3

Statistics and Topological Inference

O! it is pleasant with a heart at ease,
Just after sunset, or by moonlight skies,
To make the shifting clouds be what you please ...
Samuel Coleridge

Our central goal in this book is to explain how to use topological data analysis as a tool for scientific inference in biology. In the previous chapter, we described a strategy for assigning topological invariants to experimental data presented as a finite metric space. Moreover, we have presented theoretical justification that in ideal cases these topological invariants encode information about the shape underlying the data. But when trying to understand how to extract answers to specific scientific questions from the shape of real experimental data, many methodological questions immediately arise.

1. How confident can we be that the results of TDA applied to sampled data correctly reflect something about the underlying process generating the data?
2. How stable are the results of TDA in the face of noise and differing choices of parameters?
3. What does a particular value of a topological invariant tell us about the shape of the data?

These questions are not unique to this setting, but arise pervasively in data analysis. But the last of these questions is particularly acute in the context of topological data analysis. The geometric significance of clustering is fairly clear; the data breaks up into groups which are made up of similar points. This is not to say that it is always easy to make use of clustering for inference, but we feel like we understand the information about the shape of the data that it provides. In contrast, suppose that you compute the homology of a data set at scale 0.75 and discover that H^6 has rank 15. What then?

In this chapter, we describe answers to the three questions above using statistical techniques to analyze topological invariants computed from data. It is easy to engage in self-deception with incautious use of statistical techniques. As a consequence, our focus is on trying to understand how to sensibly and reliably use these tools to analyze data.

3.1 What Can Topological Data Analysis Tell Us?

In order to understand the use of statistics in topological data analysis, it is useful to draw a contrast with the basic approaches in classical statistics. Consider the most fundamental problem.

1. We are given a finite collection of samples $\{x_1, \ldots, x_n\} \subset \mathbb{R}$ which have been drawn independently from a Gaussian with mean μ and standard deviation σ. The probability density function of this distribution is
$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$
2. We know that the data has come from some Gaussian, and we want to estimate μ and σ.

This is a *parametric* problem; we know the answer lies in a family of unknown distributions in which each member is described by a collection of numbers. To recover the distribution, we would estimate μ and σ using the *sample mean*

$$\hat{\mu} = \frac{1}{n} \sum_i x_i$$

and the *sample variance*

$$\hat{\sigma^2} = \frac{1}{n-1} \sum_i (x_i - \hat{\mu})^2,$$

which are unbiased estimators of the mean and variance of the underlying distribution.

Deep theoretical results provide confidence in this procedure. The *law of large numbers* tells us that as n increases, the sample mean converges to μ in a suitable sense. Since $\hat{\mu}$ depends on the particular sample, it will vary, and the *central limit theorem* describes the distribution of $\hat{\mu}$; specifically, it tells us that this quantity itself has a Gaussian distribution. We can summarize the information we obtain about $\hat{\mu}$ in terms of a *confidence interval*; this is an interval $[a, b] \subset \mathbb{R}$, defined in terms of the samples, that contains the true parameter value with a specified probability. For example, the 95% confidence interval for the mean of

a Gaussian is centered around $\hat{\mu}$ and has width that depends on the standard deviation σ.

In general parametric settings, we cannot always assume that the underlying distribution is Gaussian or that we know a closed form expression for the distribution of the parameter we are estimating. As a consequence, in practice we often form confidence intervals using *the bootstrap*: this procedure estimates the distribution of the parameter by repeatedly generating samples (with replacement) from the given samples and computing the test statistic from them.

Sometimes we do not want to assume that we know a parametric family of distribution that generated the samples; this is the domain of *non-parametric statistics*. Even in these cases, the law of large numbers and central limit theorem tell us a great deal about how to estimate various summary statistics of the underlying distribution. For example, the law of large numbers tells us that the *empirical distribution* on a sample $\{x_i\}$, which assigns probability to each value proportional to its frequency, converges to the underlying distribution. For more general summary statistics, the bootstrap remains a powerful way to estimate confidence intervals in this setting. Another possibility is to try to describe the distribution using *density estimation*; for example, we could solve the optimization problem of fitting the observed samples to a mixture of Gaussians and regard this result as an approximation of the underlying distribution.

In topological data analysis, we have access to many fewer tools. As we discuss below, it is very hard to algorithmically specify the underlying geometric object, even if we assume it is a manifold, except under very restrictive hypotheses. This implies it will be hard to recover it as well. Moreover, for distributions on general metric spaces, we do not necessarily expect many of the analogues of classical statistics to hold. And writing down parametric distributions is generally difficult. As a consequence, statistical inference in topological data analysis immediately focuses on estimating distributions of summary statistics, often generated by persistent homology barcodes.

In the literature on topological data analysis, there is often an implicit (and sometimes explicit) view of *topological inference* as a process in which some sort of underlying geometric "ground truth" can be recovered. In a setup where we have access to samples which we regard as coming from a probability distribution on an underlying space, there are a number of ways of formalizing what we mean.

1. A first goal might be simply to recover the persistent homology of the underlying space (or rather, the support of the sampling distribution) from computation of persistent homology of the samples, the *empirical persistent homology*.
2. A more sophisticated version of the preceding goal would be to recover information about both the persistent homology of the underlying space and the

probability measure generating the samples. For instance, a natural way to proceed is to try to recover the persistent homology of the *level set filtration*. Given a suitable probability density ρ on $A \subseteq \mathbb{R}^n$, the super level sets

$$\Gamma_\rho(z) = \{x \in A \mid \rho(x) > z\}$$

induce a filtration as z varies.

3.1.1 Persistent Homology and Sampling

We begin by considering the first question above: can we recover the persistent homology of the underlying space from the empirical persistent homology? An initial consistency check, described in Section 3.4, is that with large enough samples we can always recover the persistent homology of the support of the probability distribution from the empirical persistent homology (Figure 3.1). The basic observation is simply that with sufficiently many samples, even regions of low probability density will be well sampled.

However, in practice we will usually not know how many samples are enough; the feature scale of the underlying object is often unknown and even when we have some estimates of the scale, experimental realities may limit the number of data points available. Thus, we need to understand the behavior of the empirical persistent homology as it converges, i.e., when we do not necessarily assume the number of samples is large. The kind of situation we might worry about is represented in Figure 3.2; an anomalous sample leads to misleading results.

Thus, we need to understand sampling variability and decide how to assemble an estimate that aggregates the empirical persistent homology from different samples; for example, we might hope to build a confidence region for the population value of the parameter. Figures 3.3 and 3.4 indicate sampling variability in persistent homology at different sample sizes. We study questions of convergence properties and confidence intervals for estimates in Section 3.5. Thinking about summaries of collections of barcodes raises interesting questions about what it means to compute the "average" barcode or to think about the variance or spread of a collection of barcodes; we will discuss these issues throughout the chapter, notably in Sections 3.3 and 3.6.

Summarizing collections of empirical barcodes is an interesting endeavor from another perspective: we might regard the probability distribution generating the samples as itself worthy of investigation, and so want to have an invariant or collection of invariants for persistent homology which explicitly encodes information about the distribution. For example, Figure 3.5 illustrates that different distributions on the same underlying space can result in very different barcodes at small sample sizes.

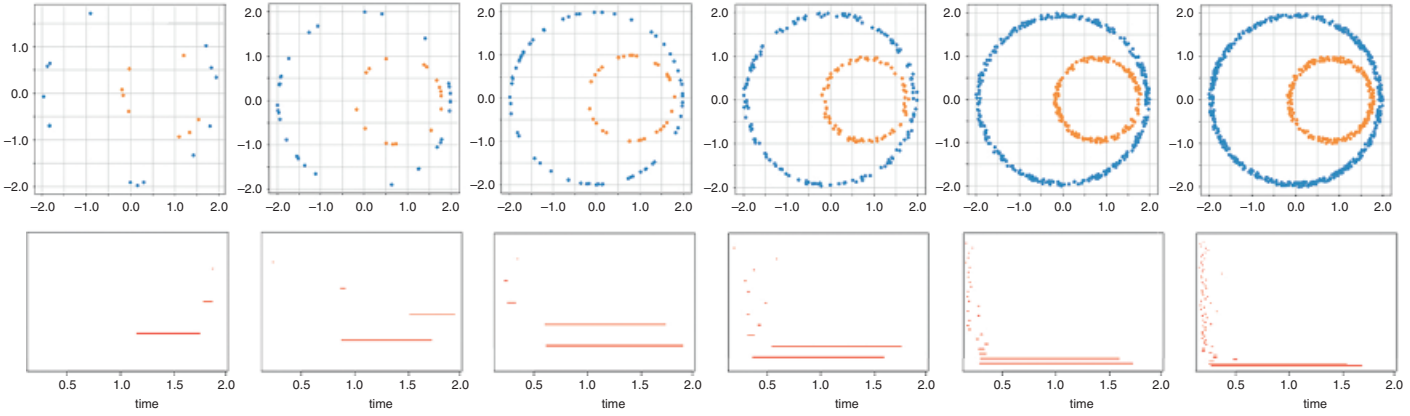

Figure 3.1 As the sample size increases, the persistent homology of the sample converges to the persistent homology of the support of the distribution, in this case the underlying space.

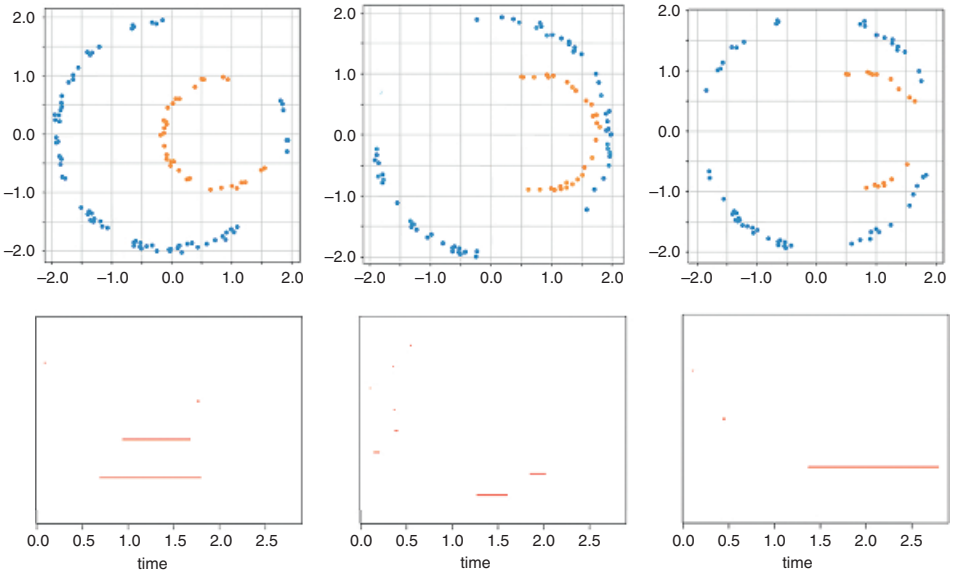

Figure 3.2 These samples were all generated from a uniform distribution on nested circles, and underneath we graph the PH_1 barcode. The barcode on the left is consistent with our expectations. But in the sample in the middle (which was a particularly anomalous sample among the many we generated), the two bars are very short and do not coexist, and on the right, there is only a single bar.

In the limiting cases where we have many samples, regions of low probability mass can make the same contribution to the topology as regions of high probability mass. And we might not regard this insensitivity to the density as a feature!

A closely related question is to understand the impact of noise in the data. One might expect the empirical persistent homology to behave well with respect to noise. After all, part of the original intuition behind persistent homology is to make homology computations robust to perturbation by integrating information across various feature scales; and this intuition is confirmed by Theorem 2.4.10, the stability theorem for persistent homology. And indeed, persistent homology is relatively stable in the face of noise concentrated around the real data; see Figure 3.6 for an example.

However, even in this case, the barcode has an increasing number of short "noise bars." The difficulties are exacerbated when we deal with data coming from a low-dimensional space embedded in a higher dimensional Euclidean space; then the noise is often the same dimension as the ambient space, which can lead to very complicated topological signals arising from the noise. These considerations motivate the study of the topology of "random" geometric complexes, which we review in Section 3.7.

