2

Topological Data Analysis

> I prefer to express myself metaphorically. Let me stress: metaphorically, not symbolically. A symbol contains within itself a definite meaning, certain intellectual formula, while metaphor is an image. An image possessing the same distinguishing features as the world it represents.
>
> *Andrei Tarkovsky*

A central dogma of topological data analysis is that data sets have shape and that describing this shape can help explain the process generating the data. As we have outlined in the preceding chapter, from this perspective clustering techniques extract "zero dimensional" information about connected components of the data set. One of the central goals of topological data analysis is to use the methods of algebraic topology to extract higher dimensional information about the shape of the data set. For example, if we suppose that the data is sampled from a manifold, a candidate goal might be to recover the homology of that manifold. More realistically, we might simply wish to recover qualitative descriptors of the data set that are robust to perturbation and capture higher dimensional information, without necessarily postulating that there is such a clean underlying geometric description. That is, we would like to set up a pipeline

$$\{\text{data}\} \to \left\{ \begin{matrix} \text{simplicial} \\ \text{complexes} \end{matrix} \right\} \to \left\{ \begin{matrix} \textit{algebraic} \\ \text{invariants} \end{matrix} \right\}.$$

To apply algebraic topology to discrete data, two major issues need to be tackled. First, we need a way to transform a discrete set of points into a richer topological space in order to have interesting topological invariants to compute. Second, the *feature scale* of the data must be accounted for; namely, we need to determine the relationship between the size of meaningful geometric features of the data and the distances between the sampled points. This second question is particularly interesting, since a priori the feature scale is often unknown. In this chapter, we explain approaches to these problems, with a primary focus on *persistent homology*

and related constructions. The basic idea is to collect information for all feature scales at once. Persistent homology originated in the work of Frosini [187], and was independently rediscovered by Robins [433] and Edelsbrunner, Letscher, and Zomorodian [154]; in Section 2.11 at the end of the chapter we provide more comprehensive references for the interested reader.

2.1 Simplicial Complexes Associated to Data

A basic and widely applicable model for the kind of data that arises in practice is a *finite metric space*; this is simply a metric space (X, ∂_X) with finitely many points. A natural geometric example of a finite metric space is a collection of points $\{x_0, x_1, \ldots, x_k\} \subset \mathbb{R}^n$ equipped with the induced Euclidean metric $\partial_{\mathbb{R}^n}$. A natural biological example of a finite metric space is a collection of gene expression vectors in $\mathbb{R}^{20000}$, with the distance between v_1 and v_2 computed by the Pearson correlation (recall Example 1.2.6).

Recall from Example 1.3.7 that any metric space (X, ∂_X) has a natural topology where the basic open sets are the balls $B_\epsilon(x) = \{z \in X \mid \partial_X(z, x) < \epsilon\}$ for all $\epsilon > 0$. As a consequence, a first thought might be to simply regard a finite metric space (X, ∂_X) as a topological space directly. Unfortunately, such a space is not very interesting – the topology is trivial, in the sense that it is discrete.

- Every point is both open and closed.
- There are no continuous maps $\gamma\colon [0, 1] \to X$ other than the constant maps. (See Figure 2.1.)
- All homological invariants except π_0 and H_0 (which just count the number of points in X) are trivial.

In order to leverage the tools of algebraic topology to study finite metric spaces, we need a different idea for assigning a topological space to (X, ∂_X). To figure out what to do, it is useful to think about the toy model in which the sampled data X was generated by drawing from some probability distribution on a nice geometric

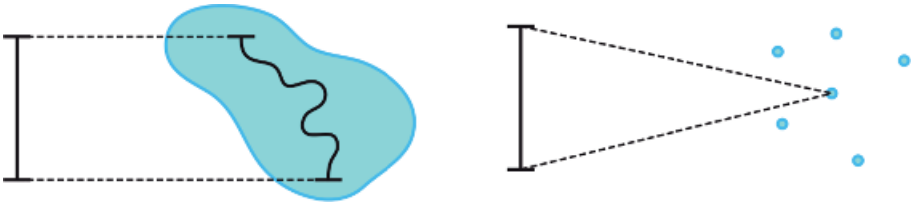


Figure 2.1 The only continuous maps from $[0, 1]$ to a discrete topological space are constant at a point.

object embedded in $\mathbb{R}^n$ (e.g., a compact smooth manifold). In this case, it is clear that we need to somehow "fill in the gaps" between the samples. If we have a rough sense of the average distance between points that are supposed to be connected, there is an evident construction: just take the union of balls around the points.

Definition 2.1.1 (Union of balls)**.** Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon \geq 0$. The *union of balls* is the union

$$\bigcup_{x \in X} B_\epsilon(x) \subset \mathbb{R}^n.$$

However, from a practical perspective, the union of balls is not ideal; it is not evidently algorithmically tractable, and it requires that (X, ∂_X) arise as a subspace of $\mathbb{R}^n$. To fix the first problem, we would like to produce an abstract simplicial complex that encodes the information of the union of balls. We can adapt this construction to the discrete setting by regarding the ϵ-balls around a finite set X as a cover. That is, the idea is to associate a k-simplex to a set of k points whose ϵ-neighborhoods intersect.

Definition 2.1.2 (Čech complex)**.** Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. The *Čech complex* $C_\epsilon(X, \partial_X)$ is the abstract simplicial complex with

1. vertices the points of X, and
2. a k-simplex $[v_0, v_1, \ldots, v_k]$ when a set of points $\{v_0, v_1, \ldots, v_k\} \subset X$ satisfies

$$\bigcap_i B_\epsilon(v_i) \neq \emptyset.$$

In fact, the Čech complex (Figure 2.2) is a special case of a standard construction from algebraic topology that associates a simplicial complex to a *cover* of a space. Recall from Definition 1.3.15 that an open cover $\{U_i\}$ of a space X is a collection of open sets such that $\cup_i U_i = X$. Given a cover $\{U_i\}$ of X, we define the *nerve* of the cover as follows.

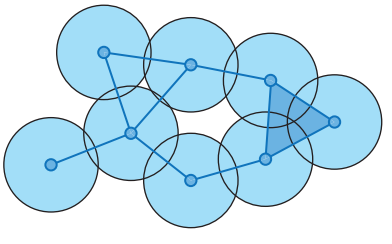


Figure 2.2 The Čech complex is a combinatorial approximation to the union of balls.

Definition 2.1.3. The *nerve* $N(\{U_i\})$ of a cover $\{U_i\}$ of X is the simplicial complex with

1. vertices corresponding to the sets $\{U_i\}$, and
2. a k-simplex $[j_0, j_1, \ldots, j_k]$ when the intersection

$$U_{j_0} \cap U_{j_1} \cap U_{j_2} \cap \ldots \cap U_{j_k} \neq \emptyset.$$

The interest of this construction is the following classical result about the relationship of the geometric realization (recall Definition 1.8.8 and Lemma 1.8.20) of this nerve to X; see e.g., [307, §15.4.3] for further discussion and a proof.

Theorem 2.1.4. *Let X be a topological space. Let $\{U_i\}$ be an open cover of X such that all non-empty finite intersections*

$$U_{j_1} \cap U_{j_2} \cap \ldots \cap U_{j_k}$$

are contractible (homotopy equivalent to a point). Then the geometric realization $|N(\{U_i\})|$ is homotopy equivalent to X.

As a corollary, we obtain the following result comparing the geometric realization of the Čech complex to the geometric Čech nerve.

Proposition 2.1.5. *Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. There exists a homeomorphism*

$$\bigcup_{x \in X} B_\epsilon(x) \cong |C_\epsilon(X, \partial_X)|$$

between the union of balls and the geometric realization of the Čech complex.

The Čech complex provides a procedure for assigning a simplicial complex to a finite metric space embedded in $\mathbb{R}^n$. However, in order to construct the Čech complex we need to be able to decide whether the intersection of ϵ-balls is non-empty. This is a non-trivial enterprise in high dimensions. Moreover, we do not wish to assume that the data points are embedded in Euclidean space at all!

To see how to proceed, it is helpful to recall our discussion of path components and single-linkage clustering for a metric space from Section 1.3. Here, for a finite metric space (X, ∂_X) and fixed $\epsilon > 0$, we defined a graph $G = (V, E)$ with

1. vertices the points of X, and
2. edges (x_i, x_j) for each pair of points x_i and x_j such that $\partial_X(x_i, x_j) \leq \epsilon$.

Recalling that a graph is a one dimensional simplicial complex, we use a mild elaboration of this construction to define a simplicial complex associated to an

arbitrary finite metric space (X, ∂_X). The Vietoris-Rips complex is the maximal simplicial complex determined by the vertices and 1-simplices specified by the graph G.

Definition 2.1.6 (Vietoris-Rips complex)**.** Let (X, ∂_X) be a finite metric space and fix $\epsilon > 0$. The *Vietoris-Rips complex* $\mathrm{VR}_\epsilon(X, \partial_X)$ is the abstract simplicial complex with

1. vertices the points of X, and
2. a k-simplex $[v_0, v_1, \ldots, v_k]$ when

$$\partial_X(v_i, v_j) \leq 2\epsilon \qquad \text{for all} \qquad 0 \leq i, j \leq k.$$

For a point cloud in $\mathbb{R}^n$, the Vietoris-Rips complex and the Čech complex can be different; for instance, notice that there is a difference between the Čech complex in Figure 2.2 and the Vietoris-Rips complex in Figure 2.3, which are generated by the same underlying metric space. The next example highlights the kind of phenomenon that leads to such differences.

Example 2.1.7. Consider the finite metric space $X = \{(0,0), (1,0), (\frac{1}{2}, \frac{\sqrt{3}}{2})\} \subset \mathbb{R}^2$. These points are the vertices of an equilateral triangle with side length 1. Choose an ϵ in the open interval $(\frac{1}{2}, \frac{\sqrt{3}}{3})$, i.e., $\frac{1}{2} < \epsilon < \frac{\sqrt{3}}{3}$. (For concreteness, $\frac{\sqrt{3}}{3} \approx 0.577$.)

1. The Vietoris-Rips complex $\mathrm{VR}_\epsilon(X, \partial_X)$ has three vertices (one for each point of X), three 1-simplices (connecting the points), and therefore has a single 2-simplex filling in the triangle.
2. In contrast, the Čech complex $C_\epsilon(X, \partial_X)$ has three vertices (one for each point of X) and three 1-simplices (connecting the points), but does not have the 2-simplex spanned by all the points since there is no point in the intersection of the balls of radius ϵ.

(See Figure 2.4 for a corresponding picture.)

The use of the Čech complex is justified by the Nerve Lemma (Theorem 2.1.4); there is no analogous result for the Vietoris-Rips complex. However, despite the

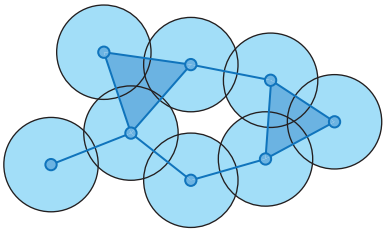


Figure 2.3 The Vietoris-Rips complex is completely determined by its 1-skeleton.

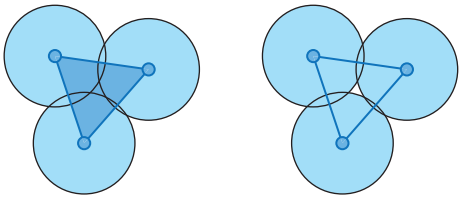


Figure 2.4 The Vietoris-Rips complex (on the left) is completely determined by its 1-skeleton, whereas the Čech complex (on the right) can potentially omit higher simplices.

fact that they are sometimes different, there is a close relationship between the Vietoris-Rips and Čech complexes.

Lemma 2.1.8. *Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. There are natural simplicial inclusions*

$$C_\epsilon(X, \partial_X) \subseteq \mathrm{VR}_\epsilon(X, \partial_X) \subseteq C_{2\epsilon}(X, \partial_X).$$

An essential property of the constructions of the Čech complex and the Vietoris-Rips complex is that they are functorial. To be precise, these constructions are functorial in both X and ϵ. (In the following discussion, we focus on the Vietoris-Rips complex; the properties of the Čech complex are analogous.) For $\epsilon < \epsilon'$ and any metric space (X, ∂_X), there is an induced simplicial map

$$\mathrm{VR}_\epsilon(X, \partial_X) \to \mathrm{VR}_{\epsilon'}(X, \partial_X),$$

since increasing the scale parameter adds more simplices.

Next, recall that a map $f\colon X \to Y$ between metric spaces (X, ∂_X) and (Y, ∂_Y) is Lipschitz continuous with constant k if $\partial_Y(f(x_1), f(x_2)) \leq k\partial_X(x_1, x_2)$. Given a Lipschitz map $f\colon X \to Y$ with Lipschitz constant k, there is an induced simplicial map

$$f\colon \mathrm{VR}_\epsilon(X, \partial_X) \to \mathrm{VR}_{k\epsilon}(Y, \partial_Y)$$

for any ϵ. Summarizing, we have the following theorem.

Theorem 2.1.9. *The construction $\mathrm{VR}_\epsilon(-)$ specifies a functor from the category of finite metric spaces and Lipschitz maps with constant 1 to Simp. The construction $\mathrm{VR}_{(-)}(X, \partial_X)$ specifies a functor from $\mathbb{R}$ to Simp.*

This means that when we vary the scale ϵ, there is a map between the associated complexes for a given data set (X, ∂_X). And if we change a data set (X, ∂_X) to produce a new data set (Y, ∂_Y) related via a Lipschitz map, there is a map connecting the associated complexes. For example, if we add some data points, so that

$Y = X \cup A$ and the metric on Y restricts to ∂_X on $X \subset Y$, then there is a map $\mathrm{VR}_\epsilon(X, \partial_X) \to \mathrm{VR}_\epsilon(Y, \partial_Y)$.

We now turn to the question of when these constructions can recover information about the underlying geometric structure of the process that generated the data.

Question 2.1.10. Let (X, ∂_X) be a finite metric space consisting of samples from a topological space A. When is $|\mathrm{VR}_\epsilon(X, \partial_X)|$ or $|C_\epsilon(X, \partial_X)|$ homotopy equivalent to A?

2.2 The Niyogi-Smale-Weinberger Theorem

In order to make sense of this question, we need to develop a precise model for sampling from a topological space A. We will introduce a definition of *geometric sampling* and study Question 2.1.10 in Chapter 3. However, to illustrate some of the geometric principles that motivate TDA, in this section we will explain an answer to the question in a very restricted context. Specifically, we describe a minimal sanity check: we explain the Niyogi-Smale-Weinberger result that given a finite metric space (X, ∂_X) consisting of sufficiently many points sampled "uniformly" from a compact Riemannian manifold $M \subset \mathbb{R}^n$, with high probability there is an isomorphism

$$H_*\left(\bigcup_{x \in M} B_\epsilon(x)\right) \cong H_*(M)$$

for some suitable choice of ϵ.

Going forward, we assume that we are given a compact manifold $M \subset \mathbb{R}^n$ that has a Riemannian structure. Recall from Section 1.11 that roughly speaking, this means that at each point of the manifold we can equip the tangent space with an inner product, and these inner products vary smoothly as we move on the manifold. As a consequence, M has a metric and there is a natural notion of volume of subspaces of M. In particular, there is a natural notion of what it means to sample from such a manifold, as the manifold is equipped with a probability measure called the *volume measure*.

