Here is the definition of Fréchet means and variances, paraphrased from Mileyko *et al.* (2011).

Definition 4.12. Let X be a metric space, let $\mathcal{B}(X)$ be its Borel σ-algebra, and let $\mathcal{P}$ be a probability measure on $(X, \mathcal{B}(X))$, and suppose that $\mathcal{P}$ has a finite second moment, that is, $\int_X \mathrm{d}(x, x')^2 \, \mathrm{d}\mathcal{P}(x') < \infty$ for all $x \in X$. Then, by the Fréchet variance of $\mathcal{P}$ we will mean

$$\mathrm{var}_{\mathcal{P}} = \inf_{x\in X} \left[F_{\mathcal{P}}(x) = \int_X \mathrm{d}(x, x')^2 \, \mathrm{d}\mathcal{P}(x') \right],$$

and the set

$$\mathbb{E}_{\mathcal{P}} = \{x | F_{\mathcal{P}}(x) = \mathrm{var}_{\mathcal{P}}\}$$

will be called the Fréchet expectation or the Fréchet mean of $\mathcal{P}$.

As stated, $\mathbb{E}_{\mathcal{P}}$ always exists as a set, but it may be empty or contain more than one point. There are results on the non-emptiness and the uniqueness of Fréchet means for manifolds (Karcher 1977, Kendall 1990). What is proved in Mileyko *et al.* (2011) is an existence result for the case $X = \hat{\mathfrak{B}}_p$.

Theorem 4.13. Let $\mathcal{P}$ be a probability measure on $(\hat{\mathfrak{B}}_p, \hat{d}_p)$, and suppose that $\mathcal{P}$ has finite second moment and compact support. Then $\mathbb{E}_{\mathcal{P}} \neq \emptyset$.

Turner, Mileyko, Mukherjee and Harer (2014) made this existence theorem algorithmic, but for a somewhat different choice of metric and metric space. By the L^2 *Wasserstein metric* on $\mathfrak{B}_\infty$, we mean the analogue of d_2 for the choice of penalty function

$$\lambda_{L^2}(I,J) = (x_1 - x_2)^2 + (y_1 - y_2)^2.$$

The resulting metric space is simpler to work with computationally, and all the existence results from Mileyko *et al.* (2011) are also shown to hold in Turner *et al.* (2014).

Another approach concerns the notion of distances between measures on metric spaces, and developed by Blumberg, Gal, Mandell and Pancia (2013) and Chazal, Cohen-Steiner and Merigot (2011). One useful choice is the so-called *Lévy–Prohorov metric*.

Definition 4.14. Given a metric space (X, d_X), the Lévy–Prohorov metric π_X on the set $\mathcal{P}(X)$ of probability measures on the measurable space X, with its σ-algebra of Borel sets $\mathcal{B}(X)$, is defined by

$$\pi_X(\mu,\nu) = \inf\{\epsilon > 0 | \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \leq \mu(A^\epsilon) + \epsilon \text{ for all } A \in \mathcal{B}(X)\},$$

where A^ϵ denotes the ϵ-neighbourhood of A, that is, the union of all balls of radius ϵ about points of A. This metric is defined for all metric spaces, and is known to induce the topology of weak convergence of measures on $\mathcal{P}(X)$ when X is a 'Polish space', that is, a separable complete metric space.

By a *metric measure space* (X, d_X, μ_X), we will mean a metric space (X, d_X) together with a probability measure μ_X on the σ-algebra of Borel sets associated to the metric d_X. Greven, Pfaffelhuber and Winter (2009) defined an analogue to the Gromov–Hausdorff metric on the collection of compact metric measure spaces. Given two metric measure spaces

$$\mathcal{X} = (X, d_X, \mu_X) \quad \text{and} \quad \mathcal{Y} = (Y, d_Y, \mu_Y),$$

the *Gromov–Prohorov distance* between $\mathcal{X}$ and $\mathcal{Y}$ is defined by Greven *et al.* (2009) to be

$$d_{GPr}(\mathcal{X},\mathcal{Y}) = \inf_{(\varphi_X,\varphi_Y,Z)} \pi_{(Z,d_Z)}((\varphi_X)_*\mu_X, (\varphi_Y)_*\mu_Y)$$

where, as in the definition of the Gromov–Hausdorff metric, $(\varphi_X, \varphi_Y, Z)$ varies over all pairs of isometric embeddings $\varphi_X : X \to Z$ and $\varphi_Y : Y \to Z$ of X and Y into a compact metric space Z.

The distributions on $\mathfrak{B}_\infty$ of interest in Blumberg *et al.* (2013) are those which are obtained by sampling n points in X, according to the probability measure μ_X, and computing persistent homology in a fixed dimension k to obtain persistence barcodes. The idea is that each one is a kind of

finite approximation to the metric measure space, and the result shows that the differences between these measures is controlled by the Gromov–Prohorov distance between the actual metric spaces. More precisely, given a metric measure space (X, d_X, μ_X), they construct probability measures $\Phi_k^n = \Phi_k^n(X, d_X, \mu_X)$ on the completed barcode spaces $\hat{\mathfrak{B}}_\infty$ described above, via the formula

$$\Phi_k^n(X, d_X, \mu_X) = (\beta_k)_*(\mu_X^n),$$

where β_k denotes the k-dimensional barcode, and $(-)_*$ denotes the push forward measure to $\hat{\mathfrak{B}}_p$. This formula makes sense because (a) β_k is a continuous map of metric spaces by Theorem 4.2, and (b) both spaces are given the structure of measurable spaces via the Borel σ-algebras. The main theorem of Blumberg *et al.* (2013) is now as follows.

Theorem 4.15 (Blumberg *et al.* 2013, Theorem 5.2). Let (X, d_X, μ_X) and (Y, d_Y, μ_Y) be compact metric measure spaces. Then we have the following inequality relating the Prohorov and Gromov–Prohorov metrics:

$$\pi_{\hat{\mathfrak{B}}_\infty}(\Phi_k^n(X, d_X, \mu_X), \Phi_k^n(Y, d_Y, \mu_Y)) \leq n d_{GPr}((X, d_X, \mu_X), (Y, d_Y, \mu_Y)).$$

This estimate permits us to prove the following convergence result.

Corollary 4.16 (Blumberg *et al.* 2013, Corollary 5.5). Let $S_1 \subset S_2 \subset \cdots \subset \cdots$ be a sequence of randomly drawn samples from (X, d_X, μ_X). We regard S_i as a metric measure space using the subspace metric and the empirical measure. Then $\Phi_k^n(S_i)$ converges in probability to $\Phi_k^n(X, d_X, \mu_X)$.

Remark 4.17. One could formulate similar results in other contexts. For example, a very interesting object of study would be to study the distributions obtained by instead choosing various landmark sets in a witness complex. This would then give an assessment of how well the space is represented by witness complexes of a fixed size.

5. Organizing data sets

5.1. Natural image patches

In this section we will describe the homological analysis of a particular data set coming out of neuroscience and image processing. The data set was constructed by Lee, Pedersen and Mumford (2003), and the analysis was done by Carlsson, Ishkhanov, de Silva and Zomorodian (2008). We first discuss the data set. It consists of images from a black and white digital camera. Each such image can be thought of as a *pixel vector* in a space whose dimension is the number of pixels used in the camera, and where the coordinate for a given pixel is a grey-scale value, which we will think of as a continuous variable, but which in fact takes integer values between 0

and 255. The 'metaproblem' proposed by Lee *et al.* was to understand in some sense the structure of the set of all images that might actually occur in taking images, versus the set of all possible pixel vectors. The usual topological method of trying to understand a space makes little sense here, due to the extremely high dimensionality of the space of images. Lee *et al.* decided instead to study the statistics of small (3×3) patches within a large database of images taken by two Dutch neuroscientists around Groningen in the Netherlands (van Hateren and van der Schaaf 1998).

Each such patch was taken as a vector in $\mathbb{R}^9$, with one grey-scale value for each of the nine pixels. One preliminary observation is that most images contain large solid regions, and therefore that patches which are constant or nearly constant will dominate the statistics of the set of patches. For this reason, Lee *et al.* decided to study only *high-contrast* patches, that is, patches $\{x_{ij}\}_{1 \leq i,j \leq 3}$ such that a certain positive definite quadratic form on the vector of differences $\{x_{ij} - \mu\}$ is large, where μ is the mean value of the values x_{ij}. The quadratic form is called the *D-norm*, and is a rescaling of the standard Euclidean form in a way which reflects the placement of the pixels in the 3×3 array. They then constructed a data set of 4×10^6 high-contrast patches, by which they mean patches whose D-norm lies in the top 20 % values which occur. They then proceeded to perform the following transformations to this data set.

(1) For any vector $\vec{x} = \{x_{ij}\}_{ij}$, let $\mu(\vec{x})$ denote the mean of the entries x_{ij}. Replace each vector $\vec{x}$ by $\vec{x}' = \vec{x} - \mu(\vec{x})$. This is done so as to understand the structure of the patch independent of the absolute values of the grey-scale values.