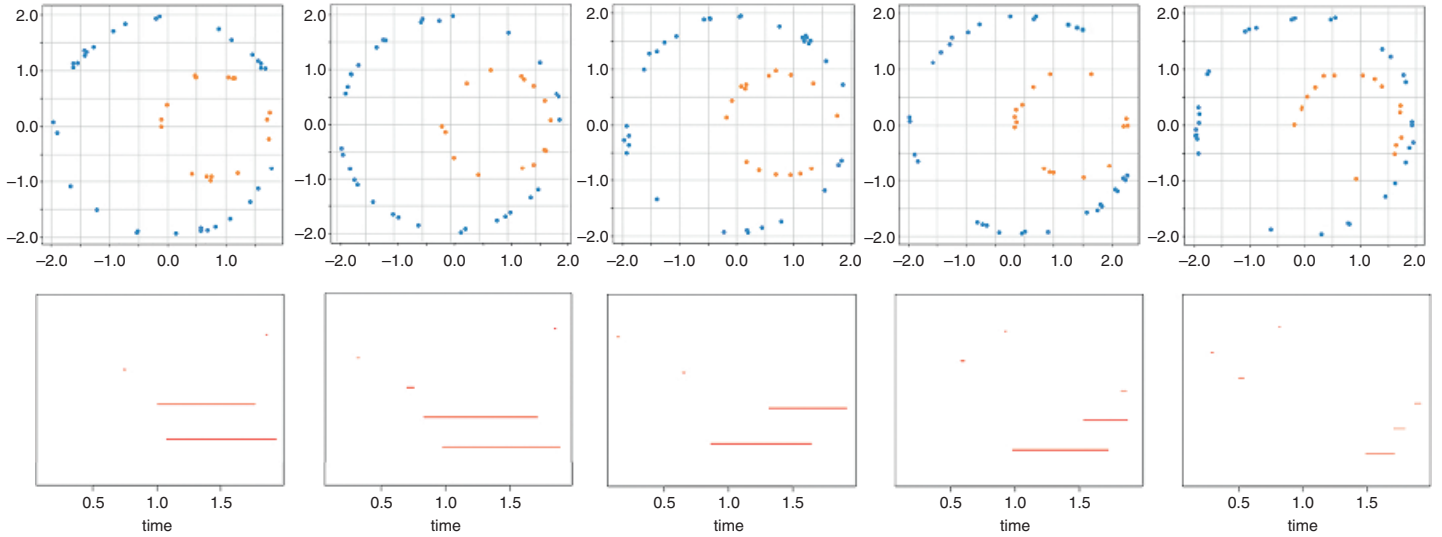

Figure 3.3 Sampling variation when the sample size is small relative to the feature scale can result in large variation in the resulting barcodes.

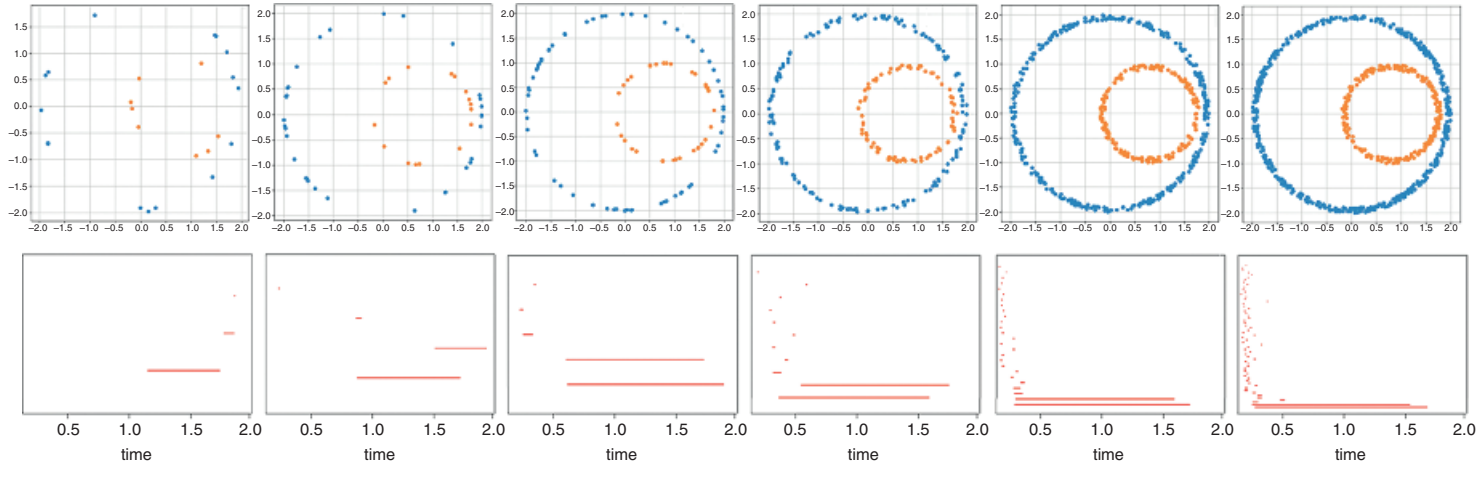

Figure 3.4 As the sample size increases, the empirical barcodes are increasingly clustered around the "true" value.

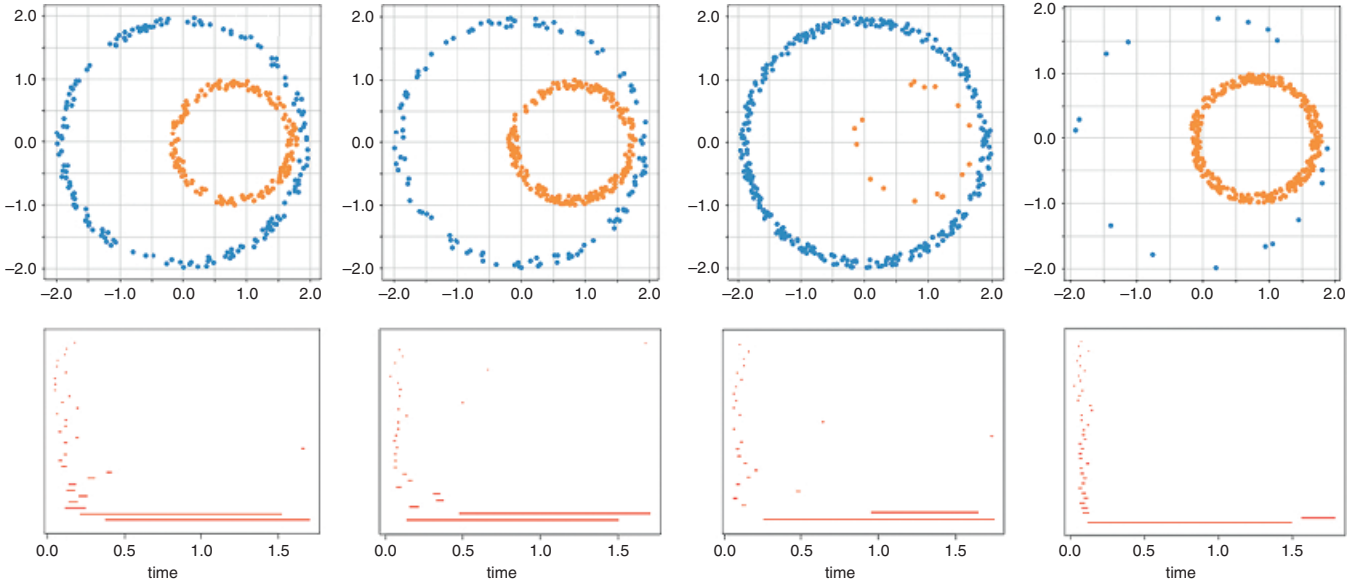

Figure 3.5 Independent identically distributed samples of fixed size from different probability distributions on the same space can result in very different barcodes.

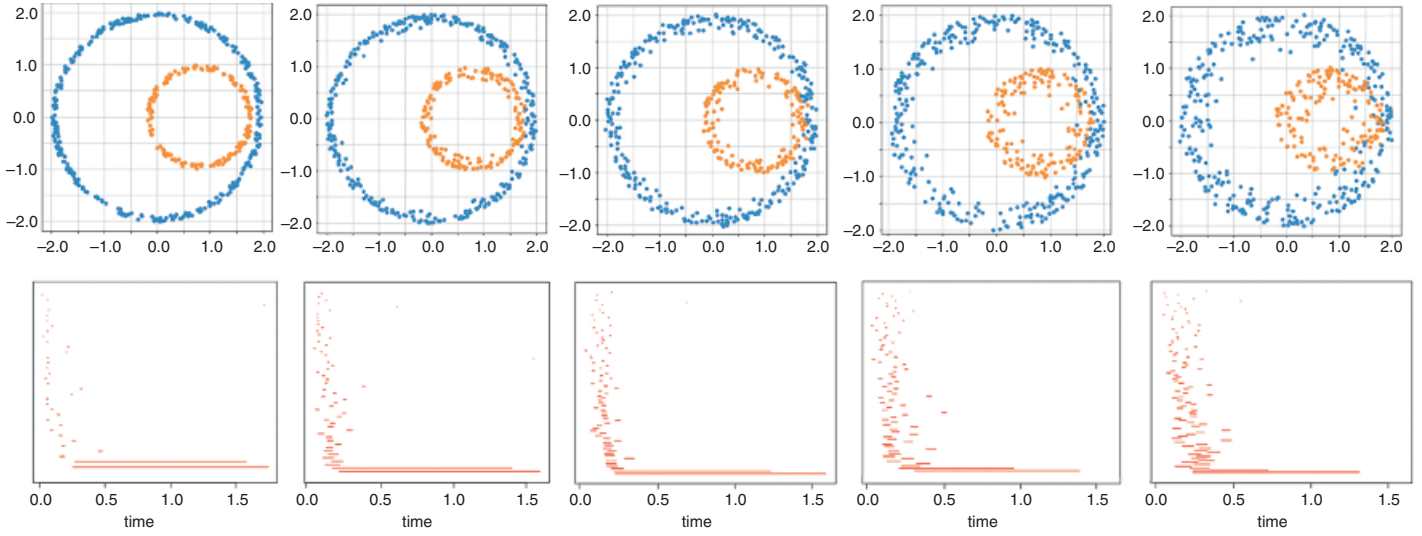

Figure 3.6 Increasing amounts of Gaussian noise centered around the underlying object cause the barcode to be filled with small spurious bars.

An even more serious problem is that not all noise is concentrated around the data. And the stability theorem has basically nothing to say about the presence of arbitrary outliers (i.e., noise points that are far from the data points). Adding a single point to a metric space (X, ∂_X) can perturb it in the Gromov-Hausdorff distance arbitrarily (recall Example 2.4.6). And we can perturb PH_i arbitrarily by adding "synthetic i-spheres" far away from the points of X. For instance, when $i = 1$, we can add 4 points at the vertices of a square with side-length k; this adds an interval $\left[\frac{k}{2}, \frac{k\sqrt{2}}{2}\right)$. Using more points, we can control the size of the interval and introduce additional intervals. (See Figure 3.7 for an example of the effect of outliers.)

This kind of instability is a well-known problem that arises even in very basic statistical inference.

Example 3.1.1. Consider computing the mean of a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}$. Specifically, let us take $\{1, 2, 3, 4, 5\}$; we find the mean is $\frac{1+2+3+4+5}{5} = 3$. Now change the point 5 to 10^{50}. To first approximation, the mean is now very close to 10^{49}. Put another way, given a set $\{x_1, \ldots, x_n\}$, one can make the mean any arbitrary value by suitably modifying a single data point! (See Figure 3.8 for a picture of this phenomenon.)

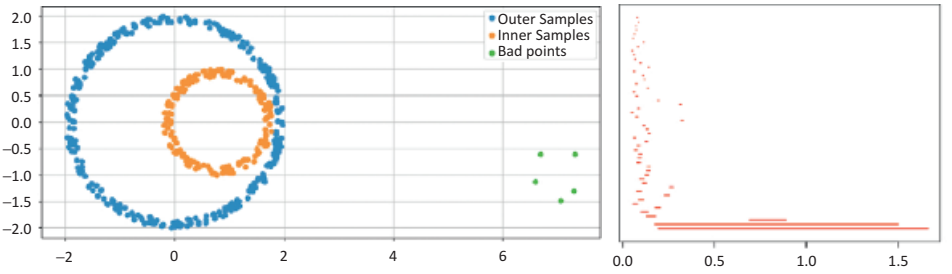

Figure 3.7 Adding a small number of points to create a circle far away from the real data can make a significant change in the barcode; a tiny number of "bad points" creates a noticeable third bar.