We want to estimate how many sampled points are necessary to estimate the homology with high probability. When sampling from the volume measure on a Riemannian manifold, it is straightforward to figure out how many points to sample so that with probability $> \kappa$ (for any fixed κ) we get an ϵ-net. Therefore, we can reduce the problem to trying to understand when a finite ϵ-net $X \subset M$ has the property that for some ϵ',

$$H_*(|C_{\epsilon'}(X, \partial_X)|) \cong H_*(M).$$

When such an isomorphism occurs depends on the size of the smallest geometric features of the manifold. That is, we need to figure out how close together points need to be in order for little balls around them to capture the structure of the manifold. For a manifold M embedded in $\mathbb{R}^n$, there are two distinct but interacting factors that control how small ϵ has to be in order for the geometric nerve to have the correct topology. We need to worry about the intrinsic curvature of the manifold, and how "twisted" the embedding into $\mathbb{R}^n$ is. See Figure 2.5 for some examples of possible embeddings of familiar geometric objects into Euclidean space.

Consider the case of S^1 embedded in $\mathbb{R}^2$. In order for the Čech nerve of an ϵ-net to have the right homotopy type, we must be able to choose an ϵ' such that

1. ϵ' is large enough to cause points of the net around the circle to be connected by 1-simplices, but
2. ϵ' is small enough so that points across the circle are not connected by "cross-cutting" 1-simplices.

The relationship between the scale ϵ and ranges of suitable values for ϵ' is controlled in part by the underlying topology of the circle – sufficiently large values for ϵ' will always result in 1-simplices that connect points across the circle. On the other hand, for very twisty embeddings, we will need to choose an ϵ' that is smaller than the size of the twists.

We think of these considerations as packaged into a quantity we refer to as the *feature scale* of the manifold. A very nice way to encode the feature scale of the manifold is to use an invariant called the *condition number*. (This is sometimes also referred to as the *reach* or *feature size*.) Any manifold embedded in $\mathbb{R}^n$ can

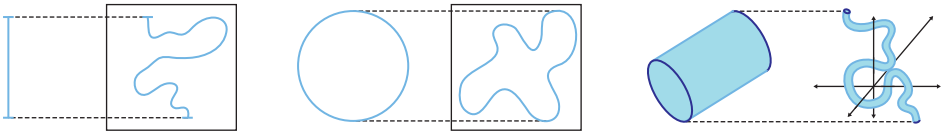


Figure 2.5 The difficulty in reconstructing a geometric object can come from both the intrinsic curvature and the twistiness of the embedding in $\mathbb{R}^n$.

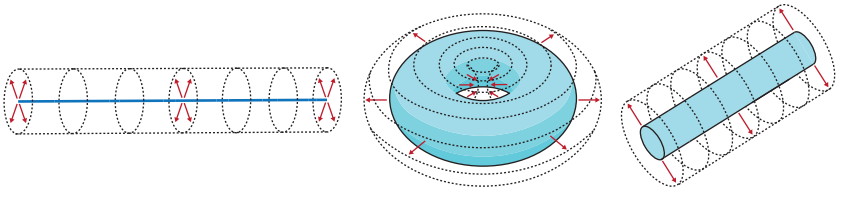


Figure 2.6 A tubular neighborhood is formed by expanding a manifold along the normal directions (perpendicular to its surface).

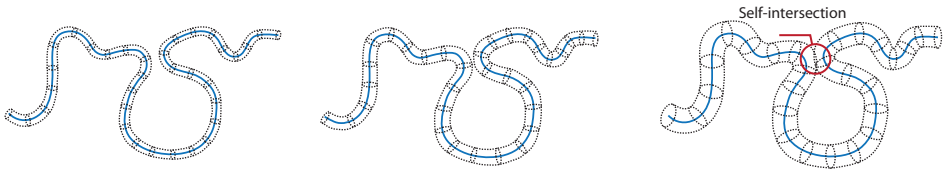


Figure 2.7 As the tubular neighborhood of a curve expands, eventually it self-intersects at the narrowest "pinch."

be thickened out to a *tubular neighborhood* of radius r; this is what one gets by extending out along the normal at any point. (See Figure 2.6 for some examples.)

The condition number is the minimum radius at which a tubular neighborhood of a manifold self-intersects; clearly, this can happen either because the manifold itself has small features (e.g., small holes) or because the embedding twists the manifold around on itself. (See Figure 2.7 for an example.)

The following theorem, due to Niyogi, Smale, and Weinberger [384], now provides a concrete result guaranteeing correct estimation of the homology.

Theorem 2.2.1. *Let M be a compact submanifold of $\mathbb{R}^n$ with condition number τ and let $\{x_1, \ldots, x_k\}$ be a set of points drawn from M according to the volume measure. Fix $0 < \epsilon < \frac{\tau}{2}$. Then if*

$$k > \beta_1 \left(\log(\beta_2) + \log\left(\frac{1}{\delta}\right)\right),$$

there is a homotopy equivalence

$$\bigcup_{z \in \{x_1, \ldots, x_k\}} B_\epsilon(z) \simeq M$$

between the union of balls and M (and in particular the homology groups coincide) with probability $> 1 - \delta$.

Here

$$\beta_1 = \frac{\text{vol}(M)}{\cos^n(\theta_1)\text{vol}(B^n_{\frac{\epsilon}{4}})}$$

and

$$\beta_2 = \frac{\text{vol}(M)}{\cos^n(\theta_2)\text{vol}(B^n_{\frac{\epsilon}{8}})},$$

where $\theta_1 = \arcsin\left(\frac{\epsilon}{8\tau}\right)$, $\theta_2 = \arcsin\left(\frac{\epsilon}{16\tau}\right)$, and $\text{vol}(B^n_r)$ denotes the volume of the n-dimensional ball of radius r.

Remark 2.2.2. Using different techniques, one can also prove an analogous result directly for the Vietoris-Rips complex [3, 315].

To get a sense for what this means, it is helpful to do an explicit example.

Example 2.2.3. The condition number of a sphere is simply its radius. So for example, for the unit circle $S^1 \subset \mathbb{R}^2$, the condition number τ is 1. Choosing $\delta = 0.01$ and $\epsilon = \frac{1}{4}$, we compute that

$$\cos^2\left(\arcsin\left(\frac{1}{32}\right)\right) \approx 1 \qquad \text{and} \qquad \cos^2\left(\arcsin\left(\frac{1}{64}\right)\right) \approx 1$$

and so

$$\beta_1 = \frac{2\pi}{\pi(\frac{1}{16})^2} = 512$$

and

$$\beta_2 = \frac{2\pi}{\pi(\frac{1}{32})^2} = 2048,$$

which means that we need at least

$$512(7.6 + 4.6) \approx 6260$$

samples.

Example 2.2.4. The condition number of a torus is the minimum of r_1 and $\frac{r_2 - r_1}{2}$, where r_1 and r_2 are the radii of the inner and outer bounding circles. We can repeat a similar computation as above, using the fact that the volume (surface area) of the torus is $(r_2^2 - r_1^2)\pi^2$; once again, we end up with a number of points in the thousands for reasonable values of δ and ϵ.

These examples frame the application of Theorem 2.2.1 in high relief. On the one hand, this result is of critical theoretical importance, and it provides a vital consistency check for combinatorial approaches to estimating the homology of manifolds from finite data. On the other hand, the explicit bounds are useless – in practice it is difficult or impossible to estimate the condition number (although see [1]) and moreover a result of 3000 points to estimate the homology of a standard circle in $\mathbb{R}^2$ is clearly much too large. (To be sure, a direct argument can be used to obtain a much tighter bound.) In applications, we will be much more concerned about the stability of the result in the face of sampling variation and noise.

Remark 2.2.5. Theorem 2.2.1 is a statement about approximating the homotopy type of a manifold via finite sampling. One might wonder how many samples are required to estimate the homeomorphism type of M. Unfortunately, even in this very restricted setting, the problem turns out to be hopeless. Assume that M is

embedded in $\mathbb{R}^n$ and the condition number is a fixed constant. Then when the dimension of M is larger than 2, the number of samples required to identify the homeomorphism type is exponential in $\text{diam}(M)^n$; see [533, §2.2] for a discussion. These concerns are relevant when studying single cell data; see Chapter 7.

2.3 Persistent Homology

The Niyogi-Smale-Weinberger theorem (Theorem 2.2.1) shows that in principle it is possible to accurately recover topological invariants of geometric objects from discrete samples. We interpret the theorem to suggest that it is reasonable to hope that in very general settings, when the distance between the samples is smaller than some *feature scale*, we can recover topological invariants of the underlying geometric object.

However, there is a key problem: the feature scale of the underlying object is usually unknowable a priori. That is, given (X, ∂_X) from M, how can we choose ϵ so that the topological invariants of $|\text{VR}_\epsilon(X, \partial_X)|$ recover information about the topological invariants of M? Moreover, choosing a single ϵ is problematic – for one thing, there might be distinct feature scales at which we can recover meaningful information, for instance if the data has regions of varying size. Another issue is that the topological invariants of $|\text{VR}_\epsilon(X, \partial_X)|$ are very unstable; small amounts of noise or sampling variation can cause large changes in the Vietoris-Rips complex and its homology. That is, at any given scale some features might not be stable with respect to noise or change of scale.

The guiding viewpoint that underlies topological data analysis is that we should simultaneously look at multiple feature scales; stable homological features that exist for a range of values of ϵ are likely to reflect the underlying signal, and this approach allows us to capture multiscale information. A naive approach to implementing this idea would simply be to vary ϵ and compute a collection of associated invariants.

1. Choose a topological invariant, e.g., the homology group $H_2(-; \mathbb{F}_p)$.
2. Select a range $[\epsilon_{\min}, \epsilon_{\max}]$, $\epsilon_{\min} < \epsilon_{\max}$. This interval reflects the smallest and largest feature scales that we will consider; a maximal choice would be to set $\epsilon_{\min} = 0$ and $\epsilon_{\max} = \text{diam}(X)$.
3. Choose values $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_m\} \subset [\epsilon_{\min}, \epsilon_{\max}]$. An easy way to do this is simply to consider the equally spaced values

$$\epsilon_i = \epsilon_{\min} + i\left(\frac{\epsilon_{\max} - \epsilon_{\min}}{m}\right),$$

but it might make sense to bunch the values around regions of interest, if we have domain knowledge about interesting feature scales.

4. Compute the collection of vector spaces

$$\{H_2(|\mathrm{VR}_{\epsilon_1}(X, \partial_X)|), H_2(|\mathrm{VR}_{\epsilon_2}(X, \partial_X)|), \ldots, H_2(|\mathrm{VR}_{\epsilon_m}(X, \partial_X)|)\}.$$

5. Compare these abelian groups; for example, make a graph of the ranks of the free parts. If these are all non-zero and all the same, it suggests that there are stable topological features of M at the feature scales in the interval $[\epsilon_{\min}, \epsilon_{\max}]$. If there is a subinterval $[a, b] \subset [\epsilon_{\min}, \epsilon_{\max}]$ on which the ranks are the same, we might conclude that there are stable topological features at those ranges of scales. (Of course, there is no guarantee that we are not seeing different features at the different scales; this procedure does not really help us match topological features across scales.)

For an example of how this might work, consider the situation depicted in Figure 2.8. When ϵ is smaller than the distance between points, the Vietoris-Rips

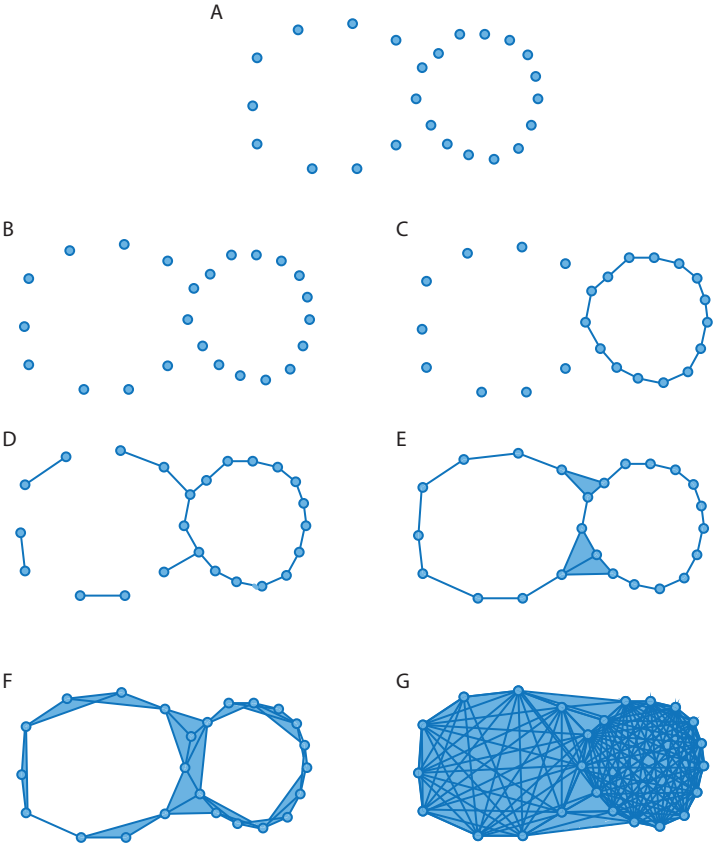


Figure 2.8 As ϵ increases, more and more simplices appear in the Vietoris-Rips complex.

complex only has 0-simplices. As ϵ increases, we first see 1-simplices appear that eventually connect the right hand circle. Then the left hand circle also appears. Finally, the circles are "filled in" by simplices crossing the circles when ϵ is large enough.

A first question is how to systematically choose the values $\{\epsilon_i\}$. Ideally, we will track places where $\mathrm{VR}_\epsilon(X, \partial_X)$ changes. Since X is finite, there are only finitely many values $\{\epsilon_i\}$ at which the simplicial complex $\mathrm{VR}_\epsilon(X, \partial_X)$ changes. We can see this because for $\epsilon > \mathrm{diam}(X)$ the Vietoris-Rips complex has all $2^{|X|}$ simplices and as ϵ increases simplices are added but never removed.

Lemma 2.3.1. *Let (X, ∂_X) be a finite metric space. Then there exist at most finitely many values $\{\epsilon_i\}$ where $\mathrm{VR}_{\epsilon_i}(X, \partial_X)$ changes, i.e., such that for all sufficiently small δ,*

$$\begin{cases} \mathrm{VR}_\epsilon(X, \partial_X) = Z & \epsilon \in [\epsilon_i - \delta, \epsilon_i) \\ \mathrm{VR}_\epsilon(X, \partial_X) = Z' & \epsilon \in [\epsilon_i, \epsilon_i + \delta] \end{cases}$$

and $Z \neq Z'$.

Therefore, we should choose $\{\epsilon_i\}$ to lie at these "inflection points" (and there is an upper bound on how many values we need to consider).

However, the most critical step is the last one; we need to find a systematic way to compare the various $\{H_2(\mathrm{VR}_{\epsilon_i}(X, \partial_X))\}$. The key insight of *persistence* is that since $\mathrm{VR}_{(-)}(X, \partial_X)$ is functorial in ϵ, for $\epsilon < \epsilon'$ we have a map of simplicial complexes

$$\mathrm{VR}_\epsilon(X, \partial_X) \to \mathrm{VR}_{\epsilon'}(X, \partial_X),$$

and for a collection $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_m$ we obtain a sequence of simplicial maps

$$\mathrm{VR}_{\epsilon_1}(X, \partial_X) \to \mathrm{VR}_{\epsilon_2}(X, \partial_X) \to \ldots \to \mathrm{VR}_{\epsilon_m}(X, \partial_X).$$

Since H_k is also a functor, applying H_k we obtain induced maps of abelian groups or vector spaces

$$H_k(\mathrm{VR}_\epsilon(X, \partial_X)) \to H_k(\mathrm{VR}_{\epsilon'}(X, \partial_X))$$

and

$$H_k(\mathrm{VR}_{\epsilon_1}(X, \partial_X)) \to H_k(\mathrm{VR}_{\epsilon_2}(X, \partial_X)) \to \ldots \to H_k(\mathrm{VR}_{\epsilon_m}(X, \partial_X)).$$

More concisely, we can package this data as follows.