(2) Replace each vector $\vec{x}'$ by $\vec{x}'/\|\vec{x}'\|_D$. This is done so as to understand the patch structure independently of the absolute value of the contrast in the patch.

We now have a data set $\mathcal{M}$ consisting of 4×10^6 points lying on a seven-dimensional ellipsoid E^7 within an eight-dimensional subspace of $\mathbb{R}^9$. This is the data set we will study.

The first inclination is to apply persistent homology directly. Unfortunately, it turns out that $\mathcal{M}$ fills out the full E^7, which would mean that were we to carry out the calculation, we would expect to find the homology of E^7. This would be unsurprising, and would not yield useful insights about $\mathcal{M}$. One can observe, though, that the density suitably measured varies a great deal within $\mathcal{M}$. This means that we can threshold by density, and obtain a data set consisting of the most frequently appearing patches, or motifs, in $\mathcal{M}$. Such information would clearly be useful, for example, in constructing compression schemes. The analysis in Carlsson *et al.* (2008)

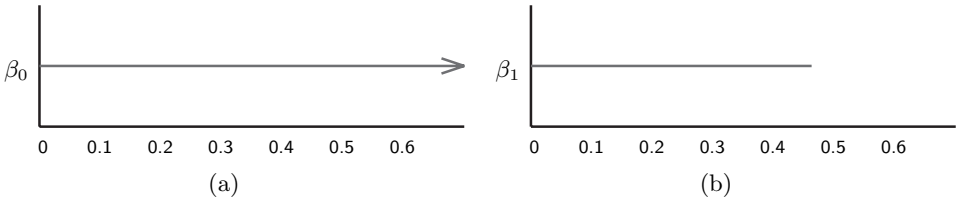


Figure 5.1. (a) Zero-dimensional and (b) one-dimensional persistence barcodes for $X(300, 30)$.

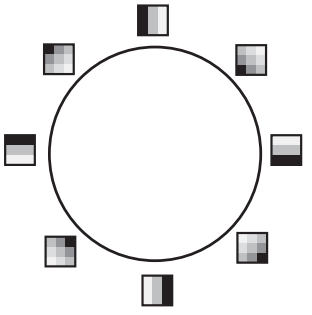


Figure 5.2. Primary circle in the data set $\mathcal{M}$.

uses a very simple proxy for density, called *co-density*. The *co-density with parameter* k on a finite metric space, denoted by $\rho_k(x)$, is defined by

$$\rho_k(x) = \{d(x, x')| \text{ where } x' \text{ is the } k\text{th nearest neighbour to } x\}.$$

Note that ρ_k varies inversely with actual density, and also that the parameter k plays a role similar to the choice of variance in a Gaussian kernel density estimator. The size of the parameter k determines whether the density is computed more or less locally, that is, using smaller or larger balls around a point x. For a data set X, we let $X(k, p)$, where k is a positive integer and p is a percentage, denote the set of points x for which the value of $\rho_k(x)$ is among the top $p\,\%$ of the values occurring in X. Figure 5.1 shows the zero- and one-dimensional persistence barcodes occurring for $X(300, 30)$, where X is a set of size 5×10^4 sampled at random from $\mathcal{M}$.

Note that the zero-dimensional barcode, in Figure 5.1(a), shows a single long bar, indicating that the space is well approximated by a connected space. The one-dimensional barcode, in Figure 5.1(b), also has a single long bar, indicating that it might very well be modelled by a circle. This is indeed the case, as suggested by the circular coordinatization of the patches in Figure 5.2. What it demonstrates is that the highest-density patches, according to ρ_{300}, consist of the discrete versions of linear functions in two variables. This is not so surprising, but it is evidence that the method works.

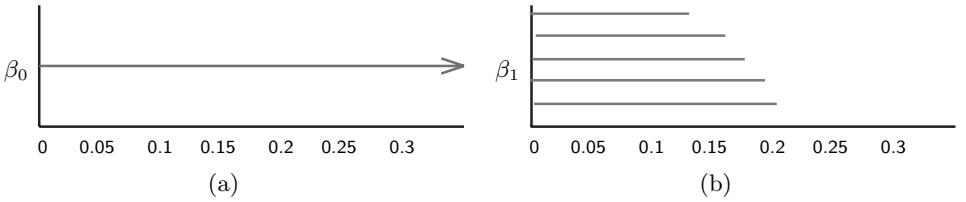


Figure 5.3. (a) Zero-dimensional and (b) one-dimensional persistence barcodes for $X(15, 30)$.

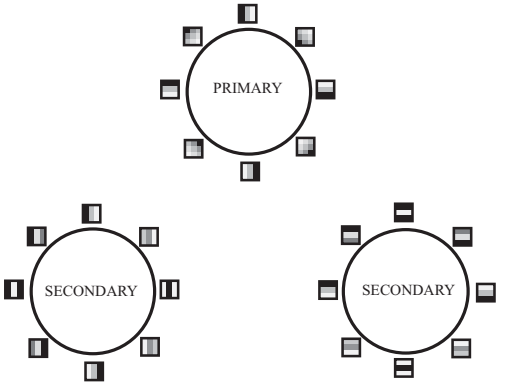


Figure 5.4. Primary and secondary circles in $\mathcal{M}$.

One could then ask what happens when we use a more local density proxy, that is, a smaller value of k. Figure 5.3 shows the result for $X(15, 30)$, where X is again a subsample from $\mathcal{M}$ of size 5×10^4. In this case, we again have a single long line in the zero-dimensional barcode, indicating connectedness, but the one-dimensional barcode has five long bars, suggesting that β_1 for this data set should be 5. We note that we obtain the same pattern, with five long bars in the β_1 barcode for different subsamples, so it appears to be recognizing a real feature of the data set. In Figure 5.4, note that the 'north' and 'south' points on the left secondary circle are identical to the north and south points on the primary circle, and the east and west points of the right-hand secondary circle are identical to the east and west points on the primary circle. This suggests that the space be identified with a collection of three circles, one labelled 'primary' and the other two labelled 'secondary', with two points in each of the secondary circles identified with two points in the primary circle, and with no overlap between the two secondary circles. This gives us the idealized pictures in Figure 5.5. The model in Figure 5.5(b) is obtained by stretching and moving the secondary circles; it is clear that $\beta_1 = 5$, since there are five independent loops.

(a) (b)

Figure 5.5. (a) Three-circle model, and (b) three-circle model deformed.

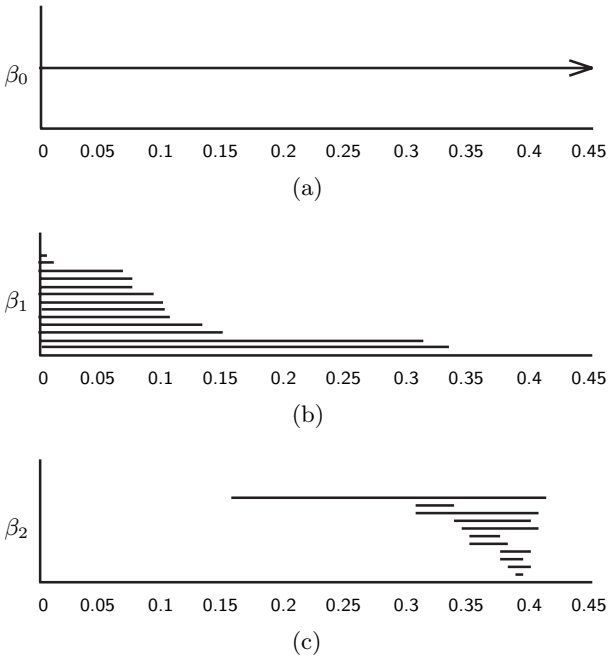
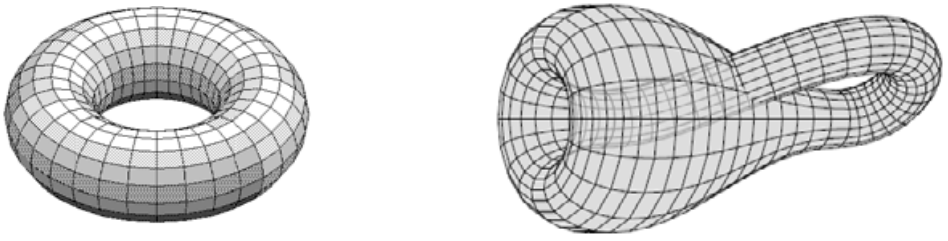


Figure 5.6. (a) Zero-dimensional, (b) one-dimensional, and (c) two-dimensional persistence barcodes.

One can now ask whether there is a natural two-dimensional object into which the data fits, and in such a way that the three-circle model naturally fits into it as well. In this case, because we are trying to obtain a two-dimensional geometry, we are actually forced to use the full data set $\mathcal{M}$, rather than a subsample. We also have to choose the k-parameter so that the density estimator is even more localized. Figure 5.6 shows the persistence barcodes in dimensions 0, 1 and 2 in this case.



(a) (b)

Figure 5.7. (a) Torus and (b) Klein bottle.