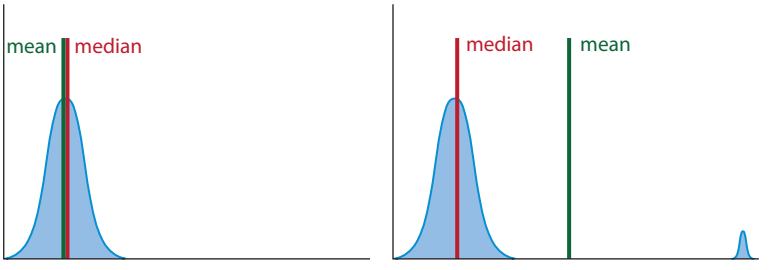

Figure 3.8 A small amount of outlying probability mass can have a large effect on the mean but cannot affect the median very much.

Traditionally, this phenomenon is the purview of *robust statistics* [251]; the mean is not robust. In contrast, the median is the classic example of a robust replacement for the mean. Changing 5 to 10^{50} does not affect the median of $\{1, 2, 3, 4, 5\}$ at all. (More generally, one needs to change at least 50% of the points in order to achieve arbitrary change in the median.) The situation with persistent homology turns out to be even worse, since whereas the mean is stable with respect to small perturbations of the distribution, barcodes of samples are not.

There are various ways to try to make persistent homology invariants more robust. One possibility is to simply preprocess the data to remove "outliers"; when there are a small number of points that are very far away from the bulk of the points, it is easy to identify them. A more principled way to do this is to consider filtering the data by density; we discuss one approach to using density estimators in Section 3.5.1, and we discuss the use of the density filtration with Mapper in Section 3.9. Another version of this strategy involves subsampling; if the number of outliers is small, most subsamples will not contain many outliers. We explain in more detail in Sections 3.4 and 3.5 how to use these ideas to bound the impact of parts of the data set with small probability mass. Finally, in Section 3.6, we discuss how to use real-valued invariants of the data and techniques from robust statistics.

3.1.2 Topological Inference

In contrast to the relative success of procedures for trying to recover information about the persistent homology of the underlying space, we cannot hope in general to identify the homotopy type of the underlying space. Any effort to identify topological spaces runs up against the fact that there is no algorithmic classification of topological spaces up to homeomorphism or homotopy type; the problem is provably uncomputable in dimensions ≥ 4, even if we restrict attention only to manifolds. (See [550, §4.1] for a nice review of these results.)

Theorem 3.1.2. *The problem of determining whether two manifolds M and N of dimension ≥ 4 presented as finite simplicial complexes are homeomorphic is undecidable.*

This result is proved by constructing a manifold whose fundamental group $\pi_1(M)$ encodes the word problem (recall Example 1.6.22 and Remark 1.6.23). Weakening homeomorphism to homotopy equivalence does not help.

Corollary 3.1.3. *The problems of determining whether two manifolds M and N of dimension ≥ 4 presented as finite simplicial complexes are homotopy equivalent or weakly homotopy equivalent are undecidable. (Even the problems of determining*

whether a given manifold is homeomorphic or homotopy equivalent to a fixed manifold Z are undecidable.)

Worse, as the allowable diameter grows, there are exponentially many possible homeomorphism types of manifolds arising as submanifolds of Euclidean space of dimension greater than 2 [533, §1.2]. Similar bounds hold for the number of possible homotopy types. As a consequence, it is not in general plausible to parametrize hypothesis classes of spaces except when imposing strong restrictions or using coarse invariants. Moreover, the explicit sample bounds for recovery of persistent homology (from Section 2.2 above and Section 3.5 below) are exponential in the intrinsic dimension of the data.

Even if we restrict ourselves to the seemingly easy problem of distinguishing spheres of different dimensions (i.e., S^{50} versus S^{51}), basic results about *concentration of measure* in high-dimensional Euclidean spaces imply that under an oblivious sampling model (i.e., when samples are drawn independently of one another) most of the mass on a sphere S^n is concentrated around a radial region which is homeomorphic to S^{n-1}. This shows that this problem requires an exponentially large number of points [533, §1.3]. More generally, as we discuss below in Section 4.6, it is very difficult to successfully estimate the dimension of very high-dimensional manifolds.

These constraints place sharp limits on the kind of geometric inference that we can expect. We have basically three options: work with low-dimensional topological features of the data and perform exploratory data analysis, work with low-dimensional data where exact topological inference is reasonable, or treat the results of topological data analysis as signals about shape that are potentially uninterpretable except as input to statistical inference or machine learning procedures. In more detail, TDA provides the following.

1. A methodology for exploratory data analysis via description and visualization of low-dimensional shape information. Arguably the most widely used TDA technique is Mapper, and indeed the standard usage pattern for Mapper is to search for meaningful clusters in the data which can then guide further experiments. We discuss this at a high level in Section 3.9. In the second part of the book, we will explain many examples of this approach, including applications to tumor classification and cell differentiation (see Sections 6.7, 7.3, and 7.4).
2. Exact information about data that truly does lie in low-dimensional topological spaces. In these cases, topological data analysis can be interpreted to provide specific geometric information about the data and is often applied in a "hypothesis testing" framework. For instance, in dimension 1, specific hypotheses about the process generating the data are reasonable, and analogues of parametric

statistics make sense. We will discuss an example of this kind of approach in phylogenetics in Section 5.2, where persistent H_1 is used to detect divergence from the "tree hypothesis" for evolutionary data and estimate recombination rates.

3. Robust "topological signals" to use as features for classification, inference, and supervised learning algorithms. Although many topological features cannot be interpreted directly (e.g., "$H^{15}(X)$ is approximately 39"), they still convey discriminative information about the data. Ideally, this approach permits integration of information from topological data analysis with other sources of information (e.g., standard parametric statistical models). Two examples of this approach that we will discuss are surface recognition via the persistent homology transform (see Section 3.8 for a general discussion and Section 9.3 for specific applications) and the use of persistent homology information to fit parameters for population genetics models (see Section 5.7).

We now explain how to integrate topological data analysis with suitable statistical techniques in order to carry out these three kinds of analyses.

3.2 Background: Geometric Sampling and Metric Measure Spaces

At the most basic level we access geometry through a metric. Therefore, we want to work with metric spaces equipped with probability measures that are compatible with the metric. We do this using the machinery of *metric measure spaces*. This framework makes it possible to extend intuitive and familiar ideas from ordinary statistics in Euclidean space to a very broad class of geometric objects.

3.2.1 Metric Measure Spaces

To express the compatibility of metric and probability measure in a precise fashion, we work with the notions of measurable spaces and measures. A *measurable space* is a set along with a collection of subsets to which we can assign "area." A *measure* on a measurable space is a rule for assigning area, i.e., a theory of integration. We rapidly review these definitions here; we recommend [56, 57] for more in-depth treatments.

Definition 3.2.1. For a set X, a *σ-algebra* is a collection Σ of subsets of X such that

1. $\emptyset \in \Sigma$,
2. given a countable set $\{U_i\}$ such that $U_i \in \Sigma$, then the union $\bigcup_i U_i$ is also in Σ, and
3. if $U \in \Sigma$, then the complement $X \setminus U$ is in Σ.

As we noted above, these closure properties are motivated by the perspective that the elements of a σ-algebra have area; for instance, given a collection of sets that have area, we should be able to measure the area of their union. Given an arbitrary collection S of subsets of X, the σ-algebra generated by this collection is the smallest σ-algebra containing S; roughly speaking, we simply add all missing unions, intersections, and complements.

Example 3.2.2. The most important example of a σ-algebra is the *Borel σ-algebra* associated to a topological space X; this is just the σ-algebra generated by the collection of open sets of X. (Equivalently, it is generated by the collection of closed sets of X.)

In fact, when (X, ∂_X) is separable (i.e., contains a countable dense subset; recall Definition 1.2.14), then the Borel σ-algebra is generated by the collection of open balls $\{B_\epsilon(x)\}$ as ϵ varies over $\mathbb{R}^{>0}$ and x over the points of X.

Definition 3.2.3. A *measurable space* is a pair (X, Σ) consisting of a set X and a σ-algebra Σ.

Example 3.2.4.

1. Let X be a countable set; the power set of X forms a σ-algebra, which we refer to as the *counting σ-algebra*. This σ-algebra is generated by the points $x \in X$.
2. Euclidean space $\mathbb{R}^n$ with the σ-algebra generated by the boxes $(a_1, b_1) \times \cdots \times (a_n, b_n)$ is a measurable space.
3. More generally, any topological space is a measurable space with the Borel σ-algebra.
4. It turns out to be technically advantageous to equip Euclidean space with a more sophisticated σ-algebra, the Lebesgue σ-algebra. This is an enlargement of the Borel σ-algebra; it includes more measurable sets, in order to force every subset of a set of measure zero to be measurable. (This enlargement is referred to as the "completion" of a σ-algebra.)

Functions between measurable spaces are defined in analogy with continuous functions.

Definition 3.2.5. Let (X, Σ) and (X', Σ') be measurable spaces. A map of sets $f\colon X \to Y$ is a *measurable function* if $f^{-1}(A) \in \Sigma$ for all $A \in \Sigma'$. A measurable function is a *measurable isomorphism* when f is an isomorphism of sets and f^{-1} is also a measurable function.

Measurable spaces support the computation of area; a *measure space* is a measurable space that has been equipped with a specific area function, a *measure*.

Definition 3.2.6. A *measure* μ on a measurable space (X, Σ) is a function

$$\mu \colon \Sigma \to \mathbb{R}^{\geq 0}$$

such that

1. $\mu(\emptyset) = 0$, and
2. for $X_i \in \Sigma$ such that $X_i \cap X_j = \emptyset$ for all i and j,

$$\mu\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} \mu(X_i).$$

A basic theorem that allows us to construct measures is that for a σ-algebra generated by a collection of subsets S, it suffices to define the measure on the sets in S. This result is closely related to the construction of the Riemann integral. (See Figure 3.9 for an example of the process.)

Example 3.2.7.

1. For a finite set X with the counting σ-algebra, the *counting measure* on X assigns to each subset $A \subseteq X$ the cardinality of A, i.e.

$$\mu(A) = \#A, \qquad A \subseteq X.$$

   This can be regarded as the measure determined by setting each point $x \in D$ to have measure 1.
2. For $\mathbb{R}^n$ with the box σ-algebra, the standard measure is determined by assigning to each rectangle its area, i.e.,

$$\mu([a_1, b_1] \times \ldots \times [a_n, b_n]) = \prod_i (b_i - a_i).$$

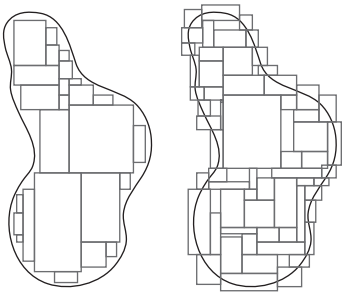

Figure 3.9 In favorable cases, the measure of an arbitrary region is bounded by the area of inner and outer covers by generating sets; the actual area is given by taking limits as the inside sum increases and the outside sum decreases. In general, the inner measure is defined in terms of the outer measure of the complement.

Given a measure μ on X, we can integrate any measurable function $f\colon X \to \mathbb{R}$ over a region $A \subseteq X$ as follows. Assuming temporarily that $f \geq 0$, we set

$$\int_A f d\mu = \sup_{\substack{B_1 \cup B_2 \cup \ldots B_\ell = A \\ B_i \cap B_j = \emptyset, i \neq j}} \sum_i \left(\inf_{x \in B_i} f(x) \right) \mu(B_i).$$

Here the sup is computed over all decompositions of A into finitely many disjoint subsets $\{B_i\}$ (in particular, ℓ will vary). If f takes both positive and negative values, we define the integral in terms of the expression above for the positive part and negative part separately and take the sum.