Definition 2.3.2. Given a fixed finite metric space (X, ∂_X), the Vietoris-Rips complex induces a functor

$$\mathrm{VR}_{(-)}(X, \partial_X) \colon \mathbb{R} \to \mathrm{Simp}$$

from $\mathbb{R}$ (regarded as the category associated to a partially ordered set) to the category of simplicial complexes. Composition with the kth homology group functor gives rise to a functor

$$H_k(\mathrm{VR}_{(-)}(X, \partial_X)) \colon \mathbb{R} \to \mathrm{Ab}.$$

It is useful to organize the resulting functors themselves into categories.

Definition 2.3.3. Let $\mathcal{C}$ be a category. The category of *filtered systems of* $\mathcal{C}$ is the category of functors $F \colon \mathbb{R} \to \mathcal{C}$ with morphisms given by natural transformations.

Clearly, any filtered system of simplicial complexes produces a filtered system of abelian groups or vector spaces. There are a variety of sources of filtered complexes that are relevant in topological data analysis, but for expositional clarity, we will focus on the Vietoris-Rips complex for the remainder of this discussion.

Example 2.3.4.

1. The Vietoris-Rips complex and Čech complex produce natural examples of filtered systems of simplicial complexes from the data of a finite metric space where we allow the scale ϵ to vary.
2. Motivated by the perspective of Morse theory, we assume the underlying data is a simplicial complex X along with a function $h \colon X \to \mathbb{R}$. There is now an induced filtered system of simplicial complexes induced by the inverse images $\{h^{-1}((-\infty, -])\}$. That is, for $b > a$, it is clear that $h^{-1}((-\infty, a])$ is a subcomplex of $h^{-1}((-\infty, b])$.

Remark 2.3.5. Note that the "Morse theoretic" perspective can be regarded as a generalization of the finite metric space approach, as follows. Given a compact subset $K \subseteq \mathbb{R}^n$, define the distance function

$$\partial_K(z) = \inf_{k \in K} \partial_{\mathbb{R}^n}(k, z).$$

Then for a finite set of points $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^n$, the filtered system of Čech complexes associated to the level sets of ∂_X is isomorphic to the filtered simplicial complex $\{C_*(X, \partial_X)\}$.

The functor $H_k(\mathrm{VR}_{(-)}(X, \partial_X))$ provides a means of addressing our problem about comparisons between the homology of the complexes as ϵ varies:

1. an element $\gamma \in H_k(\mathrm{VR}_{\epsilon_i}(X, \partial_X))$ is a k-dimensional feature at scale ϵ_i, and
2. we can determine the significance and stability of γ by finding the maximum $j > i$ such that the image of γ under the group homomorphism

$$\theta_{ij} \colon H_k(\mathrm{VR}_{\epsilon_i}(X, \partial_X)) \to H_k(\mathrm{VR}_{\epsilon_j}(X, \partial_X))$$

is non-zero.

Roughly speaking, an element $\gamma \in H_k(\mathrm{VR}_{\epsilon_i}(X, \partial_X))$ represents a k-dimensional hole in the geometric realization of the Vietoris-Rips complex at ϵ_i. If γ does not exist for $\epsilon' < \epsilon_i$, we think of this feature as being "born" at ϵ_i. When $\theta_{ij}(\gamma) = 0$, it means that the hole has been filled in by a collection of simplices with boundary γ. This suggests that it makes sense to try to figure out the "lifespan" of a particular element in homology, i.e., when it first appears and when it vanishes. More precisely, for a filtered simplicial complex $X_\bullet$, an element $\gamma \in H_k(X_i; \mathbb{F})$ is

1. *born* at i if it is not in the image of $H_k(X_{i-q}; \mathbb{F}) \to H_k(X_i; \mathbb{F})$ for any $q > 0$, and
2. *dies* at $\ell > i$ if it becomes zero in $H_k(X_\ell; \mathbb{F})$ or its image in $H_k(X_\ell; \mathbb{F})$ coincides with the image of another class that was born earlier.

Thus, we can think of the information contained in the filtered system of vector spaces as a series of elements with intervals representing their lifetime. Precisely, the persistent homology of a finite metric space can be described via a "barcode," a collection of intervals. Each interval represents the lifespan of a homological feature. (See Figure 2.9 for a simple representative example.)

Definition 2.3.6. A barcode is a multiset of non-empty intervals of the form either $[x, y) \subset \mathbb{R}$ or $[x, \infty)$. (A multiset is a generalization of a set where repeated elements are allowed, e.g., {1,1,2}.)

To be precise about the connection between persistent homology and barcodes, we require some finiteness hypotheses that always hold in practice, since we only have finitely many data points. We fix a field $\mathbb{F}$ for the remainder of this section.

Definition 2.3.7. A filtered simplicial complex is *tame* if the homology groups $H_i(-; \mathbb{F})$ are always of finite rank and change at only a finite number of indices.

By Lemma 2.3.1, the filtered complexes produced by applying the Vietoris-Rips complex construction to a finite metric space are always tame.

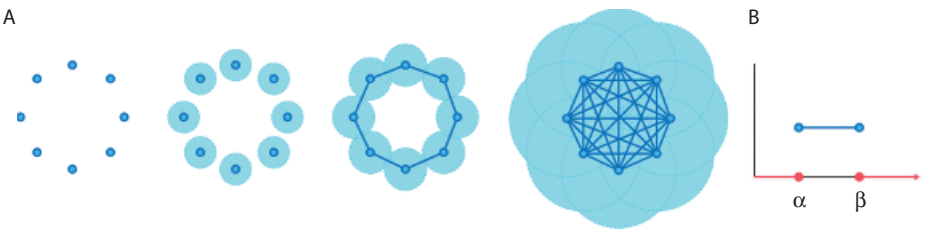


Figure 2.9 In (A), we have an idealized Vietoris-Rips filtration: when $\epsilon = \alpha$, the circle appears, and when $\epsilon = \beta$, the circle is filled in. In (B), the barcode has a single bar (representing a $\mathbb{Z}$ in homology) that appears at α and vanishes at β; this is the homology of the circle, for as long as it lasts.

Lemma 2.3.8. *Let* $X\colon \mathbb{R} \to \mathrm{Simp}$ *be a tame filtered simplicial complex. The filtered vector space produced as* $H_i(X(-);\mathbb{F})$ *has the property that*

1. *each vector space* $H_i(X(\epsilon);\mathbb{F})$ *is of finite rank and*
2. *there exists* N *such that* $H_i(X(\epsilon_1);\mathbb{F}) \to H_i(X(\epsilon_2);\mathbb{F})$ *is an isomorphism for* $\epsilon_2 > \epsilon_1 > N$.

We say such a filtered vector space is of finite type.

Remark 2.3.9. A filtered vector space of finite type can be regarded as indexed on $\mathbb{Z}$, where the integral indices correspond to values in $\mathbb{R}$ where the homology changes.

The key classification result of Zomorodian and Carlsson [551] is then the following.

Theorem 2.3.10. *Let* $\mathbb{F}$ *be a field. There is a bijection between the set of finite barcodes and the set of isomorphism classes of filtered* $\mathbb{F}$*-vector spaces of finite type.*

The basic idea of this classification is quite simple; we define *interval modules*, which are filtered systems I_{ab} of $\mathbb{F}$-vector spaces $\{V_i\}$ where for $i \in [a,b]$, $V_i = \mathbb{F}$, and all the maps $\mathbb{F} \to \mathbb{F}$ are the identity (and the others are necessarily zero). Then any filtered system of $\mathbb{F}$-vector spaces is a direct sum of interval modules; the interval modules correspond to the bars in the barcode representing the lifetime of particular elements in homology.

Theorem 2.3.10 tells us that all of the information in the filtered system of vector spaces can be encoded as barcodes. It is often useful to think of a barcode as a collection of points in $\mathbb{R}^2$, specified by the endpoints of the intervals. Such a set is referred to as a *persistence diagram*, and often it is regarded as containing the entire diagonal (consisting of size zero bars).

In conclusion, we have the "persistent homology pipeline"

$$\left\{\begin{matrix}\text{finite}\\ \text{metric}\\ \text{spaces}\end{matrix}\right\} \to \left\{\begin{matrix}\text{filtered}\\ \text{simplicial}\\ \text{complexes}\end{matrix}\right\} \to \left\{\begin{matrix}\text{barcodes / persistence}\\ \text{diagrams}\end{matrix}\right\}.$$

We now turn to some examples of the use of barcodes to describe shape. When $k = 0$, the persistent homology is describing a standard hierarchical clustering construction.

Example 2.3.11. Recall from Theorem 1.10.10 that for a simplicial complex X, $H_0(X)$ is computing the free abelian group on the components. In the case of $\mathrm{VR}_\epsilon(X, \partial_X)$ for a

finite metric space (X, ∂_X), $H_0(\mathrm{VR}_\epsilon(X, \partial_X))$ computes the single-linkage clustering at scale ϵ of (X, ∂_X).

When considering the persistent homology, observe that each cluster at time $p+i$ can be thought of as resulting from the merger of clusters at i. This is clearly closely related to the information encoded in the hierarchical clustering dendrogram associated to single-linkage clustering. (See Figure 2.10 for comparison of the barcode and dendrogram for a synthetic data set.)

In Figure 2.11, we see an idealized situation involving sampling from an object in $\mathbb{R}^2$. In practice, however, the barcodes are often not so easy to interpret. Even for geometrically simple situations, complications can arise. In Figure 2.12, we illustrate how the barcode can change due to perturbation of the data by considering a sequence of nested circles.

In Figure 2.13, persistent homology of genomic sequence data generated by coalescent simulation is shown. As explained in Section 5.7, this is a way of modeling evolutionary phenomena. Typically, one fits phylogenetic trees to the finite metric space of sequences; here, we compute the persistent homology instead. Computing the first persistent homology group detects when "non-tree-like" events are occurring, i.e., when there is genetic recombination. Another example of this kind of application of persistent homology in studying recombination rates in the evolution of bacteria is discussed in Section 5.6.3; see Figure 2.14. In both of these applications, increased recombination can be detected by a large number of bars in the PH_1 barcode.

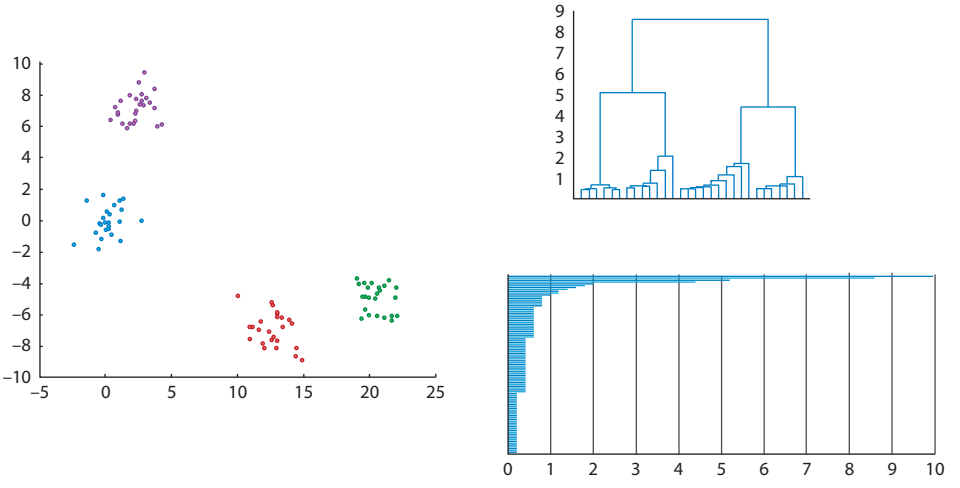


Figure 2.10 For the data set on the left, both the dendrogram and the zeroth persistent homology barcode capture how clusters merge as ϵ increases.

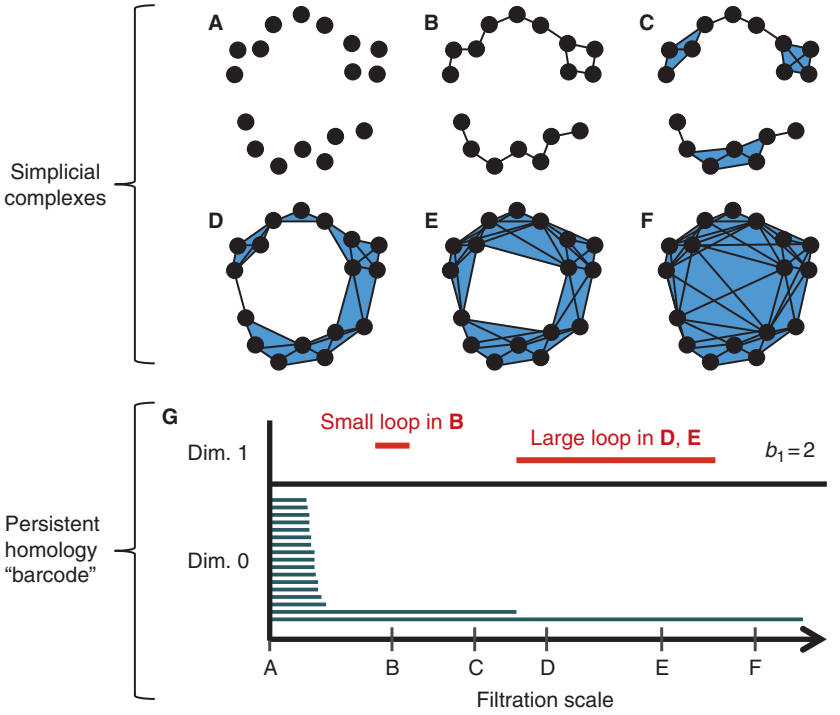


Figure 2.11 The points in panel A form a circle, with a horizontal gap separating upper and lower points. Panels A-F show the Vietoris-Rips filtration on these points as ϵ increases. Panel G shows the barcode. PH_0 (dimension 0) shows clustering of the data at different scales; each horizontal bar in the barcode is a cluster. In panel A (filtration scale 0), no points are connected; each is its own cluster (represented as 17 horizontal bars). As the scale increases, points in the simplicial complex connect, represented in the barcode as termination of a bar. There are two distinct clusters through panels B and C and one cluster in panels D, E, and F. PH_1 (dimension 1) shows loops in the data at different scales. Each bar in this part of the barcode identifies a different loop. There are two loops in this data: a short-lived loop in the top-right of the simplicial complex at scale B, and a long-lived loop appearing in panel D and persisting through panel E – this loop is represented as the long bar in the dimension 1 barcode. Robust features of the data set are captured in the barcode: the data clusters into two groups (two dimension 0 bars through scale C), and forms a loop (one long dimension 1 bar). The persistent first Betti number (b_1) is the total number of dimension 1 bars; here it is equal to 2.

In Section 8.3, we discuss an application of persistent homology to study the physical structure of DNA. Modeling DNA as a sequence of repeated units that have prescribed interaction points, persistent homology can be used to extract information about loops in the strands from a similarity matrix encoding the contact of sites with other sites. (See Figure 2.15.)