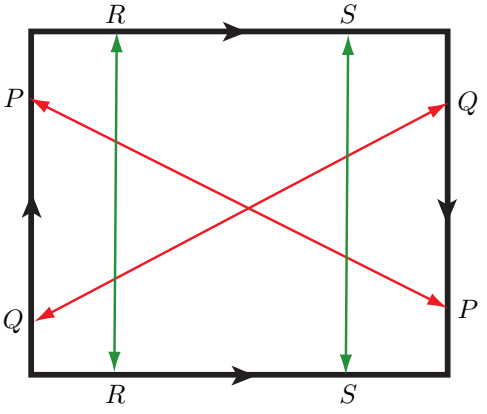


Figure 5.8. Identification space model for the Klein bottle.

Figure 5.6(a) suggests that the space is connected. Note that in the one- and two-dimensional barcodes (Figure 5.6(b,c)) there is an interval of significant length (roughly from 0.15 to 0.30) in which there are two bars in the one-dimensional barcode and a single line in the two-dimensional barcode, suggesting that for that range of values of the scale parameter we have a complex with $\beta_1 = 2$ and $\beta_2 = 1$, with $\mathbb{F}_2$ coefficients. There is a classification of all compact two-dimensional manifolds (Munkres 1975, Chapter 12), from which one can deduce that the only two possible such manifolds are the torus and the Klein bottle, as in Figure 5.7. A calculation for the field $\mathbb{F}_3$ distinguishes between these two, and the persistent homology calculations shows that the Klein bottle is the correct choice. The mod 3 Betti numbers of the torus are $(1, 2, 1)$ and for the Klein bottle they are $(1, 1, 0)$.

It is interesting to explore how the three-circle model might fit inside the Klein bottle. In order to do this, we will need a model of it more suitable for computation. One way of doing this is via an identification space model, which can be pictured as in Figure 5.8. The points labelled P, Q, R, and S

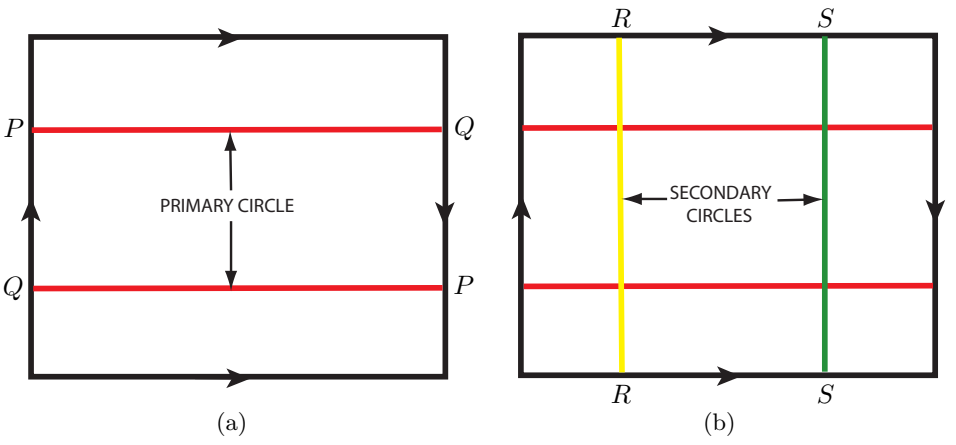


Figure 5.9. (a) Primary circle and (b) secondary circles in the Klein bottle.

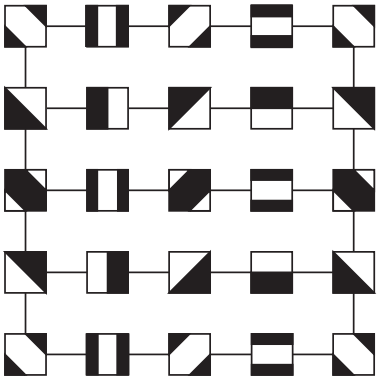


Figure 5.10. Parametrization of high-contrast patches by the Klein bottle.

are identified with other points labelled P, Q, R, and S respectively. This means we are taking the quotient topology as in Section 2.2 of the relation in which all the equivalence classes have either just one element (when the point is in the interior of the rectangle) or exactly two elements, when they are on the boundary of the disc. In this case, points along the horizontal boundary lines are identified with their counterparts on the opposite line, and the points along the vertical boundary lines are identified with the points they correspond to when we perform a reflection of the opposite segment. Given this model, we can see how the three-circle model fits inside this space.

Figure 5.10 is a schematic describing how the high-contrast patches are parametrized by the Klein bottle. Note that the parametrization respects the equivalence relation, in that patches on the boundary are identical to the corresponding patches on the opposite segment.

Having this kind of description of the geometric structure of the data set can be useful in a number of ways.

(1) *Platonic model.* We have studied discretizations of patches obtained by sampling on 3×3 grids. But we could also ask what fully continuous model might be consistent with our understanding of $\mathcal{M}$. We will do this by considering the intensity function of the patch as a quadratic polynomial of degree two, in two variables, on the unit square. We now describe a subspace of the set of all quadratic polynomials in two variables (which have six parameters: one constant, two linear, and three quadratic coefficients) by the following requirements.

- $\int_D f = 0$, where D denotes the unit square.
- $\int_D f^2 = 1$.
- The function f is of the form $f(\vec{x}) = q(\lambda(\vec{x}))$, where q is a single-variable quadratic polynomial, and where λ is a linear functional on the plane.

The first two conditions are analogous to the requirements that the mean intensity be zero and the contrast be one, which were imposed on the data set in Lee *et al.* (2003). They define a subset which is a four-dimensional ellipsoid S^4 in $\mathbb{R}^6$. The final condition, which is non-linear, defines a two-dimensional subspace of S^4 which is homeomorphic to the Klein bottle, as can be readily verified.

(2) *Compression schemes.* An understanding of what the frequently occurring data points are in a data set is often an important part of any compression scheme. For example, the *wedgelet* compression scheme (Donoho 1999) uses the information that the primary circle in our description consists of frequently occurring patches to obtain an interesting compression scheme. Maleki, Shahram and Carlsson (2008) have constructed a scheme based on the Klein bottle which outperforms wedgelets on some particular images. The rate distortion curve is given in Figure 5.11, where the Klein bottle compression scheme is the upper curve, and the lower curves are two different versions of the wedgelet scheme.

(3) *Texture recognition.* In many image processing situations, we are interested in recognizing not only large-scale features but also properties which are more related to the texture of regions of the image. For example, in studying textures, we expect to deal instead with the statistics of small features within the patch. One approach to this problem is studied in Perea and Carlsson (2014). The idea is to study all high-contrast patches occurring inside a (larger) texture patch statistically, in the hope that the statistics will tell the difference between textures.

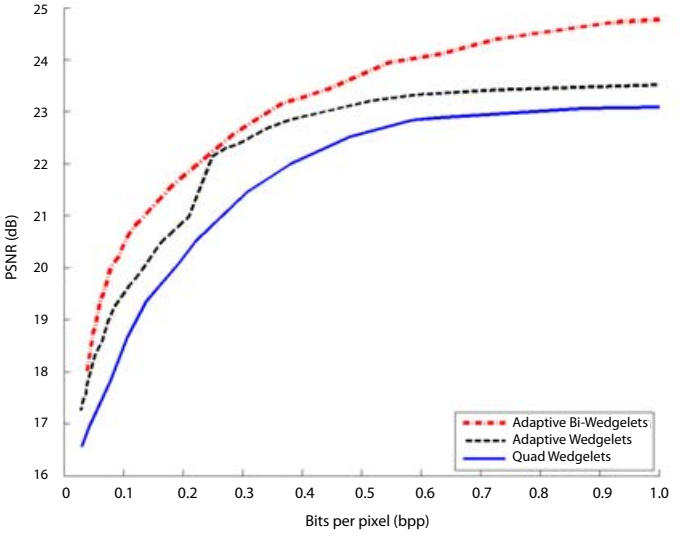


Figure 5.11. Rate distortion curve for Klein bottle-based compression.

One way to understand the statistics is to locate, for each high-contrast patch, the closest point on the Klein bottle to that patch. If we carry out this process, we obtain a large collection of points on the Klein bottle. We can then generate a probability density function on the Klein bottle by smoothing. Since the Klein bottle has a very simple geometry (it has the torus as a two-fold covering space), we can perform Fourier analysis on the probability density function, to obtain coordinates which are the Fourier coefficients. It turns out that these coefficients are able to distinguish textures in various databases quite effectively, close to state of the art. The advantage of this method is that the effect of rotation on an image gives an easy transformation on the Klein bottle, namely translation in the horizontal direction in the identification space model in Figure 5.1. The standard approach to this problem is to identify a finite set of patches (the codebook), and to compute dot products of a given patch with each of them, giving a distribution function on a finite set. The key observation here is that we can deal with infinite codebooks if the codebook is equipped with a useful geometry, as is the Klein bottle. Actual finiteness can be replaced by *finiteness of description*.

5.2. Databases of compounds

In the previous section we saw that persistent homology can be a useful tool for understanding the overall structure of a complex data set. In this section, we will discuss how persistence barcodes can instead be applied to

the individual elements, or data points, in databases where these elements themselves carry a geometric structure.