We are most interested in *probability measures*, for which we require that $\mu(X) = 1$. An important class of examples of probability measures are determined by *probability density functions*. Given a probability measure μ and a measurable function f, the integral $\int_X f$ is called the *expectation* of X.

Definition 3.2.8. Let μ be a measure on (X, Σ) and f be a measurable function $f\colon X \to \mathbb{R}$ so that $\mu(\{z \mid f(z) < 0\}) = 0$. Then there is an induced measure on X

$$\nu(A) = \int_A f d\mu, \quad A \subseteq X.$$

We say that the measure ν has *density* f with respect to μ.

Remark 3.2.9. It is standard to describe measures via probability densities when working with a basic reference measure for integration, e.g., the Lebesgue measure on $\mathbb{R}^n$ or the counting measure on a finite set. In the following discussion, we will sometimes omit specification of the measure when working with densities.

When (X, ∂_X) is a metric space, we can now use the topology induced by the metric and the Borel σ-algebra to express compatibility of metric and measure. A *Borel measure* is a measure with respect to the Borel σ-algebra.

Definition 3.2.10. A *metric measure space* with a probability measure is a metric space (X, ∂_X) that is complete and separable, equipped with a Borel probability measure μ_X. The *support* of a metric measure space is the subset $\mathrm{supp}(X)$ of X consisting of points x for which every neighborhood U of x satisfies $\mu_X(U) > 0$.

Remark 3.2.11. More generally, we can consider metric measure spaces where the measure is not a probability. We will not use such examples in this chapter, however.

Definition 3.2.10 provides a theoretical framework for describing data sampled from some kind of geometric object.

> Our working hypothesis throughout this chapter is that we have data presented as samples from an underlying metric measure space (X, ∂_X, μ_X).

Example 3.2.12.

1. A finite metric space (X, ∂_X) with the normalized counting measure
$$\mu(A) = \frac{\#A}{\#X}, \qquad A \subseteq X$$
is a metric measure space.
2. For any subset $A \in \mathbb{R}$ and a measure μ (not necessarily a probability measure) such that $\mu(A) < \infty$, A becomes a metric measure space via the uniform measure
$$\mu'(S) = \frac{\mu(S)}{\mu(A)}, \qquad S \subseteq A.$$
3. More generally, the standard probability distributions on $\mathbb{R}$ and $\mathbb{R}^n$ equip them with the structure of metric measure spaces. For example, $\mathbb{R}$ with a Gaussian density gives rise to the Gaussian measure when integrated with regard to the Lebesgue measure.
4. Manifolds also provide natural geometric examples of metric measure spaces – any compact Riemannian manifold M is a metric measure space under the volume measure [144]. Samples from the volume measure on a manifold have the property that any small region has a number of points proportional to its volume; this is a version of the uniform distribution.
5. Given any metric measure space (X, ∂_X, μ_X), any measurable subset $A \subset X$ is itself a metric measure space, where
$$\mu_A(V) = \frac{\mu_X(V)}{\mu_X(A)}$$
for $V \subset A$.

We can describe finite independent identically distributed (i.i.d.) samples as follows.

Definition 3.2.13. Let (X, ∂_X, μ_X) be a metric measure space. The *product measure* $\mu_X^{\otimes n}$ makes the metric space $(\prod_{i=1}^n X, \prod_{i=1}^n \partial_X)$ into a metric measure space, where

$$\mu_X^{\otimes n}(A_1 \times A_2 \times \ldots \times A_n) = \mu_X(A_1)\mu_X(A_2)\ldots\mu_X(A_n),$$
$$A_1 \times A_2 \times \ldots \times A_n \subseteq \prod_{i=1}^{n} X = X \times X \times \ldots \times X.$$

Thus, an i.i.d. sample of size n from (X, ∂_X, μ_X) can be described as a draw from the distribution $\mu_X^{\otimes_n}$.

We will be interested in measures induced by the application of functions (e.g., persistent homology). To be precise about this, we need the notion of the *pushforward* of a measure.

Definition 3.2.14. Let $f\colon (X, \partial_X) \to (Y, \partial_Y)$ be a measurable function between the Borel measure spaces X and Y. Then given a probability measure μ_X, the *pushforward measure* $f_*\mu_X$ on Y is specified by the formula

$$f_*\mu_X(A) = \mu_X(f^{-1}(A)),$$

for A a measurable set in Y.

Another useful way of generating new measures is by combining old ones.

Definition 3.2.15. Let μ and ν be finite Borel measures on $\mathbb{R}^n$. Then the convolution $\mu * \nu$ can be defined as

$$\mu * \nu = +_*(\mu \times \nu),$$

the pushforward of the product measure along the addition map $+\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$.

Explicitly, the convolution is given by the formula

$$(\mu * \nu)(A) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \mathbf{1}_A(x + y) d\mu(x) d\nu(y),$$

where $\mathbf{1}_A$ is the indicator function for the measurable set $A \subset \mathbb{R}^n$. Convolution with a Gaussian affords a useful general technique for smoothing distributions with complicated local structure; the width of the Gaussian controls the degree of smoothing.

Finally, we note that it is frequently useful to have a notion of size for real-valued functions on a metric measure space. To this end, we quickly recall the definition of the L_p and L_∞ norms.

Definition 3.2.16. Let (X, ∂_X, μ_X) be a metric measure space and let $f\colon X \to \mathbb{R}$ be a measurable function such that $\int_X f^p d\mu < \infty$. Then the L_p *norm* of f for $1 \leq p < \infty$ is

$$\|f\|_p = \left(\int_X |f|^p d\mu \right)^{\frac{1}{p}}.$$

When $p = \infty$, we define

$$\|f\|_\infty = \inf\{k \in \mathbb{R} \,|\, \mu(\{x \in X \,|\, f(x) > k\}) = 0\}.$$

Remark 3.2.17. When X is a finite set and μ_X is the *counting measure* (i.e., the measure that assigns probability mass $\frac{1}{|X|}$ to each point), these norms reduce to the pth root of the sum of pth powers and the max, respectively.

Remark 3.2.18. Geometric sampling on non-Euclidean metric measure spaces can be very subtle, even when dealing with the volume measure on a compact Riemannian manifold [144]. For example, consider the problem of sampling from the surface of the sphere $S^2 \subseteq \mathbb{R}^3$. In this case, there is a natural parametrization of the points of the sphere arising from spherical coordinates (r, θ_1, θ_2). A naive approach is to use the spherical coordinates to sample: sample uniformly θ_1 and θ_2 from $[0, 2\pi]$ and $[0, \pi]$ respectively and consider the map $\sigma\colon [0, 2\pi] \times [0, \pi] \to \mathbb{R}$ specified by

$$\begin{aligned} x &= \sin(\theta_2)\cos(\theta_1) \\ y &= \sin(\theta_2)\sin(\theta_1) \\ z &= \cos(\theta_2). \end{aligned}$$

Denoting by U the uniform distribution on $[0, 2\pi] \times [0, \pi]$, we have the pushforward $\sigma_* U$ which is supported on $S^2 \subseteq \mathbb{R}^3$. However, $\sigma_* U$ is concentrated at the poles and is not the distribution associated to the area form. In this case, we can simply sample uniformly in a cube around the origin in $\mathbb{R}^3 \setminus \{0\}$, discard points further than 1 from the origin, and divide by the norm. More generally, one needs to use either rejection sampling or Markov chain Monte Carlo (MCMC) techniques. These methods can be applied to general manifolds, provided one has access to an explicit and computationally tractable parameterization; of course, this is often a serious problem.

3.2.2 The Fréchet Mean and Variance of a Metric Measure Space

Most practical applications of statistics involve the use of summary statistics. As such, it is natural to look for notions of mean and variance that apply in the general context of metric measure spaces. The standard approach to this problem is the theory of the Fréchet mean and variance of a probability measure μ on a metric measure space (e.g., see [487] for an introduction to this theory). Although it

turns out that this theory is not particularly useful in barcode space (as we explain below), we nonetheless quickly review it here since understanding the pathological behavior of the Fréchet mean motivates the techniques used in practice. We restrict our attention to probability measures μ satisfying a finiteness condition.

Definition 3.2.19. Let (X, ∂_X, μ_X) be a metric measure space. Then the *Fréchet variance* as a function of $z \in X$ is the integral

$$v_\mu(z) = \int_X \partial_X(z, x)^2 d\mu(x).$$

We will assume that $v_\mu < \infty$. Then the *Fréchet mean* is defined as follows.

Definition 3.2.20. The *Fréchet mean* is the set

$$e_\mu = \operatorname{argmin}\left(\inf_z v_\mu(z)\right) \subseteq X,$$

i.e., the values $z \in X$ that achieve the infimum.

When dealing with a finite sample $\{x_1, x_2, \ldots, x_n\}$ from (X, ∂_X, μ_X), the Fréchet mean and variance of the underlying distribution are approximated using the empirical measure which assigns probability $\frac{1}{n}$ to each point in the sample. See Figure 3.10 for a simple example.

It is not at all clear that the Fréchet mean exists for general metric measure spaces; in practice, we rely on the following result.

Theorem 3.2.21. *Let (X, ∂_X, μ_X) be a metric measure space. If μ_X has compact support (e.g., if X is compact), then the Fréchet mean exists.*

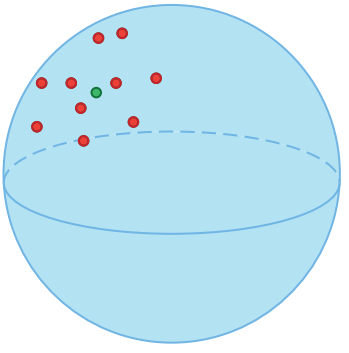

Figure 3.10 The Fréchet mean (green) of a finite sample (red) from the uniform distribution on a sphere is the point on the sphere that is the "centroid" of the sample.

More generally, the Fréchet mean can be shown to exist as long as the "tails" of μ decay sufficiently rapidly. (See [285] for a precise statement.)

The general theory of the Fréchet mean and variance provides laws of large numbers; given finite samples from μ equipped with the empirical measure, the Fréchet means of the samples converge to the Fréchet mean of μ.

Theorem 3.2.22. *Let (X, ∂_X, μ) be a metric measure space. Let $\{Z_k\}$ be a collection of i.i.d. samples $Z_k \subset X$ drawn according to μ, such that $|Z_k| \to \infty$ as $k \to \infty$. Let μ_k denote the empirical measure on Z_k. Then almost surely $e_{\mu_k} \to e_\mu$ (i.e., the probability of convergence is 1).*

The problem of understanding the convergence of derived quantities of distributions for increasing finite samples suggests that we should put a topology on the set of probability measures. We now turn to a discussion of how to construct metrics on probability distributions and on metric measure spaces.

3.2.3 Distances on Measures and Metric Measure Spaces

In order to state Theorem 2.4.10, the stability theorem for persistent homology of finite metric spaces, we used a metric on the set of isometry classes of finite metric spaces. To state the analogous stability theory describing the interaction of sampling and persistent homology, we will use a metric on the set of isomorphism classes of compact metric measure spaces. Recall that the Gromov-Hausdorff metric is defined in terms of a metric on subspaces of a fixed metric space, the Hausdorff metric. To define a metric on metric measure spaces, we will start with a metric on probability measures on a fixed metric space.

To motivate this definition, we quickly explain the notion of *weak convergence* of probability measures. For a metric space (X, ∂_X), let $\mathcal{P}(X)$ denote the set of Borel probability measures on X.

Definition 3.2.23. Let (X, ∂_X) be a metric space. A sequence $\{\mu_n\} \subset \mathcal{P}(X)$ *weakly converges* to $\mu \in \mathcal{P}(X)$ if for all bounded continuous functions $f\colon X \to \mathbb{R}$,

$$\int_X f d\mu_n \to \int_X f d\mu.$$

The idea of weak convergence is that a sequence of distributions converges when the average value of any function f converges; i.e., weak convergence means that the expectation of any random variable converges. This notion of convergence is of particular importance because it is the kind of convergence that arises in the central limit theorem.

Warning 3.2.24. Weak convergence is very different from requiring that the measure of each set converge!

Since convergence of sequences can be defined in terms of a metric (recall Definition 1.2.7), it is natural to look for a metric that controls weak convergence. We now introduce several such metrics that are useful in topological data analysis, starting with the *Prohorov distance*.