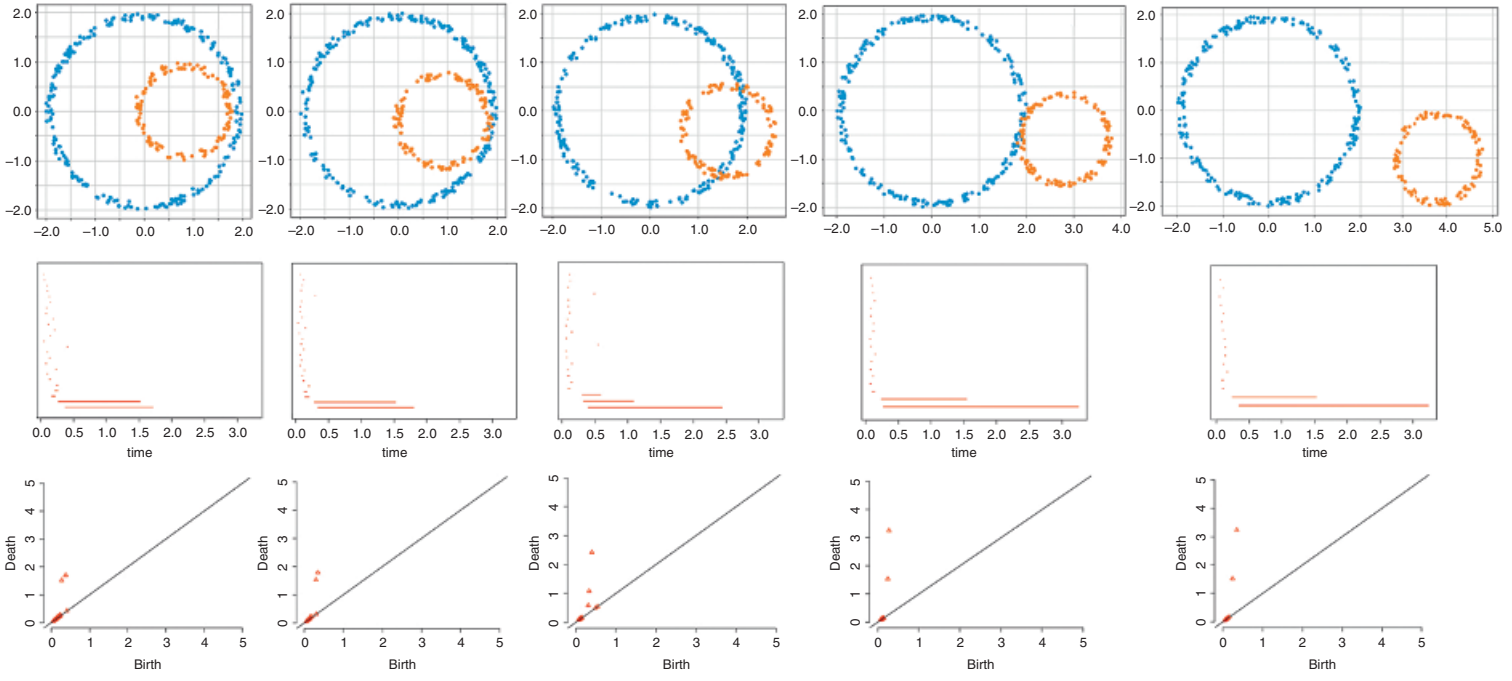


Figure 2.12 With two disjoint circles, we expect to see a barcode with two long bars, one significantly longer than the other to reflect the difference in radius. But when the circles are nested, the bars are nearly the same length as the inner circle interferes with the outer circle. Moreover, little loops connecting the two circles generate a lot of short bars. When the circles intersect non-trivially, we see an extra bar representing the loop formed by the intersection. And finally when the circles are disjoint and separated, we see the expected two bars, one longer and one shorter, corresponding to each circle.

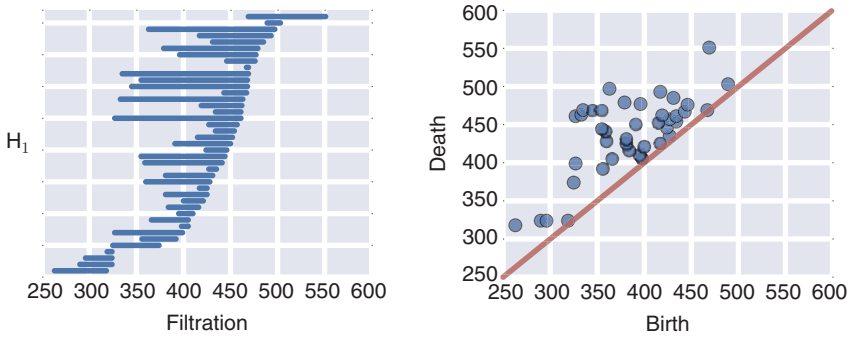
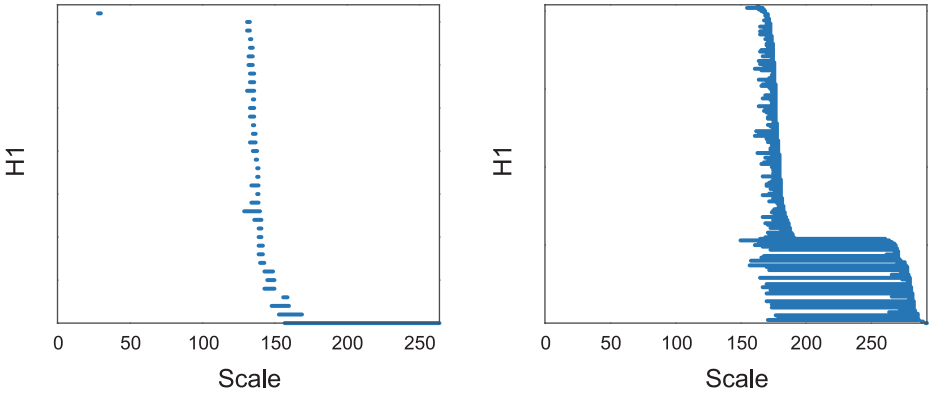


Figure 2.13 Two representations of the persistent homology of data from an evolutionary simulation; see Section 5.7 for discussion. On the left, a barcode diagram. On the right, a persistence diagram. Rather than identifying specific bars with geometric features, in this case the count of the bars conveys important information about the underlying process.



(a) *Klebsiella pneumoniae* (b) *Salmonella enterica*

Figure 2.14 Barcode diagrams reflect different scales of genomic exchange in *K. pneumoniae* and *S. enterica*. Source: [161].

There are algorithms to compute the barcodes with running time cubic in the number of simplices. See Section 2.7 and Appendix A for discussion of the computational aspects of computing persistent homology.

2.4 Stability of Persistent Homology under Perturbation

In order to use topological invariants to describe data, it is essential that small perturbations of the data give rise to small changes in the resulting invariants.

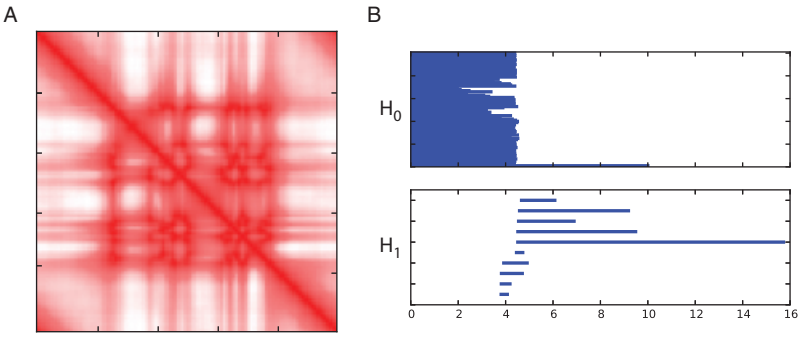


Figure 2.15 DNA can be simulated as a long polymer consisting of a large number of monomeric units interacting at specific places. Here, we show the data of a 50 Mb polymer with 10 fixed loops at random positions in the genome consisting of 1000 monomeric units. (A) The average of 5000 simulations allows us to construct a contact map. (B) Using persistent homology in a similarity matrix derived from the contact map one can clearly identify the ten loops as ten long bars in dimension one persistent classes. Source: [163].

One of the very useful aspects of persistent homology is that the set of barcodes forms a metric space; the distance between barcodes allows us to be precise about measuring changes in the output of topological data analysis. For the input, it turns out to be very useful to adopt a metric on the space of finite metric spaces, the Gromov-Hausdorff distance. These metric space structures make it possible to prove *stability theorems* that relate perturbation of the input data in the Gromov-Hausdorff metric to perturbation of the output barcodes in the barcode metric [105, 117].

These stability results are the most important theorems in the subject. In order to understand what they really say, we need to explain

1. what it means for two finite metric spaces to be close in the Gromov-Hausdorff metric, and
2. what it means for two barcodes to be close in the barcode metric.

Definition 2.4.1. Let A and B be non-empty subsets of a metric space (X, ∂_X). Then we define the *Hausdorff distance* between A and B to be

$$d_H(A,B) = \max\left(\sup_{a\in A}\inf_{b\in B}\partial_X(a,b), \sup_{b\in B}\inf_{a\in A}\partial_X(a,b)\right).$$

It is sometimes convenient to consider the equivalent formulation of the Hausdorff distance as

Figure 2.16 The Hausdorff distance is determined by the point in A with the largest distance to the closest point in B (and vice versa).

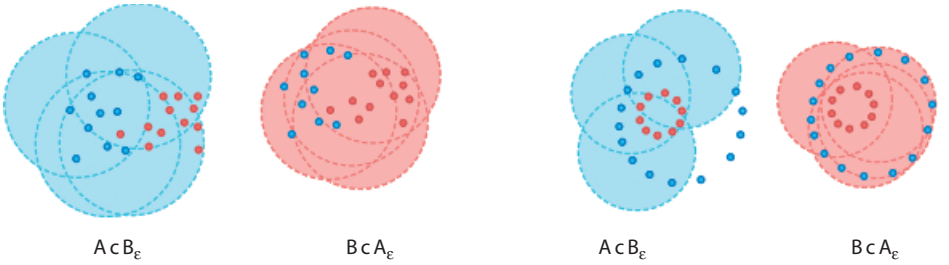


Figure 2.17 The Hausdorff distance can be computed by considering the smallest ϵ fattening of each set that contains the other.

$$d_H(A, B) = \inf_{\epsilon>0}\{B \subseteq A_\epsilon, A \subseteq B_\epsilon\},$$

where A_ϵ and B_ϵ denotes the sets of all points within distance ϵ of A and B, respectively (see Figures 2.16, 2.17).

Example 2.4.2.

1. Let $A \subset X$ and suppose that B is generated from A by perturbing each point $a \in A$ by at most ϵ; i.e., the points of B are in bijection with those of A and (denoting the bijection by θ) we have $\partial_X(a, \theta(a)) \leq \epsilon$. For instance, consider $A = \{[0, 0, 0], [1, 2, 3], [-1, 0, 5]\} \subset \mathbb{R}^3$ and $B = \{[\epsilon, 0, 0], [1, 2 + \epsilon, 3], [-1, 0, 5 - \epsilon]\}$. Then $d_H(A, B) \leq \epsilon$.
2. The Hausdorff distance is heavily influenced by the single most extreme point; given $A \subset X$, let $A' = A \cup \{x\}$. Then $d_H(A, A') = \min_{a\in A} \partial_X(x, a)$.

Lemma 2.4.3. *The Hausdorff distance imposes a metric on the set of non-empty subsets of a metric space* (X, ∂_X).

However, we cannot in general assume that the metric spaces we consider are given as subsets of a common ambient metric space. A key insight of Gromov is to circumvent this issue by considering the infimum of the Hausdorff distance over all isometric embeddings of the two metric spaces into a larger ambient metric space. Here an isometric embedding

$$\phi\colon (X, \partial_X) \to (Y, \partial_Y)$$

is an injective map $X \to Y$ such that

$$\partial_X(x_1, x_2) = \partial_Y(\phi(x_1), \phi(x_2)).$$

That is, an isometric embedding identifies X with a submetric space of Y.

Definition 2.4.4. Let (X_1, ∂_{X_1}) and (X_2, ∂_{X_2}) be compact metric spaces. The *Gromov-Hausdorff distance* between X_1 and X_2 is defined to be

$$d_{GH}((X_1, \partial_{X_1}), (X_2, \partial_{X_2})) = \inf_{\substack{\theta_1\colon X_1 \to Z \\ \theta_2\colon X_2 \to Z}} d_H(X_1, X_2).$$

Here θ_1 and θ_2 are isometric embeddings of (X_1, ∂_{X_1}) and (X_2, ∂_{X_2}) in (Z, ∂_Z) respectively (see Figure 2.18 for an example); the infimum is taken over all such (Z, ∂_Z) and embeddings θ_1 and θ_2.

We will say that two metric spaces are *isometric* if there exists an isomorphism $f\colon X \to Y$ that preserves all distances. This clearly defines an equivalence relation on the set of metric spaces.

Theorem 2.4.5. *The Gromov-Hausdorff distance is a metric on the set of isometry classes of compact metric spaces.*

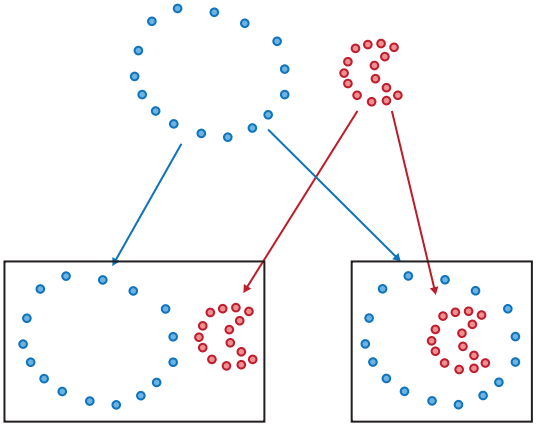


Figure 2.18 The Gromov-Hausdorff distance is computed by minimizing over all embeddings; here, the embedding on the right has a much smaller Hausdorff distance between the two image sets than the embedding on the left.

As defined above, it is hard to see how one might ever compute the Gromov-Hausdorff distance in practice. For this purpose, an alternative formulation is useful; it is also conceptually helpful in understanding what d_{GH} is measuring. Let $\mathcal{R}$ be a correspondence between X_1 and X_2, i.e., a subset of $X_1 \times X_2$ such that there exists a tuple with first coordinate x for each $x \in X_1$ and a tuple with second coordinate y for each $y \in X_2$.

The Gromov-Hausdorff distance can now be described by the formula

$$d_{GH}((X_1, \partial_{X_1}), (X_2, \partial_{X_2})) = \inf_{\mathcal{R} \subseteq X_1 \times X_2} \frac{1}{2} \left(\sup_{(x,x') \in \mathcal{R}, (y,y') \in \mathcal{R}} |\partial_{X_1}(x, y) - \partial_{X_2}(x', y')| \right).$$

Roughly speaking, the Gromov-Hausdorff distance measures the maximum distortion in the best matching between the two metric spaces.

Example 2.4.6.

1. Suppose that X' is an ϵ-net in X (recall that this means that for each $x \in X$, there exists a point $x' \in X'$ such that $\partial_X(x, x') < \epsilon$). Then $d_{GH}((X', \partial_X), (X, \partial_X)) < \epsilon$.
2. Let (X, ∂_X) be a metric space and suppose that $(X', \partial_{X'})$ is formed by adding a single point $\{z\}$ to X such that $\partial_{X'}(z, x) = \kappa > \text{diam}(X)$ for any $x \in X$. (That is, we are adding a single point to X which is "far away" from the rest of the points.) Then $d_{GH}((X, \partial_X), (X', \partial_{X'})) > \frac{\kappa}{2}$.
3. Suppose that (X, ∂_X) and (Y, ∂_Y) are isometric metric spaces. Then $d_{GH}((X, \partial_X), (Y, \partial_Y)) = 0$.