Databases of organic compounds are of fundamental importance in many biomedical applications. They are, for example, the main object of study in the area of drug discovery. The possibility of developing methods for interrogating them, and in particular to determine rapidly and simply which compounds are functionally similar to which other compounds, is something that should be explored. One approach to this problem is to somehow assign metric structures to sets of compounds. One difficulty is that sets of compounds do not fit neatly into standard formats for data, such as spreadsheets. The data specifying a particular compound consist of position coordinates for its atoms as well as information about bonds between the atoms. Since different compounds consist of different numbers of atoms as well as different numbers of bonds, there is no simple vector description with a fixed number of coordinates that can describe a set of compounds. In addition, the description is not unique, since a molecule could be rotated to produce different coordinates for the atoms. There is a standard method for encoding the structures of compounds, referred to as *simplified molecular-input line-entry system*, or SMILES. In this system, the compounds are represented as a list of symbols encoding the atoms, as well as other annotation symbols which permit the reconstruction of the compound. Three problems with this representation are that (a) there is not always a unique SMILES representation of a given molecule, (b) the conversion from SMILES to a three-dimensional representation can be ambiguous, and (c) the structure is such that there is no obvious way of assigning a distance function or similarity score to a pair of SMILES structures. Bak and Lerner (2014) have described a method for assigning distance functions to sets of compounds which circumvents these difficulties. We now describe their approach.

Any compound can be described as a finite collection of atoms with positions in three-dimensional Euclidean space, together with a collection of bonds connecting the atoms, and the lengths of the bonds. There are now two ways to assign a metric space structure to the set of atoms.

(1) The three-dimensional Euclidean distance between atoms.

(2) By regarding the bonds as the edges in a weighted graph with the weight being the bond length, we may define an edge-path metric, where the distance between two atoms α and β is the minimum over all k and all sequences $\alpha_0, \alpha_1, \ldots, \alpha_k$ for which each pair (α_i, α_{i+1}) is connected by a bond, of the sum

$$\sum_i \lambda(\alpha_i, \alpha_{i+1}),$$

where $\lambda(\alpha_i, \alpha_{i+1})$ denotes the length of the bond connecting α_i and α_{i+1}.

A key observation concerning these metrics is that they are independent of rotations of the molecules. Bak and Lerner (2014) have constructed various persistence barcodes, some using geometric constructions as in Section 3.6, and others using functional persistence (see Definition 3.18) based on various functions on the set of atoms of the molecule. The geometric constructions are of two types, namely the Vietoris–Rips complex constructed in Definition 3.2 and the α complex defined in Section 3.6. The α complex is based on the embedding of the set of atoms in three-dimensional Euclidean space. The fact that the atoms are embedded in a low-dimensional Euclidean space is what makes the α complex construction tractable. The collection of functions on the set of atoms used for the functional persistence barcodes is as follows:

(1) the centrality function $\hat{e}_2(x)$, defined in Section 3.5,
(2) the atomic mass of the atom,
(3) the partial charge of the atom.

In the case of functional persistence barcodes, a choice of scale must be made. The choices made by Bak and Lerner (2014) are a finite set of small multiples of the carbon–carbon bond length within the molecule. Given these choices, persistence barcodes are now constructed in dimensions 0, 1, 2 and 3. Let us index the set of choices of barcodes under consideration by $\beta \in B$, where B is an indexing set. Then, for each β, we assign the bottleneck distance or Wasserstein metric with $p = 2$, and label it d_β for that particular complex construction. We can then assign an associated weighted family of metrics

$$d_B^{\vec{a}} = \left(\sum_{\beta \in B} a_\beta d_\beta^2 \right)^{\frac{1}{2}}$$

to any selection of non-negative real-valued vectors $\vec{a} = \{a_\beta\}_\beta$.

Bak and Lerner study a particular class of compounds called *dihydrofolate reductase* (or DFHR) inhibitors. These are compounds that inhibit the functioning of dihydrofolate reductase in the biosynthesis of purines, thymidylate, and other important amino acids. The synthesis which is associated with DFHR is important in the development of cancer, and therefore inhibiting its functioning is a useful property in the management of cancer. DFHR inhibition is the basis for Methotrexate, which is the first historical example of structure-based drug design applied to cancer. The idea in Bak and Lerner (2014) is now to develop a database of drugs which might have some likelihood of being a DFHR inhibitor. To develop this database, they first observed that one necessary property of a DFHR inhibitor is

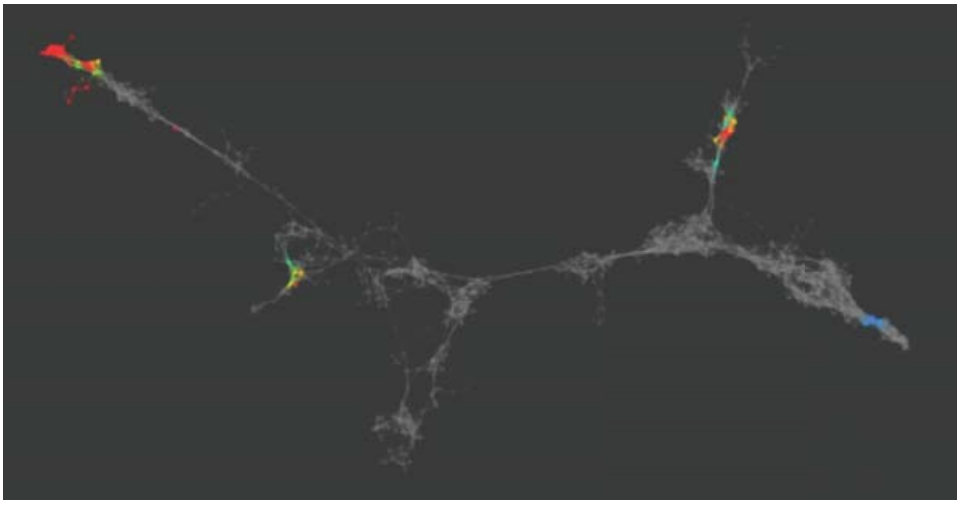


Figure 5.12. Data set constructed using the 'Mapper' methodology.

that it should contain at least one aromatic group and one hydrophobic portion. The necessity for this criterion comes from the observation that all known DFHR inhibitors have this property. They then constructed a database of three-dimensional conformations, extracted from commercial or open-source databases, or in some cases constructed by hand. There are large databases for which this has already been done, and relatively standard ways of doing this. One such method is ZINC: http://zinc.docking.org/ (Irwin and Schoichet 2005). For other molecules, that do not have a three-dimensional structure, they used a tool called OMEGA to generate conformations, and used the lowest-energy conformation. Inhibitors in species other than humans were used. The compounds were combined into a single dataset with 4000 compounds. They then worked with an optimization scheme to perform 'metric learning' to find the best values of $\vec{a}$ in selecting the metric. This optimization scheme was based on an objective function defined for the so-called *DBSCAN clustering*. The metric selected was the one optimizing this objective function. The metric they constructed produced very good localization of the actual DFHR inhibitors within the larger database of putative DFHR inhibitors. Figure 5.12 is an image of the data set constructed using the 'Mapper' methodology: see Carlsson (2009), Singh, Mémoli and Carlsson (2007), Nicolau, Levine and Carlsson (2011), and Lum *et al.* (2013). What it shows is that the DFHR inhibitors exist as four separate very localized groups. Understanding this grouping is interesting, but it means that one can now test a new compound to see where it belongs in the image network, by selecting the point in the existing data set which is closest to the new point. It will therefore act as a mechanism

for rejecting out of hand many compounds, which then do not have to be tested in the laboratory.

Bak and Lerner also used the algebraic functions defined in Section 4.2 to provide coordinates which can be used as input to more sophisticated machine learning methods. It was applied to the problem of classifying DFHR inhibitor versus non-inhibitor status based only on the barcode.

5.3. Viral evolution

The phylogenetic tree has become the standard model to capture the evolution of species since Darwin. A single tree, the 'Tree of Life', contains all species, alive and extinct, into a single structure. The revolution in genomic technologies in the last 20 years has led to an explosion of data and methods developed to infer tree structure on sets of sequences (Felsenstein 2004, Drummond, Nicholls, Rodrigo and Solomon 2002). This phylogenetic model strictly models *clonal* evolution, when genetic material is obtained from a single lineage of ancestors. In this case mutations can occur and then be transmitted via replication of the genomic material from a single parent to the offspring. It is known that there are numerous other mechanisms for the transfer of genetic material from one organism to another, where the organisms may even belong to different species. Recombination of genomic material is common in many species. But examples across species are pervasive in nature, such as hybrids in plants, horizontal gene transfer in bacteria, and fusions of genomes, as occurred with mitochondria and chloroplasts. It has been argued by Doolittle (1999) that trees do not adequately model the full range of possible mechanisms for transfer of genetic material, thereby producing conflicting results in the reconstruction of the Tree of Life. It turns out that all trees have vanishing homology groups, and that therefore homological methods might be used to provide evidence for the presence of horizontal evolutionary events within a collection of sequences, or in their genetic history. This idea was proposed by Chan, Carlsson and Rabadan (2013), who outlined a dictionary of correspondences between algebraic topology and evolutionary concepts. The power of homology to study genomic data was shown with examples arising from viral evolution.