Definition 3.2.25. Let (X, ∂_X) be a metric space equipped with two Borel measures μ_1 and μ_2. Then we define the *Prohorov distance* between μ_1 and μ_2 to be

$$d_{Pr}(\mu_1, \mu_2) = \inf\{\epsilon > 0 \mid \mu_1(A) \leq \mu_2(B_\epsilon(A)) + \epsilon \quad \text{and} \quad \mu_2(A) \leq \mu_1(B_\epsilon(A)) + \epsilon\},$$

where A varies over all closed sets in X and

$$B_\epsilon(A) = \{z \in X \mid \exists\, a \in A, \partial_X(z, a) \leq \epsilon\}.$$

To understand what the Prohorov distance means, it can be convenient to use an alternative formulation. For this, we need the notion of a *coupling*, which is a probability distribution θ on $X \times X$ such that $\theta(A \times X) = \mu_1(A)$ and $\theta(X \times B) = \mu_2(B)$ for arbitrary measurable subsets $A, B \subseteq X$.

Lemma 3.2.26. *Let (X, ∂_X) be a metric space equipped with two Borel measures μ_1 and μ_2. Then we can compute the Prohorov distance as*

$$d_{Pr}(\mu_1, \mu_2) = \inf_C \inf\{\epsilon > 0 \mid C\{(x, x') \in X \times X \mid \partial(x, x') \geq \epsilon\} < \epsilon\},$$

where C varies over all couplings.

Roughly speaking, two measures are within ϵ in the Prohorov metric when there is a matching of the space with itself such that on a region of probability mass $1 - \epsilon$ matched points are within ϵ and can vary arbitrarily on the remainder.

Proposition 3.2.27. *The distance d_{Pr} is a metric on $\mathcal{P}(X, \partial_X)$, the space of probability measures on X. If X is complete and separable, then given a sequence of probability measures $\{\mu_i\}$ that converges to a measure μ in d_{Pr}, μ_i weakly converges to μ.*

A complete and separable metric space is called a *Polish space*. In general, Polish spaces are a good setting for probability theory: not only is weak convergence metrizable, but in addition certain pathologies with product measures do not arise.

Example 3.2.28.

1. Let μ_1 and μ_2 be distributions determined by δ-functions, i.e., μ_1 has mass 1 on a point x_1 and μ_2 has mass 1 on a point x_2. Then
$$d_{Pr}(\mu_1, \mu_2) = \min(\partial_X(x_1, x_2), 1).$$
2. Let (X, ∂_X, μ_X) be a metric measure space and $Y \subset X$ have measure $> 1 - \epsilon$. Then μ_X regarded as a distribution on Y has Prohorov distance $< \epsilon$ from μ_X.

In fact, there are many metrics on $\mathcal{P}(X)$ that metrize weak convergence [195]. Optimal transport theory suggests the use of the Wasserstein or "earth-mover" metric [517]. Here, the rough idea is to imagine distributions modeled by piles of dirt; the Wasserstein distance is the minimal amount of energy (dirt times distance) that must be expended to transform one distribution into another.

Definition 3.2.29. Let (X, ∂_X) be a compact metric space equipped with two Borel measures μ_1 and μ_2. For $p \geq 1$, the *p-Wasserstein distance* between μ_1 and μ_2 is
$$d_{W_p} = \left(\inf_C \int_{X \times X} \partial_X(x, y)^p \mathrm{d}C(x, y) \right)^{\frac{1}{p}},$$
where C varies over all couplings.

Any of the Wasserstein distances metrize weak convergence of probability measures on metric spaces with bounded diameter (i.e., where the maximum distance between $x_1, x_2 \in X$ is bounded).

Lemma 3.2.30. *The distance d_{W_p} is a metric on $\mathcal{P}(X, \partial_X)$, the space of probability measures on X. Let (X, ∂_X) have bounded diameter. Then given a sequence of probability measures $\{\mu_i\}$ that converges to a measure μ in d_{W_p}, then μ_i weakly converges to μ.*

Example 3.2.31.

1. Let μ_1 and μ_2 be distributions specified by δ-functions; μ_1 has mass 1 on $x_1 \in X$ and μ_2 has mass 1 on $x_2 \in X$. Then $d_{W_p}(x_1, x_2) = \partial_X(x_1, x_2)$.
2. Let μ_1 and μ_2 be empirical distributions on finite subsets $\{x_i\} \subset X$ and $\{x_i'\} \subset X$ such that $|\{x_i\}| = |\{x_i'\}|$. Then the Wasserstein distance can be computed as
$$d_{W_p} = \min_{\theta \colon \{x_i\} \to \{x_i'\}} \left(\sum_i (\partial_X(x_i, \theta(x_i)))^p \right)^{\frac{1}{p}},$$
where θ varies over all bijections.

Remark 3.2.32. It is common in information theory and Bayesian statistics to measure the difference between distributions μ_1 and μ_2 in terms of the Kullback-Leibler divergence. Taking p and q to be probability mass functions on a discrete space X where $q(x) = 0 \implies p(x) = 0$, the Kullback-Leibler divergence is computed as

$$\sum_{x \in X} p(x) \log \frac{p(x)}{q(x)},$$

where we interpret the contribution of a term with $p(x) = 0$ to be 0. (An analogous definition can be given in the setting of measure spaces, but setting it up is sufficiently complicated that we do not pursue it here; see [195] for a discussion, where it is referred to as relative entropy.)

The Kullback-Leibler divergence has many interesting properties, but it is not a metric; it is neither symmetric nor satisfies the triangle inequality.

The Wasserstein distance and the Prohorov distance are related, in the sense that

$$d_P(\mu_1, \mu_2)^2 \leq d_{W_1}(\mu_1, \mu_2) \leq (\operatorname{diam}(X) + 1) d_P(\mu_1, \mu_2)$$

(and $d_{W_1}(\mu_1, \mu_2) \leq d_{W_p}(\mu_1, \mu_2) \leq C d_{W_1}(\mu_1, \mu_2)$ for a suitable constant C) [195]. We can convert the Prohorov and Wasserstein distances into metrics on isomorphism classes of compact metric measure spaces. The approach is to use an analogue of the technique that converts the Hausdorff distance into the Gromov-Hausdorff metric on isometry classes of compact metric spaces.

Definition 3.2.33. Let (X, ∂_X, μ_X) and (Y, ∂_Y, μ_Y) be compact metric measure spaces. The *Gromov-Prohorov distance* is defined as

$$d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) = \inf_{\phi_X, \phi_Y, Z} d_{Pr}\left((\phi_X)_* \mu_X, (\phi_Y)_* \mu_Y\right),$$

where here $\phi_X \colon X \to Z$ and $\phi_Y \colon Y \to Z$ are isometric embeddings into a metric space Z.

Definition 3.2.34. Let (X, ∂_X, μ_X) and (Y, ∂_Y, μ_Y) be compact metric measure spaces. The *Gromov-Wasserstein distance* is defined as

$$d_{GW_p}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) = \inf_{\phi_X, \phi_Y, Z} d_{W_p}((\phi_X)_* \mu_X, (\phi_Y)_* \mu_Y),$$

where (ϕ_X, ϕ_Y, Z) is as in the previous definition.

Lemma 3.2.35. *The Gromov-Prohorov and Gromov-Wasserstein distances are metrics on the set of isomorphism classes of compact metric measure spaces.*

Remark 3.2.36. Although we do not review this here, there is very interesting work on the details of the topology induced on the set of isomorphism classes of metric measure spaces by these metrics [207, 346, 347].

3.3 Probability Theory in Barcode Space

The foundation of any statistical approach to persistent homology is the notion of a probability distribution of barcodes. The set $\mathcal{B}$ of barcodes is a metric space under the bottleneck distance d_B (Definition 2.4.8) or the p-Wasserstein distance d_{W_p} (Definition 2.4.9). Therefore, $\mathcal{B}$ endowed with the Borel σ-algebra becomes a measurable space: we can work with the collection of Borel probability measures on $\mathcal{B}$. Proposition 3.2.27 shows that the Prohorov metric on the set of Borel probability measures metrizes weak convergence of probability measures when the underlying metric space is complete and separable. We begin this section by constructing subspaces of barcode space that are complete and separable.

3.3.1 Polish Spaces of Barcodes

A first thought is to consider the set of finite barcodes. It is easy to see that this barcode space is separable for either the bottleneck or Wasserstein distance; an arbitrary "bar" $[a, b)$, with $a, b \in \mathbb{R}$, can be approximated arbitrarily well by choosing rational approximations a' for a and b' for b. However, the set of finite barcodes is not complete.

Example 3.3.1. Consider a sequence of barcodes $\{X_i\}$ where $X_0 = \emptyset$ and X_i is obtained from X_{i-1} by adding a disjoint bar $[0, \frac{1}{n})$. That is,

$$X_i = \{[0, 1), [0, 1/2), [0, 1/3), \ldots, [0, 1/i)\}.$$

Working with the bottleneck distance, it is easy to check that $\{X_i\}$ is a Cauchy sequence (recall Definition 1.2.9),

$$d_B(X_i, X_j) \leq \frac{1}{\max(i, j)},$$

as the distance between X_i and X_j is bounded by the longest bar present in X_j and not in X_i (assuming that $j > i$). But $\{X_i\}$ does not converge to any element of $\mathcal{B}$; the sequence is clearly converging to a barcode with infinitely many bars! (See Figure 3.11 for a picture of this sequence.)

Instead, we can consider countable barcodes, although certain finiteness conditions are still required.

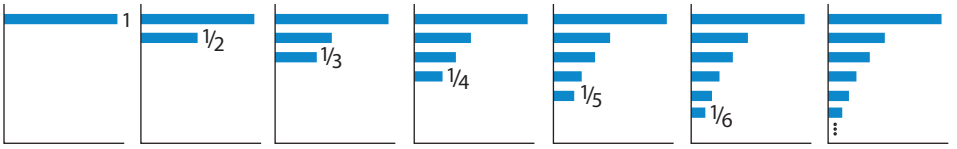

Figure 3.11 By adding shorter and shorter bars, this sequence eventually converges to a barcode with infinitely many bars!

Definition 3.3.2. Let $\overline{\mathcal{B}}$ denote the subspace of $\mathcal{B}$ consisting of those barcodes such that for all $\epsilon > 0$, the number of bars of length $> \epsilon$ is finite. We regard $\overline{\mathcal{B}}$ as a metric space with the bottleneck metric (recall Definition 2.4.8).

When working with the p-Wasserstein metric, it turns out that we need to use a slightly different finiteness condition.

Definition 3.3.3. Let $\mathcal{B}_P$ denote the subspace of $\mathcal{B}$ consisting of those barcodes B for which

$$d_{W_p}(B, \emptyset) < \infty.$$

We regard $\mathcal{B}_P$ as a metric space with the p-Wasserstein metric (recall Definition 2.4.9).

These finiteness conditions rule out phenomena like that exhibited in Example 3.3.1: we can now show that $\overline{\mathcal{B}}$ and $\mathcal{B}_P$ are complete metric spaces [60, 352].

Theorem 3.3.4. *The metric spaces $(\overline{\mathcal{B}}, d_B)$ and $(\mathcal{B}_p, d_{W_p})$ are complete and separable.*

In order to summarize distributions in $\overline{\mathcal{B}}$ and $\mathcal{B}_p$, we need to define summary statistics. In light of the discussion in the preceding section, one might hope to use the Fréchet mean and variance. Unfortunately, the Fréchet mean of a distribution of barcodes is not that useful in practice.

1. Computing the Fréchet mean is computationally expensive. An algorithm for computing an approximation to the Fréchet mean for finite sets of barcodes equipped with the empirical measure is given in [511]; however, the algorithm involves gradient descent (and so only finds local minima of the variance expression) and the rate of convergence is not well understood.
2. The Fréchet mean of a distribution μ is not necessarily unique; barcode space is positively curved [511], which means that unique geodesics do not connect all points; see Section 4.7.3. (In fact, most pairs of points are not connected by

unique geodesics, in a precise sense.) In particular Fréchet means may not be unique.
3. The Fréchet mean is very unstable; small perturbations in the sample distribution can cause the mean to jump around. To handle both this and the preceding problem, the paper [367] proposes using a distribution-valued variant of Fréchet means. Nonetheless, computation is still basically intractable.