There is an interesting body of work on the topology induced on the set of isometry classes of compact metric spaces by d_{GH}. For our purposes, one thing to observe is that any compact metric space can be approximated as the Gromov-Hausdorff limit of finite metric spaces. (See Figure 2.19 for an example of this kind of convergence.)

Lemma 2.4.7. *Given a compact metric space (X, ∂_X), let $\{X_n\}$ denote a sequence of finite $\frac{1}{n}$-nets in X. Then*

$$\lim_{n \to \infty} d_{GH}((X, \partial_X), (X_n, \partial_X)) = 0.$$

The Gromov-Hausdorff distance is a suitable means for capturing perturbations of data sets that involve bounded changes in each point, and therefore for measuring the impact of certain kinds of noise. On the other hand, Example 2.4.6 makes it clear that arbitrary changes in a constant number of points can cause arbitrary changes in the Gromov-Hausdorff distance. We will return to a discussion of this phenomenon in Chapter 3; see Section 3.4 in particular.

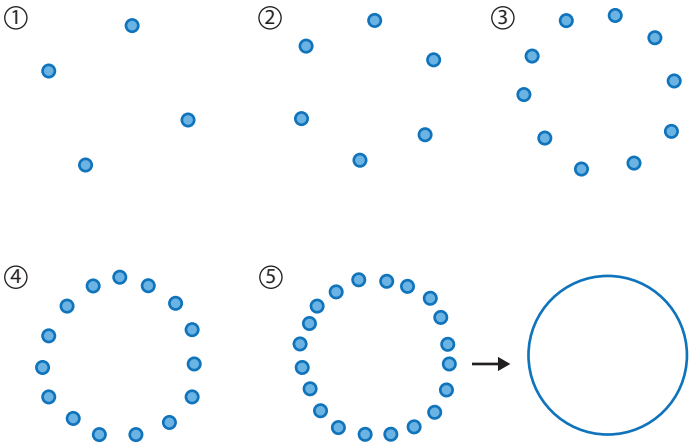


Figure 2.19 Samples of points that lie on a circle converge to the circle in the Gromov-Hausdorff distance as the sampling density increases.

We now turn to the description of various metrics on the set of barcodes. We begin with the *bottleneck distance*. Given two intervals $[a_1, b_1)$ and $[a_2, b_2)$, define

$$d_\infty([a_1, b_1), [a_2, b_2)) = \max(|a_1 - a_2|, |b_1 - b_2|).$$

We extend d_∞ to include $\emptyset$ by defining

$$d_\infty([a, b), \emptyset) = \frac{|b - a|}{2}.$$

Now given two barcodes B_1 and B_2, we define a matching between B_1 and B_2 as follows. Without loss of generality, assume that $|B_1| < |B_2|$. Then a matching is specified by a bijection $\phi\colon A_1 \to A_2$, where A_1 is a multi-subset of B_1 and A_2 is a multi-subset of B_2. We formally add $\emptyset$ to B_1 and B_2, and we regard the elements of $B_1 \setminus A_1$ and $B_2 \setminus A_2$ as matched with $\emptyset$.

Definition 2.4.8. Let B_1 and B_2 be barcodes. The *bottleneck distance* is defined to be

$$d_B(B_1, B_2) = \inf_{\phi} \sup_{Z \in B_1} d_\infty(Z, \phi(Z)),$$

where ϕ varies over all matchings between B_1 and B_2 and the supremum is taken over bars in B_1.

Roughly speaking, the bottleneck distance measures the worst discrepancy in the best matching between the two barcodes. Note that two barcodes which are a distance ϵ apart in the bottleneck distance could differ in an essentially arbitrary number of short bars of length less than $\frac{\epsilon}{2}$. Put another way, two barcodes are close

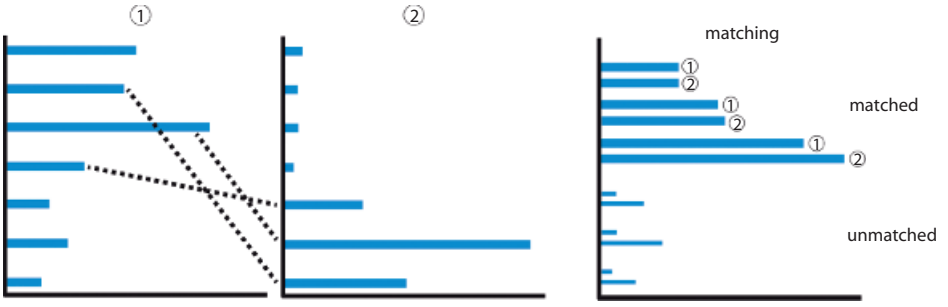


Figure 2.20 The bottleneck distance on barcodes is computed by matching long bars. Figure from experiment performed by Elena Kandror, Abbas Rizvi, and Tom Maniatis at Columbia University, with permission.

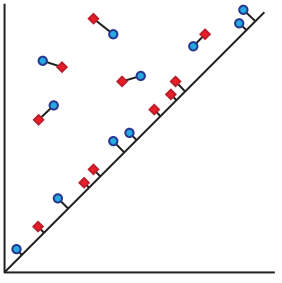


Figure 2.21 The bottleneck distance when expressed in terms of persistence diagrams is computed by matching nearby points and assigning points close to the diagonal to the nearest diagonal point.

in the bottleneck distance if after ignoring "short" bars, the endpoints of matching "long" bars are close (see Figures 2.20 and 2.21 for examples.)

There are other sensible metrics on the space of barcodes, most notably including mass transportation (Wasserstein) metrics. Since it will be convenient for later use, we will also introduce the *Wasserstein metric* here.

Definition 2.4.9. Let B_1 and B_2 be barcodes. For $p > 0$, the *p-Wasserstein distance* is defined to be

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

We can now state the stability theorem for persistent homology, arguably the most important theorem in the subject [117]. (See Figure 2.22 for an illustration.)

Theorem 2.4.10. *Let (X, ∂_X) and (Y, ∂_Y) be finite metric spaces. Then for all $k \geq 0$,*

$$d_B(\mathrm{PH}_k(\mathrm{VR}(X, \partial_X)), \mathrm{PH}_k(\mathrm{VR}(Y, \partial_Y))) \leq d_{GH}((X, \partial_X), (Y, \partial_Y)).$$

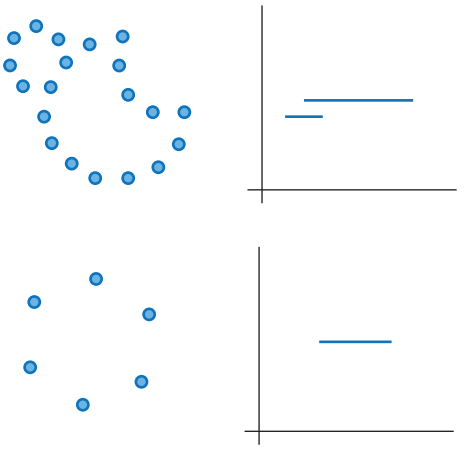


Figure 2.22 The two samples are close together in the Gromov-Hausdorff distance; although at various ϵ the homology groups are different, the barcodes are close together.

Remark 2.4.11. Analogous results hold when using the Čech complex or using the Wasserstein metric.

There are versions of the stability theorem expressed in terms of the "Morse filtration" approach to persistent homology as well. The set of functions $\{f\colon X \to \mathbb{R}\}$ can be endowed with a metric specified as

$$d_\infty(f,g) = \sup_{x\in X} |f(x) - g(x)|.$$

We say that a function $f\colon X \to \mathbb{R}$ is admissible if $H_k(f^{-1}(-\infty, t]; \mathbb{F})$ is finite rank for all $t \in \mathbb{R}$.

Theorem 2.4.12. *Let X be a topological space. Let $f, g\colon X \to \mathbb{R}$ be admissible functions. Then for all $k \geq 0$,*

$$d_B(\mathrm{PH}_k(X,f), \mathrm{PH}_k(X,g)) \leq d_\infty(f,g).$$

Using the observation of Remark 2.3.5 and the relationship between the Čech and Vietoris-Rips complex, we can regard Theorem 2.4.12 as a generalization of Theorem 2.4.10.

Remark 2.4.13. Theorems 2.4.10 and 2.4.12 are incarnations of an algebraic stability theorem, which says that for persistence modules that are κ-interleaved (which is a precise way of expressing the notion of being approximately

isomorphic), the resulting barcodes are within κ in the bottleneck metric [42, 107]. This formulation of the stability theorem allows us to substantially weaken the hypotheses necessary to apply it and also extends its reach.

2.5 Zigzag Persistence

Persistent homology is defined in situations where we have a filtered system of complexes. As we have described above, these filtrations typically arise by varying a scale parameter of some sort. Sometimes, however, we might not expect to have a filtration but rather some kind of more general diagram. That is, a natural question that arises is whether other "filtration shapes" could be used as input. We now discuss an answer to the following specific form of this question [91].

Question 2.5.1. Does a construction like persistent homology make sense when considering "filtrations" in which not all the arrows go in the same direction?

This more general kind of diagram can easily arise in practice. For example, suppose we consider many sets of samples X_i from each fixed metric space (X, ∂_X). We then can form the sequence

$$X_1 \longrightarrow X_1 \cup X_2 \longleftarrow X_2 \longrightarrow X_2 \cup X_3 \longleftarrow X_3 \longrightarrow \dots$$

where the maps are the obvious inclusions (Figure 2.23).

Applying the composite of $H_k(-;\mathbb{F})$ and the Vietoris-Rips complex functor (for some fixed ϵ) to this sequence yields a corresponding diagram of $\mathbb{F}$-vector spaces

$$\begin{array}{ccccccc} H_k(\mathrm{VR}_\epsilon(X_1)) & \longrightarrow & H_k(\mathrm{VR}_\epsilon(X_1 \cup X_2)) & \longleftarrow & H_k(\mathrm{VR}_\epsilon(X_2)) & & \\ & & & & \downarrow & & \\ & & & & H_k(\mathrm{VR}_\epsilon(X_2 \cup X_3)) & \longleftarrow & H_k(\mathrm{VR}_\epsilon(X_3)) \longrightarrow \dots \end{array}$$

In order to study these sorts of "filtrations" more carefully, we need to develop some notation for describing the pattern of arrows. To do this, we consider *zigzag diagrams* of shape S, where S is a string on the alphabet L, R.

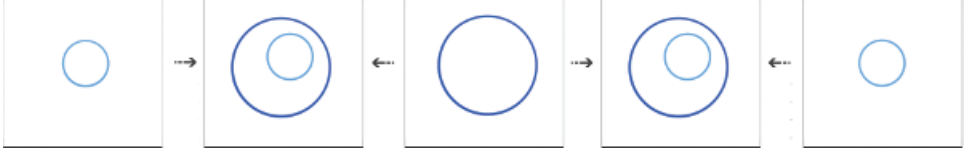


Figure 2.23 We get a natural zigzag by taking unions of samples.

Definition 2.5.2. A *zigzag diagram* (or *zigzag module*) of shape S is defined to be a sequence of linear transformations between $\mathbb{F}$-vector spaces:

$$X_1 \xrightarrow{f_1} X_2 \xrightarrow{f_2} \dots \xrightarrow{f_{k-1}} X_k,$$

where each map f_i has its direction specified by the ith letter of the string S. (This is also referred to as a zigzag module.)

The definition of a zigzag diagram is a strict generalization of the notion of a filtration. When the shape S is $RRRRRRRRR \dots R$ or $LLLLLLL \dots L$, a zigzag diagram is simply a filtered $\mathbb{F}$-module.

Example 2.5.3.

1. Let $S = RRR$. Then a zigzag diagram of shape S is a diagram
$$M_1 \to M_2 \to M_3 \leftarrow M_4$$
of vector spaces.
2. Let $S = RLRLRL$. Then a zigzag diagram of shape S is a diagram
$$M_1 \to M_2 \leftarrow M_3 \to M_4 \leftarrow M_5 \to M_6 \leftarrow M_7$$
of vector spaces.

In the original setting for persistent homology, it was intuitively clear that the "lifespan" of a homological feature was an interesting topological invariant associated to a filtration. When working with zigzag diagrams, the corresponding idea is that of a homological feature that is "consistent" across the zigzag. For example, if we are considering a zigzag of shape RL,

$$M_1 \xrightarrow{f_1} M_2 \xleftarrow{f_2} M_3,$$

then a zigzag feature should represent a collection of elements $m_1 \in M_1, m_2 \in M_2, m_3 \in M_3$ consistent in the sense that $f_1(m_1) = m_2 = f_2(m_3)$. In the context of the sampling example we started with, a zigzag feature should represent some kind of geometric property that is stable across different samples.

To work with this notion, one would again hope for an analogue of Theorem 2.3.10 that allows us to characterize homological invariants of zigzag diagrams in terms of some kind of numerical invariant like barcodes. We now switch to using the zigzag module terminology.

Definition 2.5.4. A *zigzag submodule* N of a zigzag module M of shape S is a zigzag module of shape S such that each N_i is a subspace of M_i and the maps are determined by the restrictions of the f_i.

Example 2.5.5. Let $\mathbb{F} = \mathbb{R}$, and suppose we are given the zigzag module

$$\mathbb{R} \longrightarrow \mathbb{R}^2 \longleftarrow \mathbb{R},$$

where the first map is $x \mapsto (x, 0)$ and the second map is $x \mapsto (0, x)$. Then there is a zigzag submodule

$$\mathbb{R} \longrightarrow \mathbb{R} \longleftarrow \mathbb{R},$$

where the $\mathbb{R}$ in the middle comes from the first coordinate of $\mathbb{R}^2$; the maps are now $x \mapsto x$ and $x \mapsto 0$.

We say that a zigzag submodule M is *decomposable* if it can be written as the direct sum of non-trivial submodules $\{N_j\}$ (recall Definition 1.6.40); otherwise, we say it is *indecomposable*.

Lemma 2.5.6. *Any zigzag module M of shape S can be written as a direct sum of indecomposables in a way that is unique up to permutation.*

Indecomposable zigzag modules have a very constrained form.

Definition 2.5.7. An *interval zigzag module* of shape S is a zigzag module

$$X_1 \xrightarrow{f_1} X_2 \xrightarrow{f_2} \dots \xrightarrow{f_{k-1}} X_k,$$

where for fixed $a \leq b$,

$$\begin{cases} X_i = \mathbb{F}, & 1 \leq a \leq i \leq b \leq k \\ X_i = 0, & \text{otherwise} \end{cases}$$

and the maps between the $\mathbb{F}$ are the identity map, and the zero map otherwise.

We can now state the main theorem that gives rise to zigzag barcodes.

Theorem 2.5.8. *The indecomposable zigzag modules are precisely the interval zigzag modules.*

As a consequence, we can obtain a barcode multiset which is referred to as the zigzag persistence, and tends to be represented the same way as persistence barcodes (Figure 2.24).

In Section 5.4.3, zigzag persistence is used to study HIV in tissue samples taken from central nervous system (CNS) and non-CNS regions. See Figure 2.25 for an indication of the data.