The data sets consist of genetic sequences, which are sequences from a four-element alphabet given by $\{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$, corresponding to the nucleotides adenine, guanine, cytosine, and thymine. One natural metric on the set of such sequences is the *Hamming distance*, which assigns to a pair of sequences $\{x_i\}_i$ and $\{y_i\}_i$, with $x_i, y_i \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$, the number of values of i for which $x_i \neq y_i$. Equivalently, one can describe the Hamming distance as the minimal number of substitutions which have to be made in one sequence to obtain the other. There are many variants on this distance which take into account the rates at which the various substitutions take

place and assign different numbers to each possible substitution. Several variations of these distances were studied by Chan *et al.* (2013), with similar results. The Nei–Tamura results were displayed. At this point we may apply the persistent homology techniques to such finite metric spaces, and obtain persistence barcodes.

Since what we are trying to do is to distinguish some of these metric spaces from trees, it is important to understand the behaviour of persistent homology of spaces which are somehow 'tree-like'. Let Γ denote a weighted graph, that is, a triple $(V(\Gamma), E(\Gamma), f_\Gamma)$, where $V(\Gamma)$ is a finite set, $E(\Gamma)$ is a subset of the collection of two-element subsets of $V(\Gamma)$, and where $f_\Gamma : E(\Gamma) \to (0, +\infty)$ is a weighting function on the edges. By an *edge-path* in Γ, we will mean a sequence $\{v_0, v_1, \ldots, v_n\}$ of elements $v_i \in V(\Gamma)$ such that for every $0 \leq i \leq n-1$, we have that $\{v_i, v_{i+1}\}$ is an element of $E(\Gamma)$. For any edge path $\mathfrak{e} = \{v_0, v_1, \ldots, v_n\}$, the *length* of $\mathfrak{e}$, denoted by $\lambda(\mathfrak{e})$, is given by

$$\lambda(\mathfrak{e}) = \sum_{i=0}^{n-1} f_\Gamma(\{v_i, v_{i+1}\}).$$

The distance between any two vertices v and v' of Γ, $d_\Gamma(v, v')$, is given by

$$d_\Gamma(v, v') = \min_{\mathfrak{e}} \lambda(\mathfrak{e}),$$

where the minimum is taken over all edge-paths $\mathfrak{e} = \{v_0, \ldots, v_n\}$ with the property that $v_0 = v$ and $v_n = v'$. We denote the metric space $(V(\Gamma), d_\Gamma)$ by $\mathfrak{M}(\Gamma)$. We now have the following result concerning the positive-dimensional persistent barcodes of a metric space $\mathfrak{M}(\Gamma)$.

Proposition 5.1. Let Γ be any weighted graph, and assume that the underlying undirected graph is a *tree*, that is, it admits no cycles. Then the persistence barcodes $\beta_i(\mathfrak{M}(\Gamma))$ are all trivial, that is, they are the empty multiset of bars, whenever $i > 0$.

Remark 5.2. This result uses in an essential way the result proved by Buneman (1974) that a metric space is of the form $\mathfrak{M}(\Gamma)$ if and only if it satisfies the so-called four-point condition, which asserts that a finite metric space (X, d) is of the form $\mathfrak{M}(\Gamma)$ for some weighted tree if and only if the condition

$$d(x, y) + d(z, t) \leq \max(d(x, z) + d(y, t), d(x, t) + d(y, z))$$

holds for all quadruples of elements $x, y, z, t \in X$.

The effect of this result is that one can determine that a finite metric space is not tree-like, or is isometric to a finite subset of a space of the form $\mathfrak{M}(\Gamma)$. This statement alone is not useful in studying data, since there will

be no 'exactly tree-like' metric spaces occurring in the presence of noise. However, we have Theorem 4.2, which shows us that if a finite metric space is approximately a tree-like metric space, that is, close to a tree-like metric space in the Gromov–Hausdorff metric, then its higher-dimensional barcodes will be close to the empty barcode in the bottleneck distance. This means that all the intervals contained in the barcode are of small length.

Chan *et al.* (2013) applied persistent homology in one and two dimensions to data sets arising in viral evolution, using the Vietoris–Rips and witness complex defined in Section 3.6. A summary of their results is as follows (see Chan *et al.* 2013 for details).

(1) The type of exchange of genomic material can be catalogued by the topology of the genomic data. In the case of strict clonal evolution, all information is captured by the zero-dimensional homology. Segmented viruses, that is, viruses whose genomic information is encoded in different 'chromosomes' or segments, undergo reassortments: the generation of novel viruses with combinations of segments from the different parental strains. This phenomenon is the source of most of the reported influenza pandemics in humans (the H2N2 pandemic in 1957 and H3N2 in 1968), where a novel viral strain is generated by reassorting segments of different parental strains infecting different hosts. A data set of sequences arising in avian influenza was studied from this point of view. When each distinct segment was studied, there was no higher-dimensional behaviour. However, the metric space produced by using the concatenated sequences did produce significant one-dimensional homology, which would preclude phylogeny as a complete explanation. Other viruses, such as HIV, the cause of AIDS, undergo recombination, where the genome of the offspring is a mosaic of the genome of the parents. Circulating recombinant forms (CRFs) are examples of HIV viruses with complex recombinant patterns. Persistent homology shows higher-dimensional homology groups capturing some of the complex recombinant structures.

(2) *Estimating the rate and scale of horizontal evolution.* By performing persistent homological calculations on simulations, it was observed that a lower bound for the recombination/reassortment rate was obtained by studying the counts of the number of bars in the corresponding barcode. By analysing the numbers of these events it was possible to assess the evolutionary pressures that link different pieces of genomic material together. For instance, when analysing the homology of avian influenza A, it was shown that several segments, containing the genes that encode the polymerases, are more likely to co-segregate together, possibly indicating that natural selection does not allow all combinations to be equally fit.

(3) Higher-dimensional homology generators capture genomic exchange. In cases where higher-dimensional homology was observed, a representative cycle for a given feature (or bar) was obtained. Such a cycle is a linear combination of pairs of landmark points in the case of one-dimensional homology, of triples of landmark points in the case of two-dimensional homology, and so on. The cycle then gives a list of data points which occur, and this list was studied. In the cases studied, these occurrences were consistent with the known horizontal transfer mechanisms in these situations. It should be pointed out that the representative cycle was chosen directly from the algorithms used to compute the persistent homology, and no effort was made to find the minimal cycle, that is, one with the smallest number of summands in the formal sum. Choosing the minimal cycle would be expected to give even more focused outcomes. One might also attempt to reconcile the findings of homology with results obtained via the mapping methods introduced in Singh *et al.* (2007).

(4) Explicit examples were studied: the triple reassortant avian virus that led to the outbreak in China in March 2013, the rate of reassortment of influenza A viruses infecting different hosts (humans, swine and birds), HIV recombination, and many other viruses including dengue virus, hepatitis C virus, West Nile virus, and rabies.

5.4. Time series

Time series are a very interesting class of data, where for a discrete variable t we have a point x_t in a metric space (X, d). Usually X is a Euclidean space $\mathbb{R}^n$. Given such a time series and a positive integer l, we can construct a data set out of the time series $\{x_t\}_t$ as follows. Let $\mathfrak{T}_l(\{x_t\}_t)$ denote the set of all fragments $(x_{t_i}, x_{t_{i+1}}, \ldots, x_{t_{i+l}})$ for every possible initial time t_i. From the distance function d on X, we can define a distance function d_T on $\mathfrak{T}_l(\{x_t\}_t)$ by the formula

$$d_T((x_{t_i}, x_{t_{i+1}}, \ldots, x_{t_{i+l}}), (x_{t_j}, x_{t_{j+1}}, \ldots, x_{t_{j+l}})) = \left(\sum_{s=0}^{s=l} d(x_{t_i+s}, x_{t_j+s})^2\right)^{\frac{1}{2}}.$$

To give an idea of how this works, we consider a time series given by

$$x_t = \sin(t\epsilon),$$

where $\epsilon > 0$ is a very small number, or threshold. If $l = 0$, then the data set we obtain is just the range of the sine function restricted to the set of multiples of the threshold ϵ. If ϵ is sufficiently small, it will roughly fill out the interval $[-1, 1]$, the range of the sine function. On the other hand, if $l = 1$, then the data points are pairs $(\sin(t\epsilon), \sin((t+1)\epsilon))$. A sketch of this data set in the case where $\epsilon = \pi/4$ is shown in Figure 5.13. It is easy to

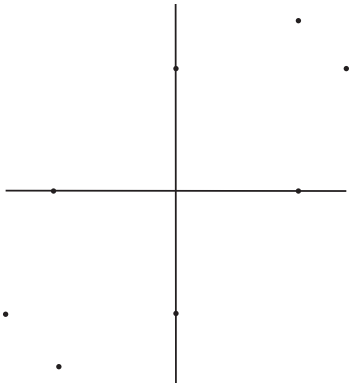


Figure 5.13. A space of time series fragments.

check, using trigonometric identities, that this set is a discrete set of points lying on the ellipse

$$x^2 - \frac{\sqrt{2}}{2}xy + y^2 = \frac{1}{2}.$$

The fact that this data set contains a loop reflects the fact that the sine function is periodic. If we performed the same construction for a time series of the form $x_t = t\epsilon$, we would instead obtain a set of points sampled from the line with equation $y = x + \epsilon$. This set has no loops, and this reflects the fact that the linear function $f(x) = x\epsilon$ is not periodic. These observations suggest the possibility that computing the one-dimensional persistent homology groups of the metric spaces $\mathfrak{T}_l(\{x_t\}_t)$ could be a strategy for detecting periodic behaviour of a given function. This method is proposed by Perea and Harer (2014), and some of its properties are explored. We summarize their results below.