As a consequence, the Fréchet mean and variance of distributions on barcode space are primarily of theoretical interest; in Section 3.6 below, we discuss various practical summary statistics.

3.3.2 Sampling and Hypothesis Testing in Barcode Space

We now describe our formalization of sampling problems in persistent homology using the analysis of the barcode space above. Specifically, we work with the following assumptions.

Hypothesis 3.3.5.

1. The data consists of independent samples from a metric measure space (X, ∂_X, μ_X).
2. For any k, the function assigning the kth persistent homology barcode to a sample $\{x_1, \ldots, x_n\} \subset X$ drawn from μ_X is a measurable map. (For example, in the case of the Vietoris-Rips complex, the stability theorem for persistent homology (Theorem 2.4.10) implies that persistent homology is continuous and hence measurable.)
3. Therefore, taking the product measure $\mu_X^{\otimes n}$ on X^n and then computing persistent homology, we obtain an induced measure $\mathrm{PH}_* \mu_X^{\otimes n}$ on $\overline{\mathcal{B}}$. This distribution represents the distribution of barcodes associated to PH_k computed from samples of size n.

A standard statistical approach would now be to assume that the distribution μ on (X, ∂_X) is parametrized by values $(z_1, z_2, \ldots, z_k)$. We might then hope to compute a joint density function in terms of a likelihood function. In this way, in principle, one could use a maximum likelihood method to estimate the parameters. However, in general, these kinds of statistical procedures are not really feasible, as we explained above in Section 3.1.2. The problem is that without stringent constraints there is no reasonable way to come up with sensible "topological hypotheses," for the following basic reasons. Theorem 3.1.2 and Corollary 3.1.3 show that the problem of specifying a topological hypothesis is ill posed. Only in certain special cases (e.g., the data is known to be low dimensional or known to be contractible) is it at all

reasonable to imagine producing a guess about the underlying topological type of the process generating the data or a parametric distribution for sampling from this topological space.

Even in the situation where a specific topological hypothesis is reasonable, it is often a challenging problem to provide an efficient algorithm for sampling from the null hypothesis. There are not natural parametric families of distributions for most metric spaces (X, ∂_X). Even in the case of a manifold, the most naive approach to specifying a distribution involves choosing coordinate charts and sewing together distributions on each chart – parametric inference and sampling is complicated in this setting. As an example of the difficulties, recall from the discussion in Section 3.2 above (notably Remark 3.2.18) that even correctly sampling from the volume measure on a compact Riemannian manifold defined by specific systems of equations requires some care. It is possible to compare the homology of observed data against samples generated from some standard random distribution on a compact geometric region bounding the empirical support. See Section 3.7 below for discussion of recent progress on theoretical understanding of the resulting distributions of barcodes; of course, simulation can also produce empirical estimates of these distributions. But more general topological hypotheses are out of reach except under stringent hypotheses about the dimension or complexity of the underlying space.

As a consequence, we focus on how to reliably estimate barcodes from samples and how to produce tractable features from barcodes. We can now reformulate more precise versions of the questions from the introduction to this section; we pose the problems in terms of how to estimate the persistent homology of a metric measure space (X, μ_X) from a sample $\{x_1, x_2, \ldots, x_k\}$ and use this estimate for inference. (For expositional convenience, we assume that $\text{supp}(\mu_X) = X$.)

1. If k is large enough, does the sample faithfully represent the persistent homology of the underlying space X? To be precise, if we take a sequence of finite samples S_n of increasing size from a metric measure space (X, ∂_X, μ_X), does the sequence $\{\text{PH}_k(S_n)\}$ converge to $\text{PH}_k(X)$?
2. Under the conditions for which the first question has a positive answer, how fast is the rate of convergence? Can we construct confidence intervals controlling the expected error in the estimate $\text{PH}_k(S_n)$?
3. Analogously, if our points are sampled from a density ρ on $A \subseteq \mathbb{R}^n$, can we recover the persistent homology of the level set filtration associated to the super level sets

$$\Gamma_\rho(z) = \{x \in A \mid \rho(x) > z\}.$$

Can we understand the rate of convergence and construct confidence intervals?

4. Given a collection of barcodes generated by samples of size k, how do we produce summaries of these barcodes? The discussion in Section 3.3 above suggests that the Fréchet mean is not useful in practice. A related question is how to produce numerical summaries that can be used as input to standard machine learning algorithms.
5. In the presence of noise, how can we ensure reliable estimation of barcodes? The stability theorem for persistent homology (Theorem 2.4.10) implies that if the noise is concentrated in the Gromov-Hausdorff metric, we can expect good behavior. But suppose the noise consists of "outliers" that are far from the data. How can we ensure that the estimates of persistent homology are not arbitrarily disrupted?

3.4 Stability Theorems for Persistent Homology of Metric Measure Spaces

We begin with analogues of the stability theorem in the context of metric measure spaces. We describe two related approaches to such a theorem. First, we consider distributions of samples. The idea is to consider the induced distributions on barcode space associated to the empirical persistent homology of subsamples of a fixed size (Figure 3.12). For samples of size n, we define the associated *distributional persistent homology* of a metric measure space (X, ∂_X, μ_X) as follows.

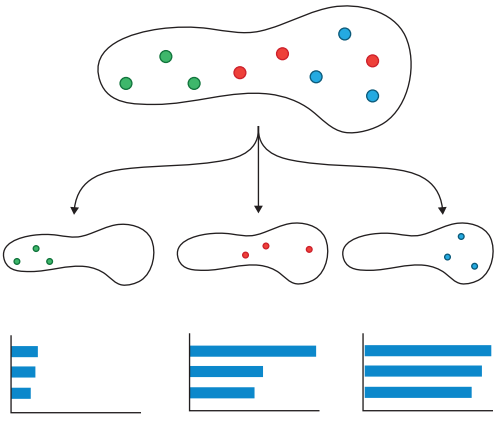

Figure 3.12 The distribution of barcodes is induced by taking many samples of a fixed size and computing their persistent homology.

Definition 3.4.1. For n and k, we define the *distributional persistent homology*

$$\Phi_k^n(X, \partial_X, \mu_X) = (\mathrm{PH}_k)_*(\mu_X^{\otimes n}),$$

the distribution on $\overline{B}$ induced by pushforward along PH_k of the product measure on the Cartesian product X^n.

In practice, we approximate Φ_k^n by sampling many blocks of size n and computing the empirical distribution; as the number of blocks approaches ∞, the law of large numbers guarantees that these approximations converge to the underlying distribution Φ_k^n. We might also subsample these blocks of size n from a larger sample from μ_X; see Figure 3.13 for an example of this.

In order for Φ_k^n to recover the persistent homology of X, the size n must be sufficiently large so that the samples can capture topological features of X; selecting n large enough requires information about the feature scale. However, even when n is too small, we can regard Φ_k^n as containing geometric information about the data, because of the following stability theorem [60].

Theorem 3.4.2. *Let (X, ∂_X, μ_X) and $(X', \partial_{X'}, \mu_{X'})$ be metric measure spaces. Fix n and k.*

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(X', \partial_{X'}, \mu_{X'})) \leq n d_{GPr}((X, \partial_X, \mu_X), (X', \partial_{X'}, \mu_{X'})).$$

Interestingly, this bound is tight (and the n is unavoidable). One way of understanding the role of n is that as n increases the invariants become finer and finer and better approximate the support of the measures, which can be far apart even though the Gromov-Prohorov distance of the metric measure spaces is small. Theorem 3.4.2 implies that the distributional invariants Φ_k^n are robust invariants, in the sense that changing X on an ϵ-probability mass arbitrarily can perturb Φ_k^n by at most $n\epsilon$. One can also formulate a Gromov-Wasserstein version of this result.

However, note that there is some subtlety to the behavior of these invariants in n; having a smaller n can make the results less sensitive to outliers since fewer noise points turn up in any given sample. On the other hand, smaller n means less resolution for detecting actual topological features of the data. Compare Figures 3.14 and 3.15.

Of course, as we have discussed, working with distributions of barcodes directly is difficult, and so in practice we will rely on ways of approximating these by distributions on $\mathbb{R}$; we will describe ways to do this in Section 3.6. Before moving on, we note two pragmatic benefits to using distributional invariants: the parameter n can be chosen to accommodate the computational power available, and the computation of Φ_k^n can be evidently parallelized with linear speedup.

We now turn to another approach to a probabilistic stability theorem which is similar in spirit to Theorem 3.4.2. We suppose we are given a data set X embedded

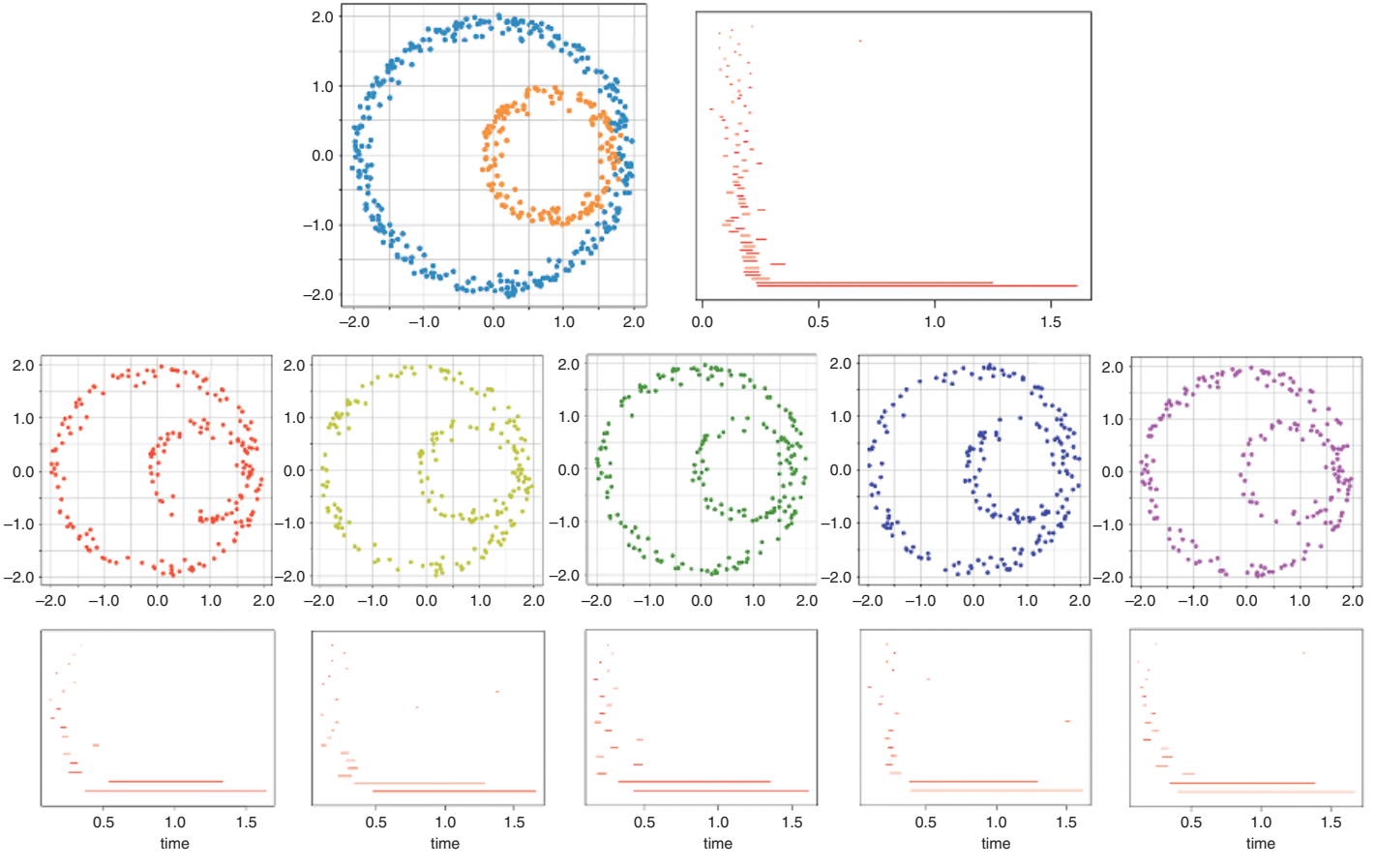

Figure 3.13 In practice, we might subsample from a large sample from the underlying distribution.