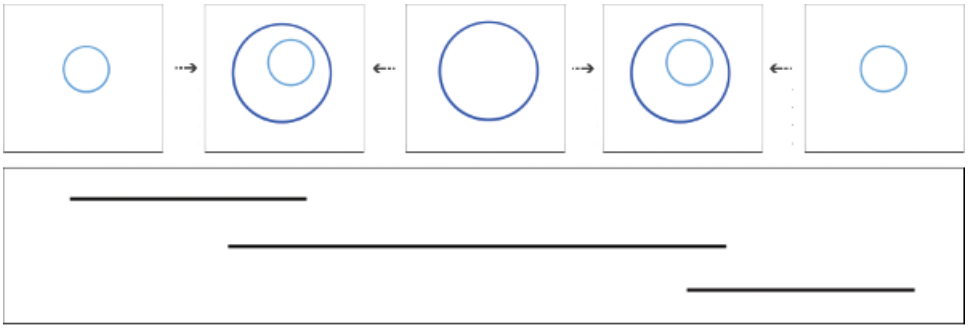


Figure 2.24 The bars represent features that persist across zigzags.

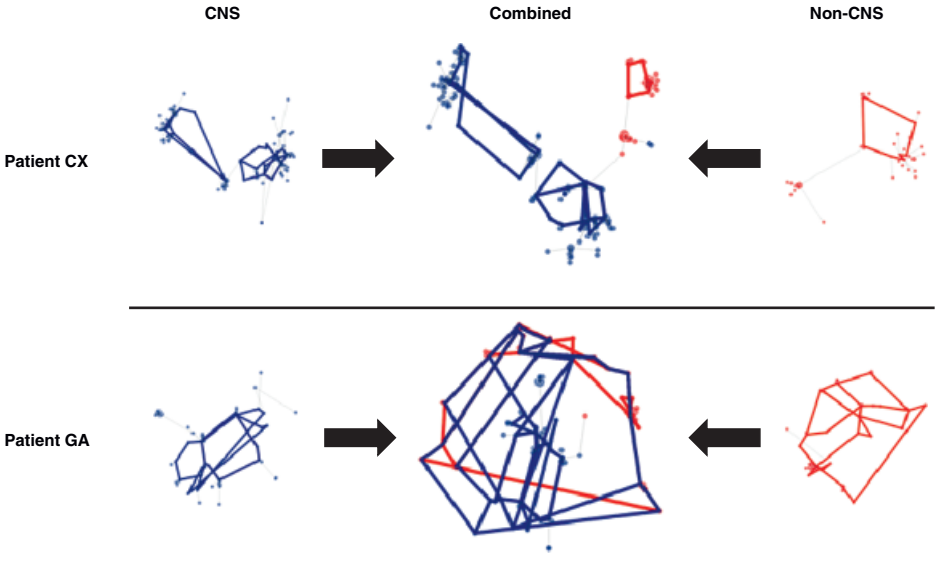


Figure 2.25 Phylogenetic networks of HIV-1 gp120 sequences obtained from Patients CX and GA. Each node represents one sequence; larger nodes show sequences that were sampled multiple times. Blue nodes were sampled from the CNS; red nodes were sampled from elsewhere in the body. The position of each node is determined by the first two principal components (computed via MDS) of genetic distance (Hamming distance). The network backbone (thin gray edges) is a minimum spanning tree, and the thick red and blue edges are generators of cycles identified by persistent homology. Red cycles denote putative recombination events that involve sequences sampled fully outside the CNS; blue cycles denote events that involve some sequences from the CNS.

We now discuss a basic zigzag that arises from metric data. In what follows, let (X, ∂_X) be a finite metric space, and choose an ordering for the points – we will denote the ordering as

$$X = \{x_1, x_2, x_3, \ldots\}.$$

Let X_k denote the subset of X consisting of the first k points in the ordering, i.e.,

$$X_1 = \{x_1\}, \quad X_2 = \{x_1, x_2\}, \quad X_3 = \{x_1, x_2, x_3\},$$

and so on. We can then define a series of distinguished scales $\epsilon_i = d_H(X_k, X)$. Notice that $\epsilon_i \geq \epsilon_{i+1}$; X_{i+1} will always be at least as close to X as X_i in the Hausdorff distance.

Definition 2.5.9. Choose real numbers $\alpha > \beta > 0$. The *Rips zigzag* consists of the zigzag module specified by the diagram of simplicial complexes

$$\begin{array}{ccccccc} \ldots & \longleftarrow & \mathrm{VR}_{\beta\epsilon_{i-1}}(X_{i-1}) & \longrightarrow & \mathrm{VR}_{\alpha\epsilon_{i-1}}(X_i) & \longleftarrow & \mathrm{VR}_{\beta\epsilon_i}(X_i) \\ & & & & & & \downarrow \\ & & & & \mathrm{VR}_{\alpha\epsilon_i}(X_{i+1}) & \longleftarrow & \mathrm{VR}_{\beta\epsilon_{i+1}}(X_{i+1}) \longrightarrow \ldots \end{array}$$

Notice that the constituent complexes in this zigzag have size controlled by the limits α and β; it was originally proposed by Morozov for the purpose of computational efficiency. Work of Oudot and Sheehy [393] provides theoretical validation for the use of this zigzag, showing that when $X \subseteq \mathbb{R}^n$ is close in Hausdorff distance to a well-behaved compact subset $Y \subseteq \mathbb{R}^n$, then there are long zigzag intervals in the Rips zigzag that permit recovery of the homology of X for suitable α and β. (As with Theorem 2.2.1, the actual numerical bounds extracted from these results are much larger than needed in practice.)

Finally, given the central importance of the stability theorem for persistent homology, one would hope for something similar in the context of zigzag persistence. In [91], stability results were proved in the context of a particular construction of zigzags from finite metric spaces, the *level set zigzag diagram*. In general, the specific form of stability results depends on the particulars of the process of constructing the zigzag. Nonetheless, theoretical considerations [67] show that essentially any reasonable procedure for producing zigzag modules will have some kind of stability theorem.

2.6 Multidimensional Persistence

The underlying idea of persistence, namely that a sensible way to cope with uncertainty about parameter settings is simply to aggregate information as the parameter changes, is a powerful and general one. But why limit ourselves to just the feature scale? There are often many parameters which we might like to apply persistence to: for example, in the motivating example for zigzag persistence, it would make sense to vary both the samples and the feature scale ϵ. And in many probabilistic

settings we want to simultaneously vary a density parameter as well as ϵ. In this section we discuss two approaches to considering persistence in multiple directions. First, we explain a systematic framework for multidimensional persistence. Then we discuss a closely related idea, the persistent homology transform.

2.6.1 Multidimensional Persistence

In many situations, it is natural to consider multiple filtrations on a data set; e.g., for a finite metric space (X, ∂_X) one filtration will come from the distance scale parameter and another from an additional property of the data. A key motivating example arises when the density of the data is not uniform: it often makes sense to consider one filtration direction generated by the distance scale and another generated by density.

Provided that these filtrations interact in a natural way, we can define multidimensional persistent homology as a generalization of the definition of persistent homology given above. Specifically, we regard $\mathbb{R}^n$ as a partially ordered set and hence a category by setting $\{a_1, \ldots, a_n\} \leq \{b_1, \ldots, b_n\}$ when each $a_i \leq b_i$.

Definition 2.6.1. A *multifiltered system of simplicial complexes* is a functor from $\mathbb{R}^n$ to simplicial complexes. A *multifiltered vector space* is a functor from $\mathbb{R}^n$ to $\mathbb{F}$-vector spaces.

Explicitly, for $n = 2$, a multifiltered complex $\{X_{\alpha,\beta}\}$ is specified by a commutative diagram

$$\begin{array}{ccc} X_{a,b} & \longrightarrow & X_{x_1,y_1} \\ \downarrow & & \downarrow \\ X_{x_2,y_2} & \longrightarrow & X_{c,d} \end{array}$$

for any $x_1, x_2 \in [a, c]$ and $y_1, y_2 \in [b, d]$.

Example 2.6.2. Suppose we have a finite metric space (X, ∂_X) and a codensity function $\gamma\colon X \to \mathbb{R}$, where γ is small at higher density points and large at sparse points. For example, γ could be a normalized count of the distance to the kth-nearest neighbor. (Here k is a parameter that has to be chosen.) Then we define a functor

$$\mathbb{R} \times \mathbb{R} \to \mathrm{Simp}$$

via the formula

$$(\epsilon, \delta) \mapsto \mathrm{VR}_\epsilon(\gamma^{-1}(-\infty, \delta]).$$

Given any multifiltered complex, by passing to homology, we obtain a multifiltered vector space, the multidimensional persistent homology. There is again

a structure theorem for multifiltered vector spaces, but in contrast to the one dimensional case, the irreducible objects are not easily described. As a consequence, there is no tractable analogue of the barcode in this context which completely describes the isomorphism type of the multifiltered vector space, and so no easy summarization of the results of computing multidimensional persistent homology.

A number of possible solutions to this problem have been proposed: even though there is no complete invariant, there are many interesting invariants which capture partial information that are relevant to data analysis.

1. Zomorodian and Carlsson proposed the rank invariant: this is the numerical invariant obtained by taking the ranks of the maps in the filtration [92].
2. Lesnick and Wright studied in detail the "fibered barcode," a version of the rank invariant (introduced under a different name in [99]), which is the collection of invariants obtained by choosing lines through the filtrations and computing the one dimensional persistence in that direction [325]. They have developed a tool, Rivet [324], that supports exploratory data analysis in this context, displaying the rank invariant as well as the bigraded Betti numbers. See Figures 2.26 and 2.27 for examples.

2.6.2 The Persistent Homology Transform

In the general spirit of persistence, one approach to choosing lines through the filtration is to consider the collection of all of them at once. We now discuss an implementation of this idea in the restricted context of surfaces embedded in Euclidean space.

Beyond difficulties in computing persistent homology, as we discuss in detail in Chapter 3, it can be difficult to interpret the results of persistent homology computations even for data embedded in comparatively low-dimensional Euclidean spaces $\mathbb{R}^n$ for $n > 3$. One approach to this issue is to restrict attention to spaces embedded in $\mathbb{R}^2$ or $\mathbb{R}^3$; such examples arise when considering surfaces, for instance. In the setting of cancer genomics, motivating examples arise from the imaging of tumors, as we discuss in a bit more detail in Section 3.8.

When working in $\mathbb{R}^2$ or $\mathbb{R}^3$, filtrations generated by a height function seem particularly useful. However, one issue with filtrations generated by height functions is that they depend on a choice of orientation – along which direction do we measure height? Just as the basic idea of persistence is to consider all scales at once, a simple approach is to consider all possible orientations at once. We now explain a direct approach to considering a kind of multidimensional persistence in this setting [510].

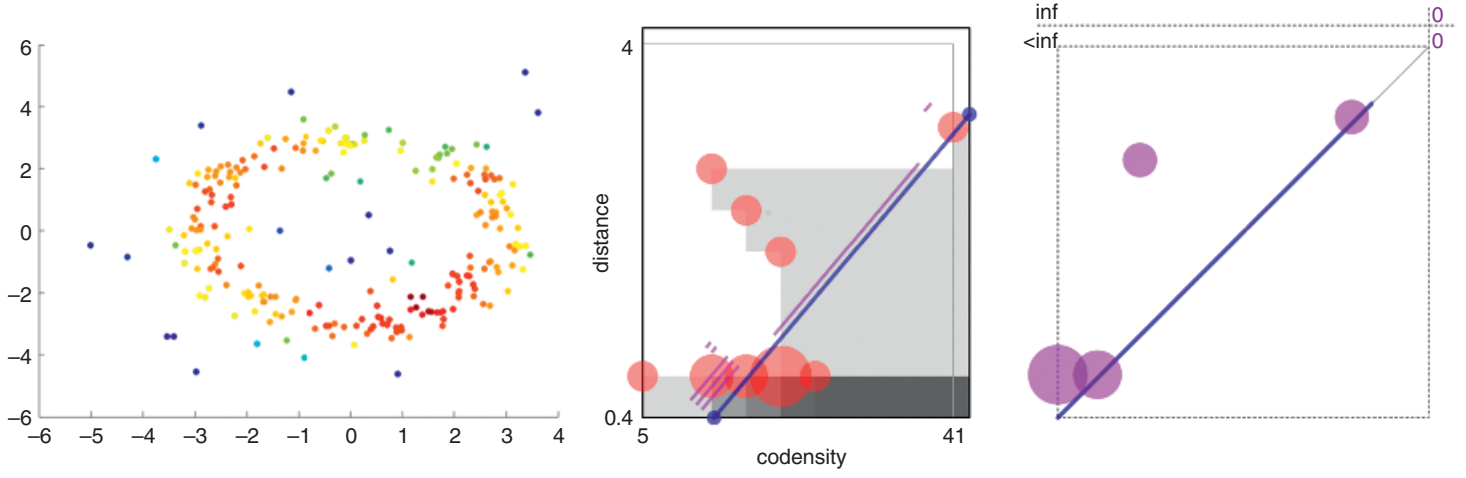


Figure 2.26 Persistence along a line that increases codensity and feature scale; in dimension zero, there is a single long bar that appears in this direction. In dimension 1, we see a single point away from the diagonal; simultaneously filtering by scale and density reduces the impact of noise. The size of the dots indicates multiplicity of the point, and the shading reflects the dimension of the vector space.

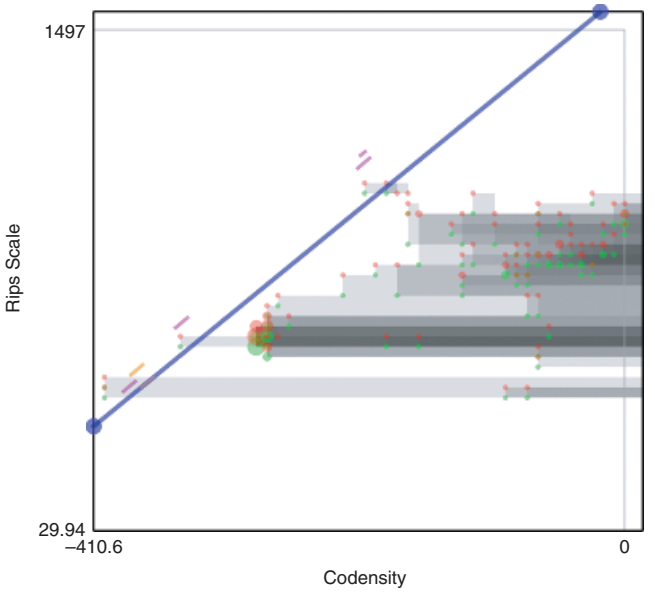


Figure 2.27 Multidimensional persistence for the HIV data set. From Monica Nicolau, Arnold J. Levine, Gunnar Carlsson, *Proceedings of the National Academy of Sciences* Apr 2011, 108 (17), 7265–7270. Reprinted with Permission from *Proceedings of the National Academy of Sciences*.

Suppose that our data is presented as a finite simplicial complex M embedded in $\mathbb{R}^d$. For each direction, represented by a point $v \in S^{d-1}$, we define a filtration of M as

$$M(v)_\epsilon = \{x \in M \mid x \cdot v \leq \epsilon\}.$$

We can now consider summarizing M by considering the persistent homology of each of these filtrations in aggregate. Specifically, we have the following definition.

Definition 2.6.3. The *persistent homology transform* of $M \subseteq \mathbb{R}^d$ is the function

$$\mathrm{PHT}\colon S^{d-1} \to \mathcal{B}^d$$

specified by the assignment

$$v \mapsto [\mathrm{PH}_0(M(v)_\bullet), \mathrm{PH}_1(M(v)_\bullet), \ldots, \mathrm{PH}_d(M(v)_\bullet)].$$

The main theorem of [510] shows that in dimensions 2 and 3, we can use the collection of persistent homologies here to uniquely characterize the input object.