For any real-valued function f on the real line, we will define the *sliding window embedding* attached to f and with parameters $M \in \mathbb{N}$ and $\tau \in (0, +\infty)$, denoted by $SW_{M,\tau}f(t)$, by

$$SW_{M,\tau}f(t) = (f(t), f(t+\tau), \ldots, f(t+M\tau)),$$

where $SW_{M.\tau}f$ is an $\mathbb{R}^{M+1}$-valued function on $\mathbb{R}$. By choosing a finite set of points $T \in [0, L]$, we obtain a point cloud attached to the function f, which we denote by $\mathfrak{C} = \mathfrak{C}(f, M, \tau, T)$. It is the persistent homology of $\mathfrak{C}$, using the Vietoris–Rips complex, which will reflect periodic behaviour. To formulate the main result of Perea and Harer (2014), we make some definitions. For any L^2 function on the circle S^1, we let $S_n f$ denote its Nth Fourier truncation, that is, the linear combinations of $\sin(nt)$ and $\cos(nt)$, for $n \leq N$, occurring in the Fourier decomposition for f. For any finite set $T \subseteq S^1$, we can then apply the sliding window embedding for $S_N f$,

based on M and τ, and restrict it to T, to obtain a point cloud in $(M+1)$-dimensional Euclidean space $Y = Y(f, T, N, M, \tau)$. We then apply a process which mean-centres every data point in Y, and then normalizes it to lie on the unit sphere in $\mathbb{R}^{M+1}$. That is, for each data point $\underline{x} = (x_0, x_1, \ldots, x_M)$, we first transform it to a vector $\hat{\underline{x}} = (\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_n)$, where

$$\hat{x}_i = x_i - \frac{1}{M+1}\left(\sum_i x_i\right).$$

Finally, we form $\underline{x}^* = \hat{\underline{x}}/\|\hat{\underline{x}}\|$. We will denote the point cloud which has been transformed in this way by $\overline{Y}(f, T, N, M, \tau)$. We next assume the function f is L-periodic on $\mathbb{R}$, so that

$$f\left(x + \frac{2\pi}{L}\right) = f(x).$$

We then set

$$\tau_N = \frac{2\pi}{L(2N+1)}.$$

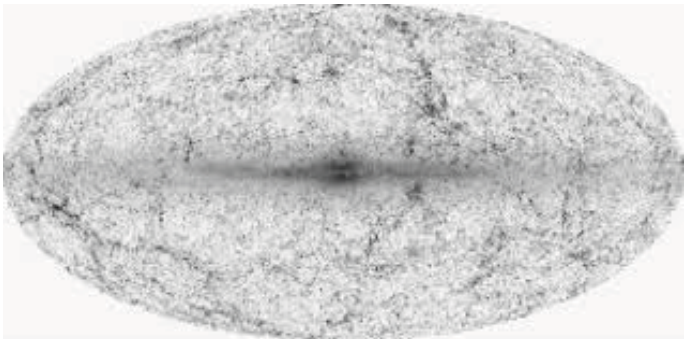
It is then shown in Perea and Harer (2014) that the limit of the sequence $\operatorname{dgm} \overline{Y}(f, T, N, 2N, \tau_N)$ exists as a point $\operatorname{dgm}_\infty(f, T, w)$ in the completed barcode space $\hat{\mathfrak{B}}_\infty$, based on the bottleneck distance, defined in Section 4.4, where dgm denotes the operation which assigns to a point cloud its one-dimensional persistence diagram, or barcode, and where $w = 2\pi/L$. For any barcode $\beta = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ in $\mathfrak{B}_\infty$, we define its *maximal persistence* $\operatorname{mp}(\beta)$ by the equation $\operatorname{mp}(\beta) = \max_i(y_i - x_i)$. It is easy to check that mp extends to a function on $\hat{\mathfrak{B}}_\infty$. Perea and Harer (2014) have also proved that one may take a limit over families of subsets $T_\delta \subseteq S^1$, where T_δ is δ-dense in the sense that every point in S^1 is within a distance δ of some point in T_δ, and where $\delta \to 0$, to obtain a limit $\operatorname{dgm}_\infty(f, w)$. One of the principal results of Perea and Harer (2014) is as follows.

Theorem 5.3. Let f be an L-periodic continuous function on S^1 such that $\hat{f}(0) = 0$ and $\|f\|_2 = 1$, and suppose that the persistence diagram is computed for homology with coefficients in $\mathbb{Q}$. Then we have that

$$\operatorname{mp}(\operatorname{dgm}_\infty(f, w)) \geq 2\sqrt{2} \max_{n \in \mathbb{N}} |\hat{f}(n)|,$$

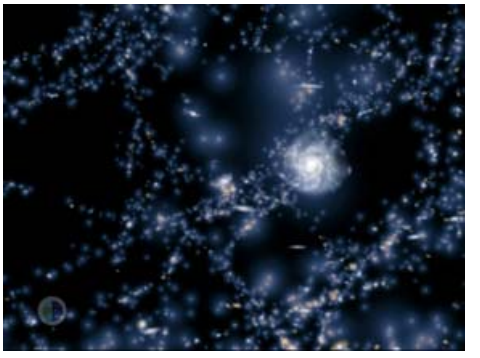
where $\hat{f}(n)$ is given by $\hat{f}(n) = \frac{1}{2}a_n - \frac{i}{2}b_n$ if $n > 0$, $= \frac{1}{2}a_{-n} + \frac{i}{2}b_{-n}$ is $n < 0$, and $= a_0$ if $n = 0$, where a_n and b_n denote the coefficients of $\cos(nt)$ and $\sin(nt)$, respectively, in the Fourier expansion of f.

This is a very interesting relationship, and these observations are used as the basis for a test for periodicity based on persistence barcodes of the point clouds constructed from the function f. Perea and Harer (2014) made



(a)

(b)



(c)

Figure 5.14. Distribution of stars, galaxies and other structures: (a) 2Mass redshift survey, (b) NASA/Goddard, (c) ESO/IAC.

comparisons with more conventional methods for detecting periodicity, with favourable results.

5.5. Structure of the cosmic web

A very interesting object of study in cosmology and astrophysics is the distribution of stars, galaxies, and other structures. Far from being uniform, it exhibits very intricate local structure, as indicated by the images in Figure 5.14, from Huchra *et al.* (2005) (a), NASA's Goddard Space Flight Center (b), and the European Southern Observatory and Instituto de Astrofisica de Canarias (c). Galaxies and mass exist in a wispy web-like spatial arrangement consisting of dense compact clusters, connected by elongated filaments, and sheet-like walls, which surround large near-empty void regions; see Bardeen, Bond, Kaiser and Szalay (1986) for a discussion of the relevant cosmology. To understand this kind of 'three-dimensional texture', one can attempt to study it as a topological problem. It is usually formulated via an assumption that the presence of matter is obtained by sampling

from a density function ρ defined on three-dimensional space. The density ρ would itself be a complicated object, given the complexity of the resulting sampling. The first recognition that topological methods could be useful in this setting occurred in Gott, Dickinson and Melott (1986) and Hamilton, Gott and Weinberg (1986). The idea was that one should study the level sets of ρ from a topological point of view. These level sets are two-dimensional surfaces, for which there is a natural integer invariant called the *genus*, which is defined as one half of its first Betti number. Gott *et al.* (1986) and Hamilton *et al.* (1986) considered the genus of the level surfaces as a useful numerical invariant of this situation. Note that the genus differs from the Euler characteristic, defined below, by a constant. Another family of spaces one could study is the *excursion sets* at various levels r, defined as $\rho^{-1}([r, +\infty))$. It was recognized in Sousbie (2011) and Sousbie, Pichon and Kawahara (2011) that persistent homology gives a method which allows us to track topological behaviour of excursion sets as the thresholds for these excursion sets decrease.