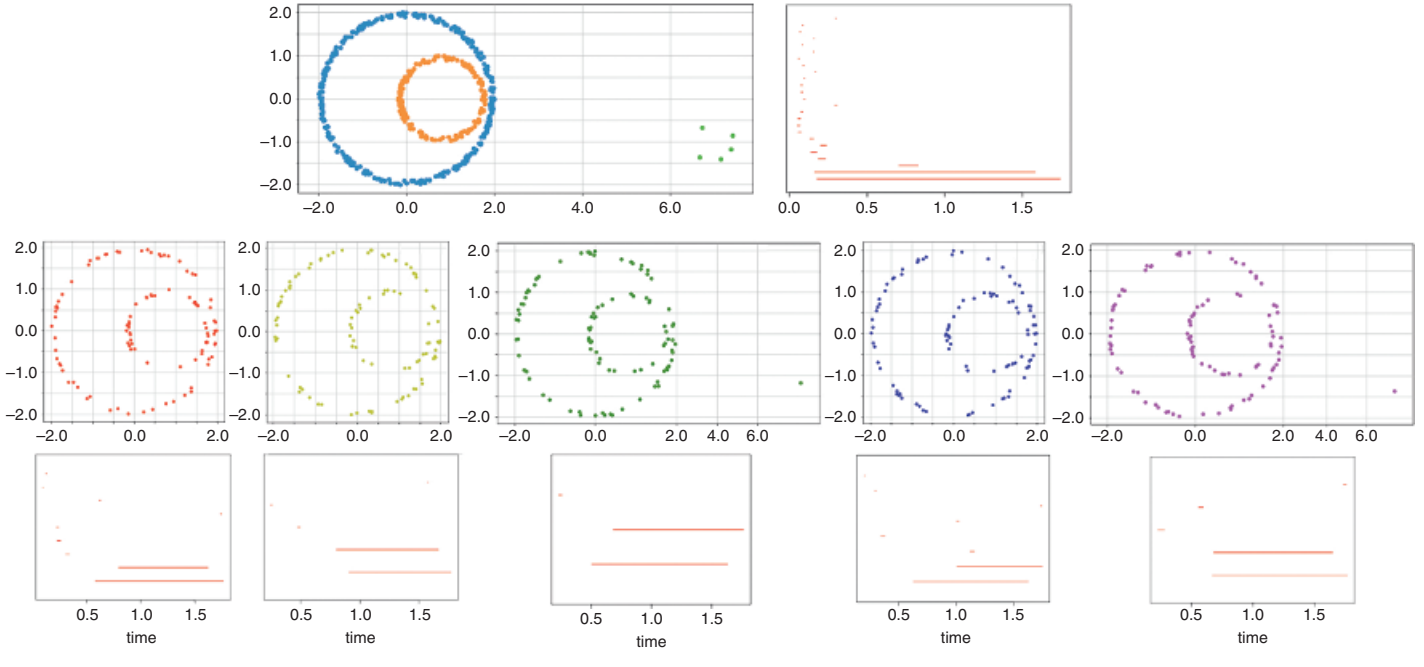

Figure 3.14 Samples of size 100 are quite clean, showing just two long bars (although note the way the bars move around relative to one another).

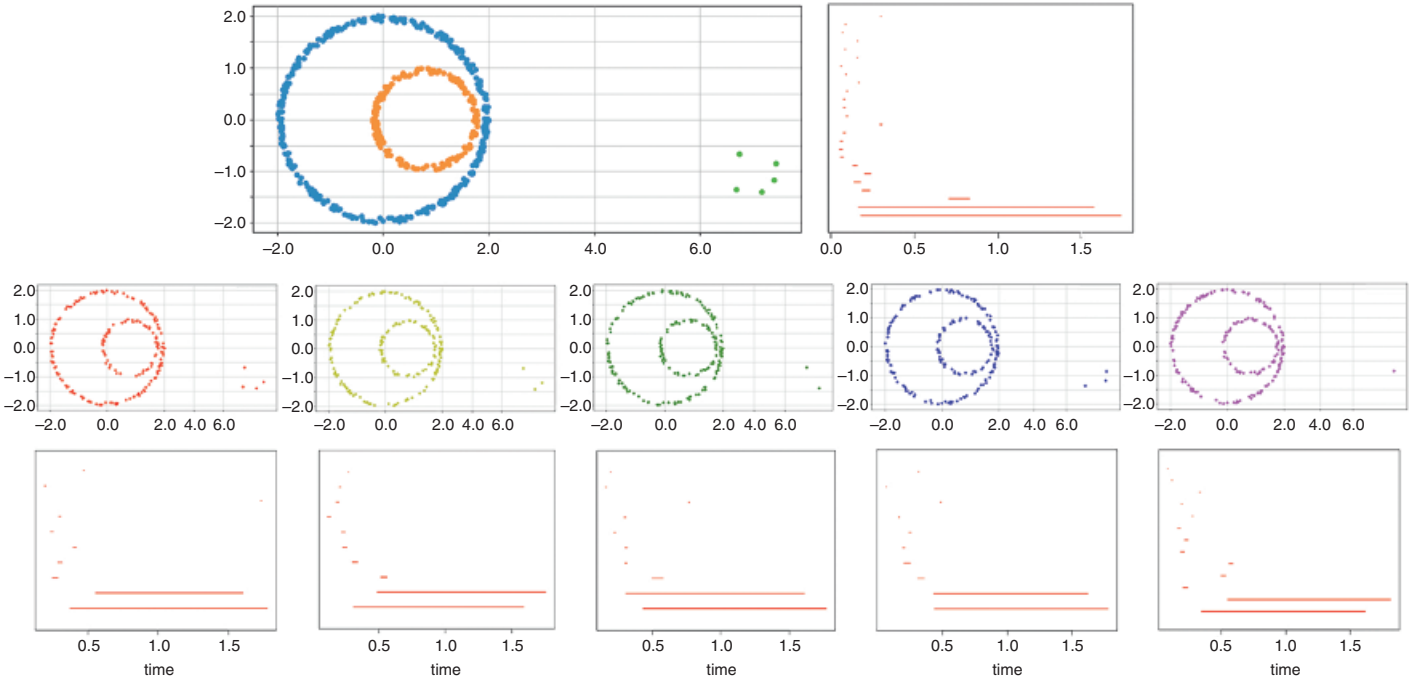

Figure 3.15 Samples of size 200 have more stability in the position of the two long bars but also have a lot more short noise bars.

in $\mathbb{R}^n$. Recall from Remark 2.3.5 that the filtered complex associated to the Čech complexes on X can be alternatively described in terms of the filtration imposed by the distance function. Specifically, let C be a compact subset of $\mathbb{R}^n$. The distance function $D\colon \mathbb{R}^n \to \mathbb{R}$ is defined as

$$D(x) = \inf_{z \in C} \partial_{\mathbb{R}^n}(x, z).$$

The sublevel sets $\{x \mid D(x) \leq \epsilon\}$ as ϵ varies are precisely the filtration imposed by the geometric Čech complexes of C. (When working with a finite metric space (X, ∂_X), the inf is replaced by the minimum.)

Estimating the persistent homology of the filtration for $X \subset \mathbb{R}^n$ via samples from some distribution on X is very sensitive to outliers. The work of [104, 108] proposes to handle this by replacing the distance function D (which captures the distance to the support of X) by a generalization that incorporates the measure on X. This generalization is referred to as the *distance to a measure*. For a continuous distribution, we have the following definition.

Definition 3.4.3. Let (X, ∂_X, μ_X) be a compact metric measure space. Let $F_x(t) = \mu_X(\{z \,|\, \partial_{\mathbb{R}^n}(x, z) \leq t\})$. Then for $0 < m < 1$ we define the *distance to a measure* to be

$$\delta_{\mu_X, m}(x) = \sqrt{\frac{1}{m} \int_0^m F_x^{-1}(u)^2 du},$$

where here

$$F_x^{-1}(u) = \inf_t \{t \mid F_x(t) \geq u\}.$$

Here m is a resolution parameter that is a measure of the feature scale; choice of suitable values of m is once again an issue in practical use. The idea of the parameter m is that we are averaging density-biased approximations to the distance over a range controlled by m. Along these lines, for finite samples, the distance to a measure has a much simpler expression.

Lemma 3.4.4. *Given a finite sample $Y = \{x_1, x_2, \ldots, x_n\} \subseteq X$, the distance to a measure function for the empirical distribution on Y for m is*

$$\delta_m(x) = \sqrt{\frac{1}{k} \sum_{z_\alpha \in N_k(x)} \partial_{\mathbb{R}^n}(z_\alpha, x)^2},$$

where here k is the smallest integer $\geq mn$ and $N_k(x)$ denotes the k nearest neighbors of x in Y.

Notice that when m is very small, the distance to a measure function is very close to the distance function D. The advantage of the distance to a measure is that

it is Wasserstein stable, in the sense that the L_∞ norm distance is bounded by the 2-Wasserstein distance. Specifically, we have the following theorem.

Theorem 3.4.5. *Suppose that μ_1 and μ_2 are two probability measures on $\mathbb{R}^n$. Then for mass parameter $0 < m < 1$, we have*

$$\|\delta_{\mu_1,m} - \delta_{\mu_2,m}\|_\infty \leq \frac{1}{\sqrt{m}} d_{W_2}(\mu_1, \mu_2).$$

In turn, the bottleneck distance between the persistence diagrams associated to the distance filtrations on these two functions is bounded by the L_∞ norm.

Corollary 3.4.6. *Suppose that μ_1 and μ_2 are two probability measures on $\mathbb{R}^n$. Then for mass parameter $0 < m < 1$, we have*

$$d_B(P_{\delta_{\mu_1,m}}, P_{\delta_{\mu_2,m}}) \leq \|\delta_{\mu_1,m} - \delta_{\mu_2,m}\|_\infty \leq \frac{1}{\sqrt{m}} d_{W_2}(\mu_1, \mu_2).$$

As a consequence, we can conclude that the persistent homology estimate associated to the distance to a measure filtration is robust to outliers having low probability mass. (See Figure 3.16 for an example demonstrating robustness in the face of outliers.)

Furthermore, one can show [108] that the distance to a measure is statistically well behaved in the sense that a uniform law of large numbers applies to establish that it can be approximated by finite samples. Moreover, there are natural confidence intervals describing how well it is approximated by empirical estimates. Another interesting aspect of the distance to a measure is that, since for small m it approaches the ordinary distance function to X, in principle it can be used for geometric inference. On the other hand, computing the distance to a measure is difficult in practice due to problems associated to estimating level sets. See [79] for recent work that provides better algorithms and also extends the methodology to arbitrary metric spaces.

We now turn to the issue of understanding the way that the empirical persistent homology converges to the persistent homology of the underlying space.

3.5 Estimating Persistent Homology from Samples

Suppose we take a sequence of finite samples S_n of increasing size from a metric measure space (X, ∂_X, μ_X). It is straightforward to see that in fact $\{\mathrm{PH}_k(S_n)\}$ does converge to $\mathrm{PH}_k(\mathrm{supp}(\mu_X))$ almost surely, provided that X is bounded: Lemma 1.2.20 shows that for any compact metric measure space (X, ∂_X), there

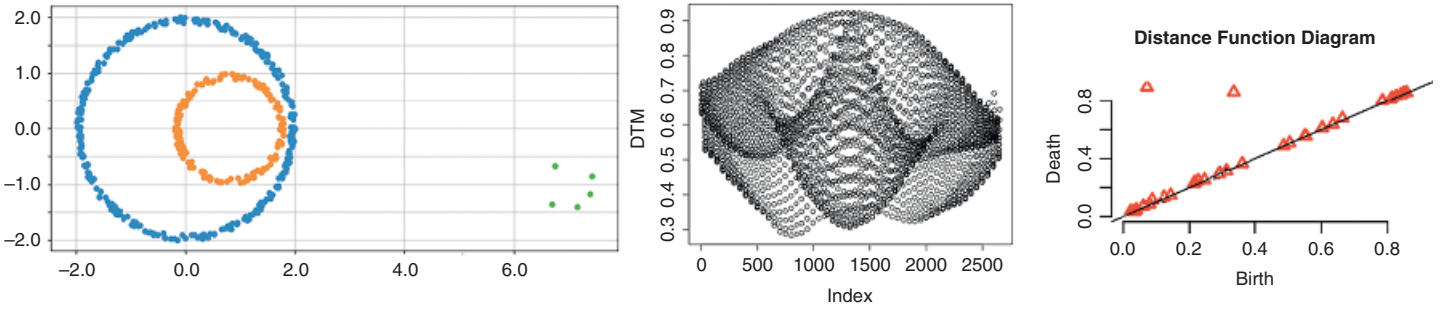

Figure 3.16 The level set persistence of the distance to a measure recovers the persistent homology of the two circles and ignores the outliers. (The middle panel plots the distance function.)

exists a finite ϵ-net X_ϵ for each $\epsilon > 0$. If we were given a sequence $\{X_n\}$ such that as $n \to \infty$, X_n is an $\frac{1}{n}$-net,

$$\{X_n\} \longrightarrow X$$

in the Gromov-Hausdorff metric and so

$$\{\mathrm{PH}_k(X_n)\} \longrightarrow \mathrm{PH}_k(X)$$

in the barcode metric. The point now is that for any ϵ, there exists an n sufficiently large so that any sample of size $> n$ is with high probability an $\frac{1}{n}$-net. This implies the following result.