Theorem 2.6.4. *Let $d = 2$ or $d = 3$. Then* PHT *specifies an injective function from the set of finite simplicial complexes $M \subset \mathbb{R}^d$ to the set of functions from S^{d-1} to $\mathcal{B}^d$.*

2.7 Efficient Computation of Persistent Homology

In order for topological data analysis to be useful in practice, it must be possible to efficiently compute invariants like PH of real data sets. For example, one reason for the ubiquity of linear regression, PCA, and clustering in data analysis is the ease of computation, even for large data sets. Moreover, since many applications of TDA are in the context of exploratory data analysis, it is important that repeatedly recomputing with different parameters be feasible. In this section, we give an overview of the source of computational difficulty in applying TDA; Appendix A has a more detailed discussion of specific software packages.

As a baseline for comparison, we note the following.

1. Computing the single-linkage clustering dendrogram for a finite metric space (X, ∂_X) where $|X| = n$ can be done in time proportional to $n \log n$.
2. Similarly, Mapper (described in Section 2.8) can also be computed very efficiently.

Persistent homology is another matter. As is evident from the discussion of the computation of homology, persistent homology cannot be computed much more efficiently than matrix multiplication on matrices with dimensions given by the number of simplices – and for non-sparse matrices, practical algorithms for matrix multiplication are roughly cubic.

To compute persistent homology, we proceed as follows. Suppose that we have a filtered simplicial complex X. We choose a total ordering of the simplices of X that is compatible with the filtration on X; i.e., $\sigma < \tau$ if σ appears in a lower filtration than τ. (The order of simplices within a given filtration level is arbitrary.) Let n denote the number of simplices of X. We now form the $n \times n$ matrix D defined by the formula

$$D_{i,j} = \begin{cases} 1, & \text{if } \sigma_i \text{ is a codimension 1 face of } \sigma_j \\ 0, & \text{otherwise.} \end{cases}$$

We now define $\text{low}(j)$ to be the row number of the last 1 in column j; we set $\text{low}(j) = 0$ if column j consists only of zero entries. We will say that the matrix D is *reduced* if $\text{low}(j_1) \neq \text{low}(j_2)$ for $j_1 \neq j_2$. The following algorithm reduces the matrix D:

```
for j = 1 to n
  while there exists k < j such that low(k) = low(j) != 0:
     add column k to column j
```

The algorithm clearly terminates, since each step decreases low in a given column. We can extract the persistence diagram from the reduced form of D by observing that the pairs $(j, \text{low}(j))$ specify persistence intervals.

The serious issue that arises here is the dependence of the running time on the number of simplices. For example, for the Vietoris-Rips complex (or the Čech complex), this can be a problem when the feature scale ϵ approaches the maximum distance between any pair of points in the data set.

Lemma 2.7.1. *Let (X, ∂_X) be a finite metric space, and choose $\epsilon > \text{diam}(X)$, i.e.,*

$$\forall x, y \in X, \qquad \epsilon > \partial_X(x, y).$$

Then $\text{VR}_\epsilon(X, \partial_X)$ has $2^{|X|}$ simplices.

The inexorable conclusion of Lemma 2.7.1 is that in order to efficiently compute persistent homology, it will be necessary to control the number of simplices. One way to do this is to only work with low-dimensional homology; state of the art implementations (see Appendix A) can handle thousands of points when computation is limited to H_1. A general approach to this problem is simply to study the Vietoris-Rips complex over a range $[0, \epsilon_{\max}]$ that ensures a tractable number of simplices at $\epsilon_{\max}$. Another technique is to take many subsamples from (X, ∂_X) such that each subsample results in tractable persistent homology computations, and then combine the persistent homology of the subsamples in some way to estimate the persistent homology of X. This idea is part of the motivation for zigzag persistence, notably the Rips zigzag of Definition 2.5.9. Because zigzag persistence can be used in contexts where we control the size of the maximal complex, modern implementations can be used on data sets with thousands of points. Moreover, techniques for combining such subsamples in a systematic way along with methods for understanding error and variability in the results lead us naturally into the domain of statistical methods; we explore this in detail in the next chapter.

Another possibility is to construct a smaller complex. An early approach to this is the weak witness (or weak Delaunay) complex [458]. The idea is to choose as vertices a set of *landmarks* but use all of the data points to determine the complex.

Definition 2.7.2. Let (X, ∂_X) be a finite metric space. Consider a set of points

$$A = \{x_0, x_1, \ldots, x_k\} \subset X.$$

Then a point $w \in X$ is a *weak witness* for A if $\partial_X(w, x_i) \leq \partial_X(z, x_i)$ for all i and $z \in X - A$.

Roughly speaking, the witness complex will only include simplices for which weak witnesses exist.

Definition 2.7.3. Let $L \subset X$ be a subset of the finite metric space (X, ∂_X). The *witness complex* is the simplicial complex specified by the rule that a simplex $[x_0, x_1, \ldots, x_k]$ for $x_i \in L$ is in the complex if all subsimplices admit weak witnesses.

In practice, the landmarks are often picked randomly or using an algorithm to maximize dispersion (Figure 2.28).

Although very attractive from the perspective of efficiency, the witness complex has problematic theoretical properties:

1. There do not appear to be good stability theorems for the witness complex,
2. the dependence on choice of landmarks is not well understood [105], and
3. the witness complex can fail to reconstruct the homotopy type even in simple examples [215].

In light of these issues, we believe that the only way to extract information from witness complexes is by using the statistical techniques outlined in the next chapter.

For points embedded in $\mathbb{R}^n$, other "small" complexes come from consideration of the Voronoi tesselation of $\mathbb{R}^n$. For example, the Delaunay complex is the simplicial complex obtained as the nerve of the cover of $\mathbb{R}^n$ given by the sets U_x for $x \in X$, where

$$U_x = \{z \in \mathbb{R}_n \,|\, \partial_X(x, z) \leq \partial_X(x', z)\ \forall x' \in X\}.$$

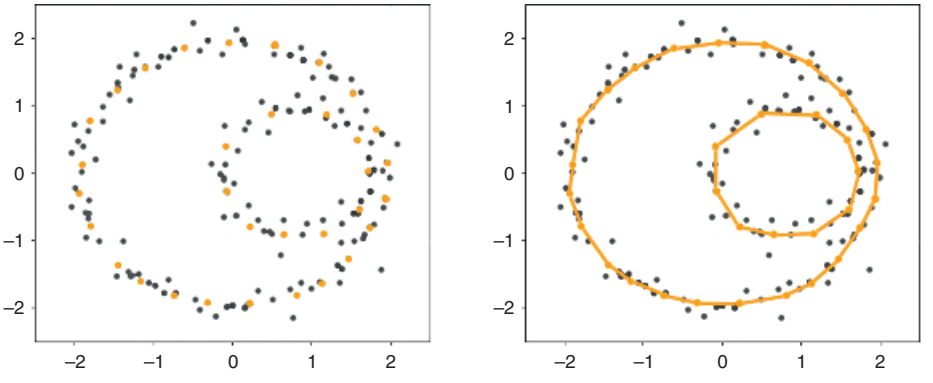


Figure 2.28 The landmark points give rise to concise simplicial circles that capture the topology of the data.

In low dimensions, the Delaunay complex can be computed very efficiently and faithfully recovers the homotopy type of X, although the dependence on the ambient dimension is exponential in general. A variant of this is called the α-complex, which again can be computed efficiently in low dimensions. (Both of these complexes can be computed for data sets consisting of thousands of points via state the art packages.)

As another example, using techniques from the theoretical computer science literature about approximation of metric spaces, the paper [455] explores how to build a hierarchical collection of approximations to suitable finite metric spaces such that for any given accuracy the computation time is linear in the number of points X. Here, suitable means that the metric space has constant doubling dimension, which is a measure of how the volume of balls changes as the radius changes. Note however that metric spaces with doubling dimension d admit low distortion embeddings into $\mathbb{R}^d$; from a practical perspective, it is not clear when these complexes are useful.

2.8 Multiscale Clustering: Mapper

For very large data sets, the techniques of topological data analysis described above can be computationally infeasible. For example, the number of simplices in the Vietoris-Rips complex can grow too rapidly for computation of higher (persistent) homology to be practical. (See Section 2.7 and Section 3.4 for various ways to address this problem.) Another issue is that the output of persistent homology can be hard to interpret for large high-dimensional data sets. An approach to answering these questions when handling very large data sets is to consider integration of ideas of persistence with clustering.

In this section, we describe a method for multiscale clustering: this is the Mapper algorithm of Singh, Mémoli, and Carlsson [462]. Roughly speaking, the idea of Mapper is to define a function on the data set, for example a measure of local density, and then perform clustering at different ranges of values of this function and keep track of how the clusters change as these ranges vary.

The basic framework assumes the data is presented as a finite metric space (X, ∂_X) and we choose

- a *filter function* $f\colon X \to \mathbb{R}^n$, and
- a cover $\mathcal{C} = \{U_\alpha\}$ of the range of f in $\mathbb{R}^n$; typically this cover is taken to be a collection of overlapping closed boxes. In the case of $n = 1$, a typical cover is a collection of closed intervals.

We now proceed as follows. This algorithm amounts to a discretization of the Reeb graph (see Section 1.12) at each scale.

1. Cluster each inverse image $f^{-1}(U_\alpha) \subseteq X$, regarded as a metric subspace of X, for all $U_\alpha \in \mathcal{C}$; denote by $C_{\alpha,i}$ the ith cluster. (Any clustering algorithm can be used that takes as input only the interpoint distances and does not require specification of the number of clusters; single-linkage clustering is a standard choice.)
2. Form a graph where the vertices are given by the clusters $C_{\alpha,i}$ as α and i vary and there is an edge between $C_{\alpha,i}$ and $C_{\alpha',j}$ when

$$C_{\alpha,i} \cap C_{\alpha',j} \neq \emptyset \qquad \text{(clusters overlap).}$$

3. Finally, we assign a color to each vertex in the graph corresponding to a particular cluster $C_{\alpha,i}$ according to the average value of f on $x \in C_{\alpha,i}$.

The results are of course dependent on the choice of filter function and the cover; this algorithm is well adapted to the methodology of *exploratory data analysis*, where we are trying to understand the data without an explicit hypothesized model to describe it. For the cover, it is standard to try successive refinements of the range of f, sometimes equally spaced, but often with increased resolution in areas where we expect more interesting behavior to occur. Standard filter functions include density measures and eccentricity measures; these depend on the data, and we will see in the examples and applications many different useful choices of filter function.

Example 2.8.1.

1. Let (X, ∂_X) be any finite metric space, $f\colon X \to \mathbb{R}$ an arbitrary function, and $\mathcal{C} = \{(-\infty, \infty)\}$. Then the output of Mapper is simply the graph consisting of a point for each cluster of (X, ∂_X), no edges, and the clusters colored with the average value of f on the cluster. (See Figure 2.29 for an example.)
2. Let (X, ∂_X) be any finite metric space, $f\colon X \to \mathbb{R}$ an arbitrary function, and $\mathcal{C} = \{[0, 1], [2, 3]\}$. Writing

$$X_{[0,1]} = f^{-1}([0, 1]) \qquad \text{and} \qquad X_{[2,3]} = f^{-1}([2, 3])$$

   the output of Mapper is the union of a collection of vertices for the clusters of $X_{[0,1]}$ and a collection of vertices for the clusters of $X_{[2,3]}$. Again, there are no edges, since the cover does not overlap, and the colors represent the average values of f on the cluster corresponding to the vertex.
3. Now consider the previous example, but modify the cover to be $\mathcal{C} = \{[0, 1], [0.5, 3]\}$. In this case, there are potentially edges between the vertices for clusters that overlap.
4. Consider points sampled densely from a unit circle in $\mathbb{R}^2$, let $f\colon X \to \mathbb{R}$ be the function $(x, y) \mapsto x$ that takes a point to its x-coordinate, and take $\mathcal{C}$ to be a series of overlapping subsets of $[-1, 1]$. (Specifically, we take ten intervals which overlap by 25% on each side.) Then the Mapper graph recovers the circle; see Figure 2.29.

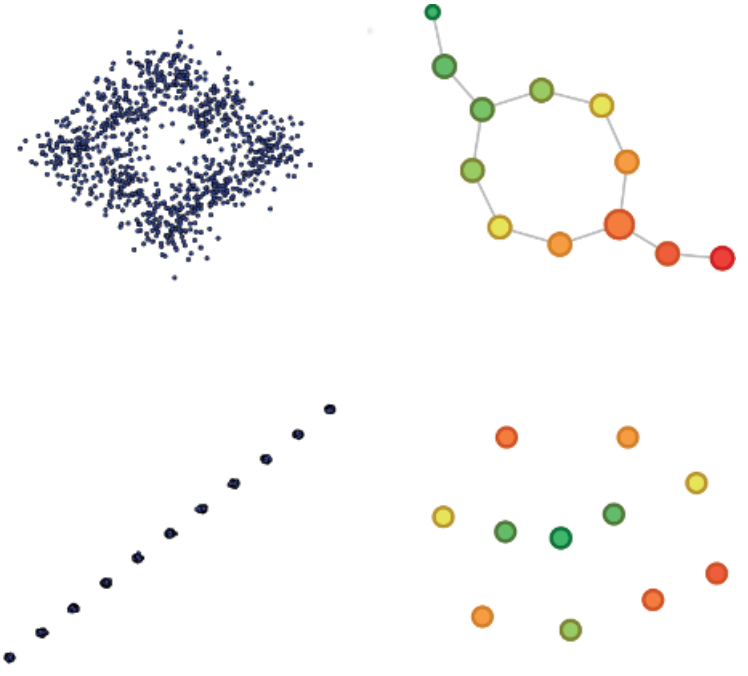


Figure 2.29 Top: The filter function is a projection onto the x-axis and there are 10 overlapping charts; the Mapper graph recovers the topology of the circle. Bottom: When there is a single chart that covers the domain, Mapper just returns the results of clustering, colored by the filter function (in this case, distance from the mean of the data). From Abbas H. Rizvi et al., *Nature Biotechnology* 35, 551–560 (270). © 2017 Nature. Reprinted with Permission from Springer Nature.

In practice, Mapper has turned out to be very useful for identifying clinically significant subsets of the data that are hard to find with traditional clustering methods. It has also been an effective way to represent the structure of the data set across feature scales. To give a sense of what this means, we illustrate with some examples of the use of Mapper on real data.

Example 2.8.2. An early and celebrated example of the application of Mapper was work on a breast cancer data set, by Nicolau, Levine, and Carlsson [383]. The data here is presented as a finite metric space comprising vectors of expression data in $\mathbb{R}^n$. Expression-based classification of tumors is a well-studied problem and has been the subject of a vast number of papers (e.g., see [236, 512]); clustering is the standard technique here. However, there is reason to worry about the efficacy of basic clustering techniques: for example, different tumors activate or suppress pathways with varying strengths, and there is widely variable infiltration of healthy cells into tumor samples. As a consequence, one expects clinically significant features to appear at varying scales.

The analysis used samples from 295 breast cancers as well as additional samples from normal breast tissue; see Figure 2.30. The original expression vectors were in $\mathbb{R}^{24479}$, but a preprocessing step projected them into $\mathbb{R}^{262}$.

- The distance metric was given by the correlation between (projected) expression vectors.
- The filter function used was a measure taking values in $\mathbb{R}$ of the deviation of the expression of the tumor samples relative to normal controls.
- The cover was overlapping intervals in $\mathbb{R}$.