In order to develop theory, and to understand the appearance of observed results, one can assume that ρ is itself a random function, or a random field, obtained under some stochastic process. The idea is that one could attempt to understand expected values for the Betti numbers of the excursion sets of random fields. Random fields have been studied in great detail in Adler (1981) and Adler and Taylor (2007). A case of particular interest is that of a *Gaussian random field*, in which the assumption is made that the distribution assumed by the values of the random field at any particular point or any finite family of points in the domain is Gaussian. Gaussian fields in a cosmological context are of key importance because of the following facts.

(1) The primordial universe was, to high precision, a spatial Gaussian random field, which has been shown observationally by numerous cosmic microwave background experiments such as COBE (Bennett *et al.* 2003), WMAP (Spergel *et al.* 2007), and more recently Planck (Abergel *et al.* 2011).

(2) There is also a fundamental physical reason to expect that all primordial structure is nearly completely Gaussian. This concerns the generation of primordial density perturbations during the early inflationary phase of the universe (at $t \sim 10^{-36}$ seconds after the Big Bang), during which quantum fluctuations were blown up to macroscopic proportions.

(3) A more technical mathematical reason why Gaussian fluctuations are expected is the central limit theorem. Given the fact that the fluctuations at each scale are independently distributed, it would naturally give rise to a Gaussian field.

This assumption permits us to understand many properties of the random field, including some topological ones.

There is another useful integer-valued summary of the homology of any finite-dimensional space called the *Euler characteristic* of the space. It is defined to be the alternating sum

$$\chi(X) = \sum_i (-1)^i \beta_i(X).$$

Note that there is an Euler characteristic for each different field of coefficients. The Euler characteristic is a very useful invariant due to the fact that it is so readily computable, as is demonstrated in the following result.

Proposition 5.4. Let X be any finite simplicial complex. Then $\chi(X)$ is equal to the alternating sum

$$\sum_i (-1)^i \sigma_i(X),$$

where $\sigma_i(X)$ denotes the number of i-dimensional simplices in the complex X.

What this means is that the Euler characteristic can be computed directly by counting simplices, without performing any linear algebra. The price one pays is, of course, that much information about the actual homology is lost.

Example 5.5. Let S^n denote the n-sphere. Then $\chi(S^n) = 0$ if n is odd, and 2 if n is even. So, while the Betti numbers can distinguish between all spheres, the Euler characteristic can only determine the parity of the dimension of the sphere.

It turns out that for many Gaussian random fields, as shown by Adler and Taylor, it is possible to determine the distribution of the statistic given by the Euler characteristic of the excursion sets, in terms of certain parameters defining the random field. These beautiful results are described in detail in Adler and Taylor (2007). An important case (when the Gaussian field is characterized by its *power spectrum*) was obtained in the cosmological literature (Bardeen *et al.* 1986, Hamilton *et al.* 1986). What this suggests is the possibility that we can begin to evaluate models for the stochastic processes from which the distribution in the cosmic web is generated, provided that we can develop sufficient theory to do so.

More recently, Park *et al.* (2013) and van de Weygaert *et al.* (2011) have shown that the Betti numbers of Gaussian random fields carry strictly more information concerning the random field than the Euler characteristic alone. Specifically, it is found that an invariant of the Gaussian field called the *slope of the power spectrum* affects the shape of the curves of the Betti numbers (as the threshold r changes), while the shape of the curve of Euler characteristics is independent of it.

Persistent homology has also been studied in this way. Adler *et al.* (2010) have defined the analogue of the Euler characteristic for persistent homology and studied it for Gaussian random fields. The analogue of the Euler characteristic for persistent homology is defined as follows. For a fixed barcode $\beta = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define

$$\tau(\beta) = \sum_i (y_i - x_i),$$

and then define the analogue of the Euler characteristic, χ^{pers} by the formula

$$\chi^{\text{pers}}(X) = \sum_i (-1)^i \tau(\beta_i(X)),$$

where $\beta_i(X)$ denotes the i-dimensional persistence barcode for X. This quantity has the same properties which make the Euler characteristic highly computable, and in Adler *et al.* (2010) a result is proved which computes χ^{pers} for Gaussian random fields.

Finally, van de Weygaert *et al.* (2011) have taken a different topological approach to analyse the topology at various scales of a set of discrete points in two- or three-dimensional space. The points are to be thought of as individual stars, or perhaps galaxies, and the idea is to construct the α complex (see Section 3.6) associated to them at various scales. This method is used by van de Weygaert *et al.* (2011) to study the Betti numbers produced by various stochastic process models, including some which are explicitly proposed for cosmological problems. In addition, they point out that persistent homology is applicable to this family of complexes, and will likely be applied in future investigations.

Acknowledgements

The author is very grateful for helpful conversations with many people, including R. Adler, A. Bak, A. Blumberg, E. Carlsson, J. Carlsson, F. Chazal, J. Curry, V. de Silva, P. Diaconis, H. Edelsbrunner, R. Ghrist, L. Guibas, J. Harer, S. Holmes, M. Lesnick, A. Levine, P. Lum, B. Mann, F. Mémoli, K. Mischaikow, D. Morozov, S. Mukherjee, J. Perea, R. Rabadan, H. Sexton, P. Skraba, G. Singh, R. van de Weijgaert, M. Vejdemo-Johansson, S. Weinberger, and A. Zomorodian.

REFERENCES[1]

A. Abergel *et al.* (2011), 'Planck early results XXIV: Dust in the diffuse interstellar medium and the galactic halo', *Astron. Astrophys.* **536**, A24.

R. Adler (1981), *The Geometry of Random Fields*, Wiley Series in Probability and Mathematical Statistics.

R. Adler and J. Taylor (2007), *Random Fields and Geometry*, Springer Monographs in Mathematics.

R. Adler, O. Bobrowski, M. Borman, E. Subag and S. Weinberger (2010), Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applications: A Festschrift for Lawrence D. Brown*, Vol. 6, pp. 124–143.

A. Bak and M. Lerner (2014), in preparation.

J. Bardeen, J. Bond, N. Kaiser and A. Szalay (1986), 'The statistics of peaks of Gaussian random fields', *Astrophys. J.* **304**, 15–61.

C. Bennett, R. Hill, G. Hinshaw, M. Nolta, N. Odegard, L. Page, D. Spergel, J. Weiland, E. Wright, M. Halpern, N. Jarosik, A. Kogut, M. Limon, S. Meyer, G. Tucker and E. Wollack (2003), 'First-year Wilkinson Microwave Anisotropy Probe (WMAP)* observations: Foreground emission', *Astrophys. J.* Supplement Series **148**, 97.

A. Blumberg, I. Gal, M. Mandell and M. Pancia (2013), Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. arXiv:1206.4581

O. Bonnet (1848), 'Mémoire sur la théorie generale des surfaces', *Journal de l'École Polytechnique* **19**, 1–146.

P. Buneman (1974), 'A note on the metric properties of trees', *J. Combin. Theory* B **17**, 48–50.

D. Burago, Y. Burago and S. Ivanov (2001), *A Course in Metric Geometry*, Vol. 33 of *Graduate Studies in Mathematics*, AMS.

G. Carlsson (2009), 'Topology and data', *Bull. Amer. Math. Soc.* **46**, 255–308.

G. Carlsson and V. de Silva (2010), 'Zigzag persistence', *Found. Comput. Math.* **10**, 367–405.

G. Carlsson and A. Zomorodian (2009), 'The theory of multidimensional persistence', *Discrete Comput. Geom.* **42**, 71–93.

G. Carlsson, T. Ishkhanov, V. de Silva and A. Zomorodian (2008), 'On the local behavior of spaces of natural images', *Internat. J. Computer Vision* **76**, 1–12.

G. Carlsson, V. de Silva and D. Morozov (2009), Zigzag persistent homology and real-valued functions. In *Proc. 25th Annual Symposium on Computational Geometry*, ACM, pp. 247–256.

G. Carlsson, A. Zomorodian, A. Collins and L. Guibas (2005), 'Persistence barcodes for shapes', *Internat. J. Shape Modeling* **11**, 149.

J. Chan, G. Carlsson and R. Rabadan (2013), 'Topology of viral evolution', *Proc. Nat. Acad. Sci.* **110**, 18566–18571.

[1] The URLs cited in this work were correct at the time of going to press, but the publisher and the authors make no undertaking that the citations remain live or are accurate or appropriate.

F. Chazal, D. Cohen-Steiner, L. Guibas, F. Memoli and S. Oudot (2009), Gromov–Hausdorff stable signatures for shapes using persistence. In *Eurographics Symposium on Geometry Processing 2009. Computer Graphics Forum* **28**, 1393–1403.

F. Chazal, D. Cohen-Steiner and Q. Merigot (2011), 'Geometric inference for probability measures', *Found. Comput. Math.* **11**, 733–751.

D. Cohen-Steiner, H. Edelsbrunner and J. Harer (2007), 'Stability of persistence diagrams', *Discrete Comput. Geom.* **37**, 103–120.

D. Cohen-Steiner, H. Edelsbrunner, J. Harer and Y. Mileyko (2010), 'Lipschitz functions have L^p-stable persistence', *Found. Comput. Math.* **10**, 127–139.