In the Mapper graph, the samples divide into two branches. The lower right branch itself has a subbranch (referred to as c-MYB+ tumors), which are some of the most distinct from normal and are characterized by high expression of genes including c-MYB, ER, DNALI1 and C9ORF116. Interestingly, all patients with c-MYB+ tumors had very good survival and no metastasis. These tumors do not correspond to any previously known breast cancer subtype; the grouping seems to be invisible to classical clustering methods – for example, hierarchical clustering fails to identify this particular subset of tumors (see bottom left of Figure 2.30). We will study this example in detail in Section 6.7.

Example 2.8.3. Another interesting application of Mapper is to the study of the differentiation process from murine embryonic stem cells to motor neurons. The process is demonstrated in Figure 2.31; over time, undifferentiated embryonic cells become differentiated motor neurons when retinoic acid and sonic hedgehog (a differentiation-promoting protein) are applied.

The data generated corresponds to RNA expression profiles from roughly 2000 single cells.

- The distance metric was provided by correlation between expression vectors.
- The filter function used was multidimensional scaling (MDS) projection into $\mathbb{R}^2$; as we review in Section 4.2, this is a procedure for embedding an arbitrary metric space in a lower dimensional Euclidean space.
- The cover was overlapping rectangles in $\mathbb{R}^2$.

As can be seen in Figure 2.32, the Mapper diagram neatly identifies various regions characterized by their state in the differentiation process; in contrast, conventional clustering directly applied to the raw metric data does not produce clusters that encode information about the progress of differentiation. We will study this example in Section 7.3.

One potential concern for applications is the fact that the Mapper algorithm is not stable in the sense that we have described for persistent homology. For one thing, choice of parameters for the clustering algorithm can lead to unstable results; for example, when hierarchical clustering is used, the results are very sensitive to the choice of cutoff parameter. Worse, it is possible to construct examples of metric spaces (X, ∂_X) and a cover $\mathcal{C}$ such that two very similar filter functions give rise to very different results.

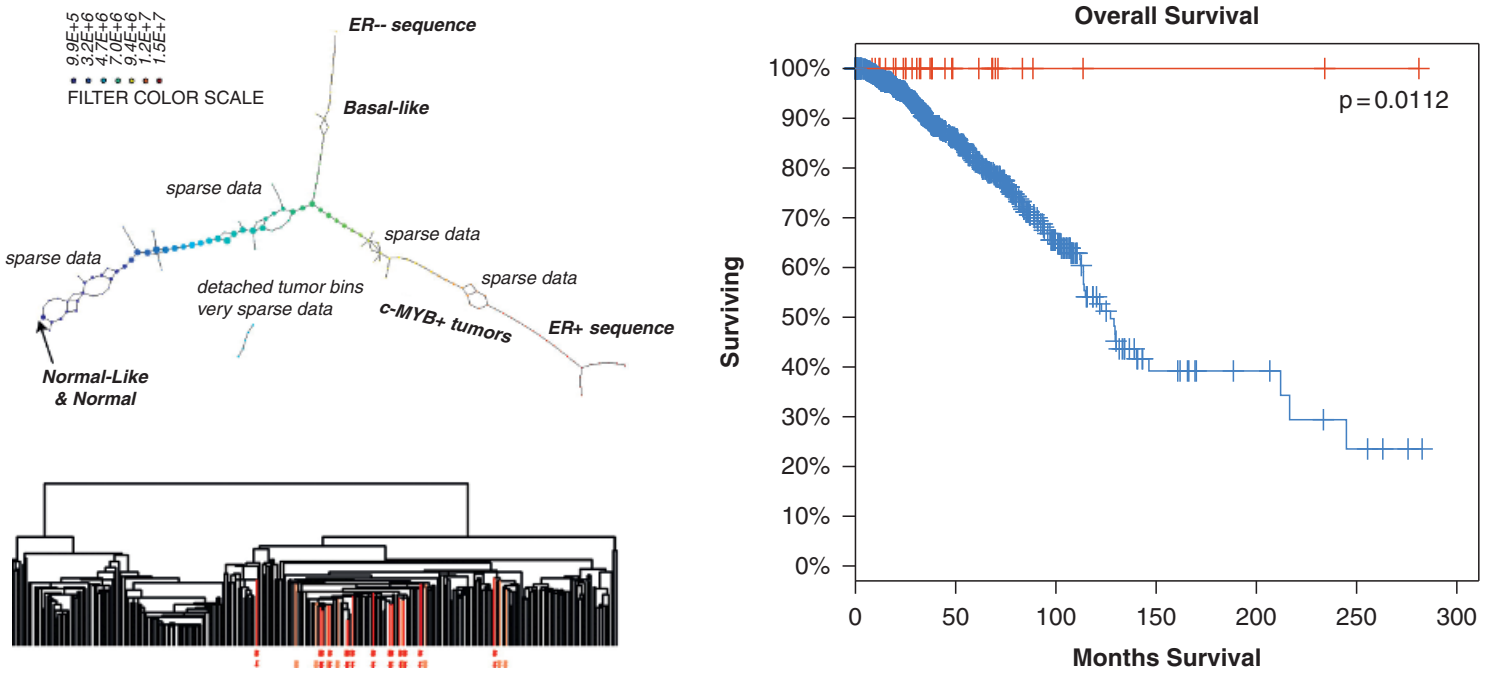


Figure 2.30 **Mapper applied to breast cancer expression data**. Top left: Mapper representation of the gene expression data from 295 breast tumors. Blue color indicates samples similar to normal tissue. Tumors with expression profiles that deviate significantly from those of normal tissue appear in the two arms on the right side. The upper arm is characterized by low expression of estrogen receptor (ER−). The lower right branch contains samples with high expression of c-MYB+ and cannot be identified using standard clustering techniques, as indicated on the lower left. Independent validation using 960 breast invasive carcinomas from "The Cancer Genome Atlas" of two of the highest expressed genes in c-MYB+ tumors, DNALI1 and C9ORF116, show very good prognosis for these tumors. Source: [383].

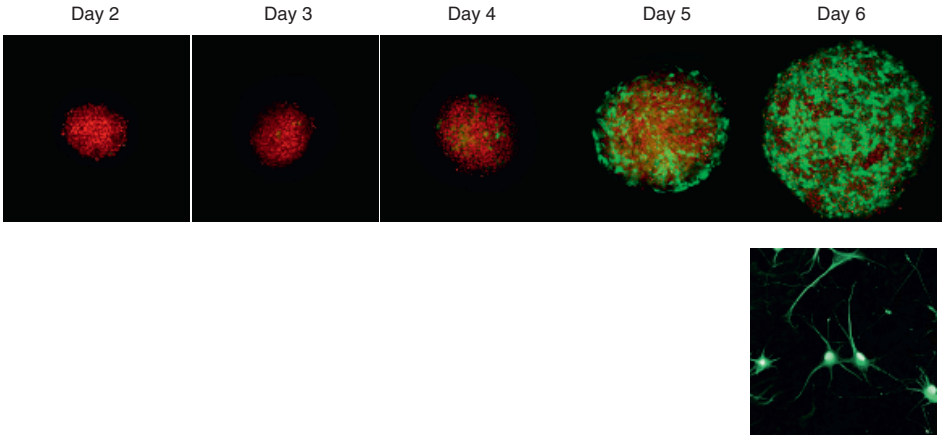


Figure 2.31 Over time, embryonic stem cells differentiate into distinct cell types. These pictures capture the in vitro differentiation of mouse embryonic stem cells into motor neurons over the course of a week. Embryonic stem cells are marked in red, and fully differentiated neurons in green. Figure from experiment performed by Elena Kandror, Abbas Rizvi and Tom Maniatis at Columbia University.

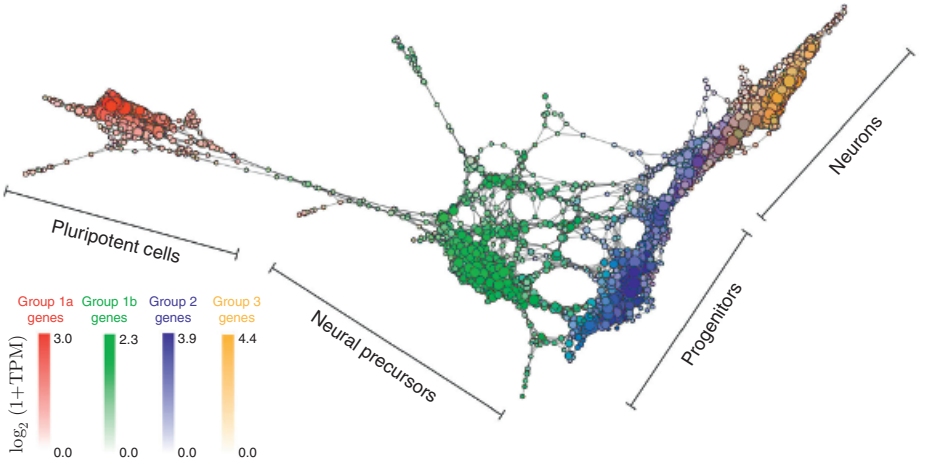


Figure 2.32 The different regions in the Mapper graph nicely line up with different points along the differentiation timeline. Source: [431].

Effectively, the issue is that a mismatch between the scale of change in the data and the width of the overlap of inverse images can give rise to dramatic changes in the Mapper graph in response to small shifts in filter function or cover. (See Figure 2.33 for a representative example of this phenomenon.)

There are various different approaches to handling this instability in practice.

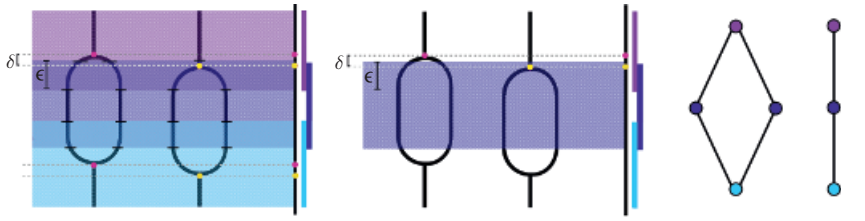


Figure 2.33 Small perturbation of the data relative to the cover can lead to large changes in the Mapper graph.

1. As we discuss in Section 3.9 below, various approaches motivated by standard considerations in statistics give us tools to establish confidence in the robustness of Mapper output.
2. Another possibility is to reintroduce persistence in the cover direction: the idea is to consider a tower of successive refinements of covers. With a suitable metric on such towers of covers, one can prove a stability theorem in this context [143].

The notion of refinement of covers also gives rise to a way to make precise the connection between Mapper and the Reeb graph. Specifically, consider the sequence of covers $\mathcal{C}_\epsilon$ consisting of all intervals of size ϵ. Then as $\epsilon \to 0$, the resulting Mapper graph converges to the Reeb graph [366].

2.9 Towards Persistent Algebraic Topology

In this chapter, we have focused primarily on ways of associating homological invariants to data sets; our focus reflects the majority of existing work on topological data analysis. From a pragmatic perspective, this choice of emphasis is very natural. Homology groups are distinguished in part by being computable; as we have seen, given a topological space presented as the geometric realization of a simplicial complex, there is an efficient algorithm for computing its homology.

In contrast, computing homotopy groups is an intractable problem. Computing the homotopy groups of spheres is a basic and unsolved problem in algebraic topology. From an algorithmic standpoint, we have the following hardness results.

1. Even for a finite complex X, $\pi_1(X)$ is uncomputable in general. (This problem is equivalent to solving the "word problem" in groups, which asks for an algorithm to determine whether two expressions in a generators and relations presentation of a group are equal.)
2. For simply connected finite complexes X and fixed k, computing $\pi_k(X)$ can be done in time polynomial in the number of simplices of X [85], although the complexity is completely infeasible for realistic use.

3. If k is allowed as part of the input (i.e., not fixed at the outset), even computing the ranks of $\pi_k(X)$ is a $\#P$-complete problem [14] (and therefore likely to be exponential, provided that current beliefs about computational complexity are true).

Notwithstanding, one can define and study persistent homotopy groups. This is an interesting endeavor for several reasons. For one thing, it is possible to use partial computations of such persistent homotopy groups to distinguish topological features of data [59]. From a theoretical perspective, consideration of persistent homotopy groups leads to efforts to understand *persistent algebraic topology*.

In classical algebraic topology, homology groups are homotopy invariants and thus capture information about the homotopy type of the space. In fact, a version of Whitehead's theorem (Theorem 1.6.31) shows that a map $f\colon X \to Y$ between simply connected CW complexes that is an isomorphism on homology groups is a homotopy equivalence. There are corresponding questions about the relationship between persistent homology and some kind of persistent homotopy equivalence.

1. What is the right notion of persistent homotopy equivalence and persistent weak equivalence? Is there an analogue of the Whitehead theorem (Theorem 1.6.31)?
2. Can we axiomatically characterize persistent homology in an analogous fashion to the way we can axiomatically characterize ordinary homology?
3. How should we think about the stability theorem (Theorem 2.4.10) in these terms?

Although it is not totally clear what candidate answers for these questions might look like, the stability theorem and the importance of the metric structure on barcodes suggests that what we are seeing is the outline of some kind of "approximate algebraic topology." See [58] for the beginnings of foundations for such a theory.

2.10 Summary

- We may assign mathematical structure to a data set by viewing the points of the set as points in a suitable metric space (X, ∂_X).
- This chapter focuses on two ways to assign a simplicial complex to a finite metric space (X, ∂_X). For a given $\varepsilon > 0$, we have the Čech complex $C_\varepsilon(X, \partial_X)$ (see Definition 2.1.2) and the Vietoris-Rips complex $\mathrm{VR}_\varepsilon(X, \partial_X)$ (see Definition 2.1.6). These complexes are functorial in ϵ.
- Given a finite metric space (X, ∂_X) uniformly sampled from a compact Riemannian manifold M, the Niyogi-Smale-Weinberger Theorem (see Theorem 2.2.1) shows that it is possible to recover topological invariants of the underlying

geometric object M, provided the distance between sampled points is smaller than some feature scale.

- The feature scale of data is unknown a priori. The idea of persistent homology is to keep track of how homological features change as the scale parameter varies.
- To investigate persistence, we examine filtered systems of simplicial complexes (see Definition 2.3.3), which arise via the functoriality of $\mathrm{VR}_\varepsilon(X, \partial_X)$ in ε.
- In order to use topological invariants to describe data, we must guarantee that small perturbations in the data correspond to commensurately small changes in the resulting invariants. To measure the size of these changes in the data, we use the Gromov-Hausdorff distance (see Definition 2.4.4). To measure changes in the barcodes, we use the bottleneck distance (see Definition 2.4.8). The stability theorem for persistent homology (Theorem 2.4.10) bounds the size of changes in barcodes by the size of changes in the data.
- Zigzag persistence is the study of persistent homology considering filtrations of different shapes where the arrows have different orientations. This approach may be helpful in controlling the number of simplices, allowing efficient computation of persistent homology.
- In some cases, a single data set may give rise to multiple filtrations. For example, we might filter by both scale and density. This is the focus of multidimensional persistence.
- The Mapper algorithm is a method for multiscale clustering that has been effectively applied to identify clinically significant information in data sets that traditional clustering may miss. Mapper performs clustering at different scales, keeping track of changes in the clusters as the scale varies.

2.11 Suggestions for Further Reading

Topological data analysis is a young field, and for many aspects of it the original papers remain the best reference. However, there have been a number of excellent introductory articles, ranging from brief treatments (e.g., [193, 326, 535]) to more comprehensive (and technical) overview articles [90, 103, 156, 157]. There are also now a number of good books [111, 155, 194, 392, 550], with slightly different areas of emphasis.