J. Dalbec (1999), 'Multisymmetric functions', *Beiträge Algebra Geom.* **40**, 27–51

H. Derksen and J. Weyman (2005), 'Quiver representations', *Notices Amer. Math. Soc.* **52**, 200–206.

D. Donoho (1999), 'Wedgelets: Nearly minimax estimation of edges', *Ann. Statist.* **27**, 859–897.

W. Doolittle (1999), 'Phylogenetic classification and the universal tree', *Science* **284** (5423), 2124–2128.

A. Drummond, F. Nicholls, A. Rodrigo and W. Solomon (2002), 'Estimating mutation parameters, population history, and genealogy simultaneously from temporally spaced sequence data', *Genetics* **161**, 1307–1320.

H. Edelsbrunner, D. Letscher and A. Zomorodian (2002), 'Topological persistence and simplification', *Discrete Comput. Geom.* **28**, 511–533

S. Eilenberg (1944), 'Singular homology theory', *Ann. of Math.* **45**, 407–447.

L. Euler (1741), 'Solutio problematis ad geometriam situs pertinentis', *Commentarii Academiae Scientiarum Petropolitanae* **8**, 128–140.

L. Euler (1758*a*), 'Elementa doctrinae solidorum', *Novi Commentarii Academiae Scientiarum Petropolitanae* **4**, 109–140. *Opera Omnia* (1) **26**, 72–93.

L. Euler (1758*b*), 'Demonstratio nonnullarum insignium proprietatum quibus solida hedris planis inclusa sunt praedita', *Novi Commentarii Academiae Scientiarum Petropolitanae* **4**, 140–160. *Opera Omnia* (1) **26**, 94–108.

J. Felsenstein (2004), *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA.

M. Fréchet (1944), 'L'intégrale abstraite d'une fonction abstraite d'une variable abstraite et son application à la moyenne d'un élément aléatoire de nature quelconque' *Rev. Sci.* **82**, 483–512.

M. Fréchet (1948), 'Les éléments aléatoires de nature quelconque dans un espace distancié' *Ann. Inst. Henri Poincaré* **10**, 215–310.

P. Gabriel (1972), 'Unzerlegbare Darstellungen I', *Manuscr. Math.* **6**, 71–103.

A. Greven, P. Pfaffelhuber and A. Winter (2009), 'Convergence in distribution of random metric measure spaces', *Probab. Theory Rel. Fields* **145**, 285–322.

J. Gott, M. Dickinson and A. Melott (1986), 'The sponge-like topology of large-scale structure in the universe', *Astrophys. J.* **306**, 341–357.

A. Hamilton, J. Gott and D. Weinberg (1986), 'The topology of the large-scale structure of the universe', *Astrophys. J.* **309**, 1–12.

J. Hartigan (1975), *Clustering Algorithms*, Wiley Series in Probability and Mathematical Statistics.

A. Hatcher (2002), *Algebraic Topology*, Cambridge University Press.

J. van Hateren and A. van der Schaaf (1998), 'Independent component filters of natural images compared with simple cells in primary visual cortex', *Proc. R. Soc. Lond.* B **265**, 359–366.
J. Huchra, T. Jarrett, M. Skrutskie, R. Cutri, S. Schneider and L. Macri (2005), The 2MASS redshift survey and low galactic latitude large-scale structure. In *Nearby Large-Scale Structures and the Zone of Avoidance* (K. P. Fairall and P. A. Woudt, eds), ASP Conference Series, Vol. 329.
J. Irwin and B. Schoichet (2005), 'ZINC: A free database of commercially available compounds for virtual screening', *J. Chem. Inf. Model.* **45**, 177–182.
H. Karcher (1977), 'Riemannian center of mass and mollifier smoothing', *Comm. Pure Appl. Math.* **30**, 509–541.
W. Kendall (1990), 'Probability, convexity, and harmonic maps with small image I: Uniqueness and fine existence', *Proc. London Math. Soc.* **61**, 371–406.
J. Kogan (2007), *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press.
A. Lee, K. Pedersen and D. Mumford (2003), 'The non-linear statistics of high contrast patches in natural images', *Internat. J. Computer Vision* **54**, 83–103.
D. Lipsky, P. Skraba and M. Vejdemo-Johansson (2011), A spectral sequence for parallelized persistence. arXiv:1112.1245
J. Listing (1848), *Vorstudien zur Topologie*, Vandenhoeck und Ruprecht.
P. Lum, G. Singh, J. Carlsson, A. Lehman, T. Ishkhanov, M. Vejdemo-Johansson, M. Alagappan and G. Carlsson (2013), Extracting insights from the shape of complex data using topology. *Nature Scientific Reports* **3**, # 1236.
S. Mac Lane (1998), *Categories for the Working Mathematician*, second edition, Vol. 5 of *Graduate Texts in Mathematics*, Springer.
A. Maleki, M. Shahram and G. Carlsson (2008), Near optimal coder for image geometries. In *Proc. 15th IEEE International Conference on Image Processing* (ICIP), pp. 1061–1064.
Y. Mileyko, S. Mukherjee and J. Harer (2011), 'Probability measures on the space of persistence diagrams', *Inverse Problems* **27**, 1–22.
J. Milnor (1963), *Morse Theory*, Princeton University Press.
J. Munkres (1975), *Topology: A First Course*, Prentice Hall.
M. Nicolau, A. Levine and G. Carlsson (2011), Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Nat. Acad. Sci.* **108**, 7265–7270.
P. Niyogi, S. Smale and S. Weinberger (2008), 'Finding the homology of submanifolds with high confidence from random samples', *Discrete Comput. Geom.* **39**, 419–441.
C. Park, P. Pranav, P. Chingangram, R. van de Weygaert, B. Jones, G. Vegter, I. Kim, J. Hidding and W. Helwing (2013), 'Betti numbers of Gaussian fields', *J. Korean Astron. Soc.* **46**, 125–131.
J. Perea and G. Carlsson (2014), 'A Klein bottle-based dictionary for texture representation', *Internat. J. Computer Vision* **107**, 75–97.
J. Perea and J. Harer (2014), Sliding windows and persistence: An application of topological methods to signal analysis. arXiv:1307.6188v1
H. Poincaré (1895), 'Analysis situs', *Journal de l'École Polytechnique* (2) **1**, 1–123.

B. Riemann (1851), Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse. Dissertation, Göttingen.

V. Robins (1999), 'Towards computing homology from finite approximations', *Topology Proceedings* **24** (1), 503–532.

G. Segal (1968), 'Classifying spaces and spectral sequences', *Inst. Hautes Études Sci. Publ. Math.* **34**, 105–112.

V. de Silva and G. Carlsson (2004), Topological estimation using witness complexes. In *Proc. First Eurographics Conference on Point-Based Graphics*, pp. 157–166.

G. Singh, F. Mémoli and G. Carlsson (2007), Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Proc. Eurographics Symposium on Point-Based Graphics 2007* (M. Botsch and R. Pajarola, eds).

T. Sousbie (2011), 'The persistent cosmic web and its filamentary structure I: Theory and implementation', *Mon. Not. R. Astron. Soc.* **414**, 350–383.

T. Sousbie, C. Pichon and H. Kawahara (2011), 'The persistent cosmic web and its filamentary structure II: Illustrations', *Mon. Not. R. Astron. Soc.* **414**, 384–403.

D. Spergel, R. Bean, O. Doré, M. Nolta, C. Bennett, J. Dunkley, G. Hinshaw, N. Jarosik, E. Komatsu, L. Page, H. Peiris, L. Verde, M. Halpern, R. Hill, A. Kogut, M. Limon, S. S. Meyer, N. Odegards, G. Tucker, J. Weiland, E. Wollack and E. Wright (2007) 'Three-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Implications for cosmology', *Astrophys. J.* Supplement Series **170**, 377–408.

K. Turner, Y. Mileyko, S. Mukherjee and J. Harer (2014), Fréchet means for distributions of persistence diagrams. arXiv:1206.2790v2

A. Vandermonde (1774), Remarques sur les problèmes de situation. In *Mémoires de l'Académie Royale des Sciences pour 1771*, Paris, pp. 556–574.

R. van de Weygaert, G. Vegter, H. Edelsbrunner, B. Jones, P. Pranav, C. Park, W. Hellwing, B. Eldering, N. Kruithof, E. Bos, J. Hidding, J. Feldbrugge, E. ten Have, M. van Engelen, M. Caroli and M. Teillaud (2011), Alpha, Betti, and the Megaparsec Universe: On the topology of the cosmic web. In *Transactions on Computational Science XIV* (M. L. Gavrilova *et al.*, eds), Vol. 6970 of *Lecture Notes in Computer Science*, Springer, pp. 60–101.

A. Zomorodian (2005), *Topology for Computing*, Vol. 16 of *Cambridge Monographs on Applied and Computational Mathematics*, Cambridge University Press.

A. Zomorodian (2010), The tidy set: A minimal simplicial set for computing homology of clique complexes. In *Proc. 2010 Annual Symposium on Computational Geometry*, ACM, pp. 257–266.

A. Zomorodian and G. Carlsson (2005), 'Computing persistent homology', *Discrete Comput. Geom.* **33**, 247–274.