Theorem 3.5.1. *Let (X, ∂_X, μ_X) be a metric measure space. Let $\{S_n\}$ be a sequence of finite samples drawn from μ_X such that $|S_n| \to \infty$. Then almost surely* $\mathrm{PH}_k(S_n)$ *converges to* $\mathrm{PH}(\mathrm{supp}(\mu_X))$ *in the barcode metric (or Wasserstein metric).*

Theorem 3.5.1 focuses attention on the rate of convergence of $\{PH_k(S_n)\}$. The key issue is to analyze the number of samples needed to obtain an ϵ-net with high probability (for some fixed ϵ). Such estimates require knowledge of the feature scale; we need to be able to compute how likely we are to sample in a ball around any given point. Estimates for compact Riemannian manifolds were given by Niyogi-Smale-Weinberger [384] (as explained in our discussion of Theorem 2.2.1), and elaborated on and extended by [170]. We describe the problem in the framework of the latter, which is more general and is expressed explicitly in terms of the language of *confidence regions*.

A confidence region is the multivariate analogue of the basic statistical notion of a confidence interval, which we now review. Returning to our example of estimating parameters of a Gaussian, we suppose that we have a sample $\{x_1, \ldots, x_n\}$ from a Gaussian distribution with mean μ and standard deviation σ. As discussed above, to estimate μ we compute the empirical mean $\hat{\mu}$ from the samples. We know that as n increases, it is very likely that $\hat{\mu}$ will be a good approximation of μ. One way to make that precise is to talk about a confidence interval.

Definition 3.5.2. A confidence interval $[a, b]$ with confidence level α for the parameter θ is specified by two random statistics a and b such that the probability that $\theta \in [a, b]$ is α.

For example, we know that $\hat{\mu}$ is distributed according to the t-distribution around μ with parameters determined by $\hat{\sigma}$, and using this fact we can derive the confidence interval for μ

$$\left[\hat{\mu} - \frac{c\hat{\sigma}}{\sqrt{n-1}}, \hat{\mu} + \frac{c\hat{\sigma}}{\sqrt{n-1}}\right],$$

where c is chosen such that the probability in the tail of the distribution larger than c has mass $\frac{1-\alpha}{2}$.

We now turn to the analogous notions for persistence diagrams. Associated to a specific c, the confidence set around a barcode B is a subset of the set of barcodes within a distance c of B. We can visualize this as the union of squares with side-length $2c$ is centered at each point of the persistence diagram. Points where the bounding box intersects the diagonal can be interpreted as noise. (Alternatively, we can put a band of width $(\sqrt{2})c$ around the diagonal.) See Figure 3.17 for an example.

To define a confidence set with probability α, we need to find c such that the true parameter is within c of the empirical barcode with probability larger than α. To formulate this, it turns out to be useful to talk about asymptotic confidence sets, defined as follows.

Definition 3.5.3. Fix a reference barcode $\mathcal{B}$ and denote by $\widehat{\mathcal{B}}_n$ the empirical barcode computed from a sample of size n. For $0 < \alpha < 1$, the *asymptotic* $1 - \alpha$ *confidence set* is the collection of regions determined by a (usually decreasing) sequence $c_n > 0$, where

$$\limsup_{n\to\infty} \Pr(d_B(\mathcal{B}, \widehat{\mathcal{B}}_n) > c_n)) < \alpha.$$

(Recall that lim sup denotes the limit of the supremums of the remaining terms in the sequence.)

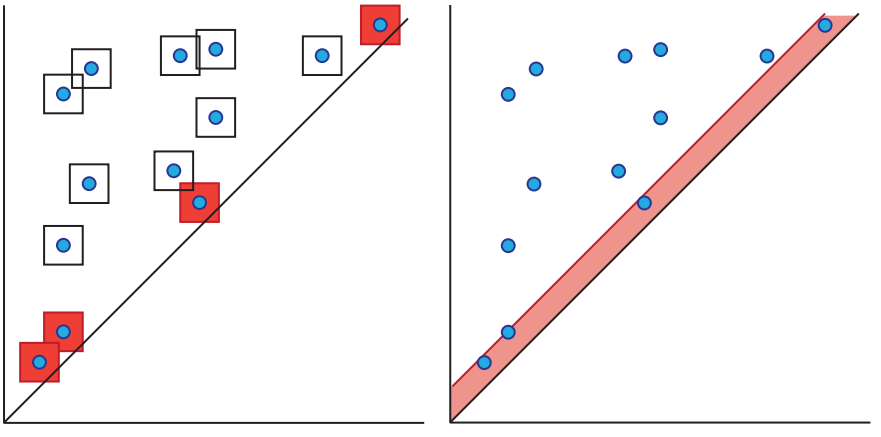

Figure 3.17 The confidence interval around the persistence diagram (in blue) is given by the boxes; a band around the diagonal contains "noise."

As one would expect, the rate of convergence (i.e., how large n has to be in order to obtain sufficiently small c_n) depends on the details of the density and the feature scale of the underlying manifold space. Going forward, we will assume that M is a compact manifold of dimension d embedded in $\mathbb{R}^k$ ($k > d$), that the condition number (recall Section 2.2) of M is positive, and that the samples are drawn from a probability density on $\mathbb{R}^k$ which is supported on M, smooth, and bounded away from 0.

Remark 3.5.4. More generally, it suffices for M to be a compact and rectifiable (piecewise smooth) subset of Euclidean space and to have a relatively weak differentiability criterion for M.

To bound the convergence of the confidence intervals for persistence diagrams, we define

$$\rho(x,t) = \frac{\Pr(B_{\frac{t}{2}}(x))}{t^d} \qquad \text{and} \qquad \rho(t) = \inf_{x \in M} \rho(x,t).$$

Then $\rho = \lim_{t \to 0} \rho(t)$ captures relevant information about the local variation in the probability measure on M.

We now fix our space $M \subset \mathbb{R}^k$ and let $\mathcal{P}$ denote the persistent homology of the sublevel sets of the function

$$\partial_M(z) = \inf_{y \in M} \partial_{\mathbb{R}^k}(y,z).$$

(Recall from Remark 2.3.5 that this is a version of the Čech complex.) For a sample of size n, let $\widehat{\mathcal{P}_n}$ denote the empirical persistent homology, i.e., the persistent homology of the sublevel sets of ∂_M restricted to $\mathcal{P}_n$. We have the following analogue of Theorem 2.2.1.

Proposition 3.5.5. *Under the hypotheses above,*

$$\Pr(d_B(\mathcal{P}, \widehat{\mathcal{P}_n}) > t) \leq \frac{2^d}{\rho(\frac{t}{2})t^d} e^{-n\rho(t)t^d}.$$

The associated confidence region is the collection of boxes of side length t centered at the points of the persistence diagram $\mathcal{P}_n$.

In particular, setting

$$t_n = \left(\frac{4 \log n}{\rho n} \right)^{\frac{1}{d}},$$

we have that

$$\Pr(d_B(\mathcal{P}, \widehat{\mathcal{P}_n}) > t_n) < \frac{2^{d-1}}{n \log n}.$$

Making use of this result involves estimating ρ, which can be done using the plug-in estimator

$$\hat{\rho}_n = \min_i \frac{P_n(B_{\frac{r_n}{2}}(x_i))}{r_n^d},$$

where r_n is a sequence of numbers approaching 0 and P_n denotes the empirical measure for the sample $\{x_1, x_2, \ldots, x_n\}$.

There are a number of other methods of obtaining similar confidence interval estimates that are of broader interest; we turn to discussion of those in the remainder of the section.

3.5.1 Estimating Persistent Homology by Density Estimation

Another approach to computing the persistent homology from samples of a density in Euclidean space is to use standard techniques for density estimation to approximate the support of the density (e.g., see [429] for a modern theoretical analysis). Given a suitable probability density ρ on $\mathbb{R}^d$, the problem of estimating the superlevel sets

$$\Gamma_\rho(z) = \{x \in \mathbb{R}^d \mid \rho(x) > z\}$$

is a classical question in statistics. The path-connected components of $\Gamma_\rho(z)$ have long been studied in the context of unsupervised clustering and classification [229].

From the perspective of persistence, a natural question is to try to estimate the persistent homology of the level set filtration determined by the inclusions

$$\Gamma_\rho(z_2) \subseteq \Gamma_\rho(z_1)$$

for $z_1 < z_2$. A standard approach is to use a kernel density estimator; this is a smoothed version of the empirical density. The specific choice of kernel function employed is not important for our discussion, except for the following properties. We require a function $K\colon \mathbb{R} \to \mathbb{R}$ such that

1. $\int K = 1$,
2. the kernel has mean 0,
3. $\sup_x K(x) = K(0)$, and
4. K is Lipschitz for some constant ℓ.

Typically we will think of a smooth symmetric kernel, e.g., the Gaussian kernel $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

For a bandwidth parameter h (this controls the amount of smoothing), define the measure

$$K_h(A) = h^{-d} \int_A K(h^{-1}t)dt.$$

Given the density ρ and associated measure P on $\mathbb{R}^d$, we want to study the convolution $P_h = K_h * P$, which we regard as a smoothed version of P. Denote the level set persistent homology of P_h by $\mathrm{PH}_k(P_h)$.

We can form an empirical approximation as follows. The density of the convolution is

$$p_h(x) = \int_M \frac{1}{h^d} K\left(\frac{\partial_{\mathbb{R}^d}(x,u)}{h}\right) dP(u),$$

and so the standard estimator given points $\{x_1, \ldots, x_n\}$ is given by

$$\hat{p_h}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\partial_{\mathbb{R}^d}(x,x_i)}{h}\right). \tag{3.1}$$

We can now compute the persistence diagram associated to the level set filtration determined by the estimated density $\hat{p_h}$, which we will denote by $\mathrm{PH}_k(\hat{P}_h)$.

Remark 3.5.6. We note that this is estimating a somewhat different quantity than the persistent homology of the support of ρ; instead, we are in some sense directly estimating the homology of the support of ρ using the persistent homology of the level set filtration. This increases the robustness of the result, due to smoothing.

For simplicity, we assume that the support of the distribution P is contained in the Euclidean box $[-c, c]^d \subseteq \mathbb{R}^d$. Standard arguments show that $\hat{p}_h$ converges to p_h; this follows from Hoeffding's inequality, for example. (And stronger statements can be derived from tighter refinements of this sort of bound.) Translating this into a statement about persistence diagrams, we obtain the following result.

Theorem 3.5.7. *Under the hypotheses above, for fixed α and for any distribution P supported on the box $[-c, c]^d$*

$$\Pr(d_{W_p}(\mathrm{PH}_k(\hat{P}_h), \mathrm{PH}_k(P_h)) > \delta_n) \leq \alpha$$

and where δ_n is a solution to the equation

$$2\left(\frac{4c\ell\sqrt{d}}{\delta_n h^{d+1}}\right)^d e^{-\frac{n\delta_n h^{2d}}{2K(0)^2}} = \alpha.$$

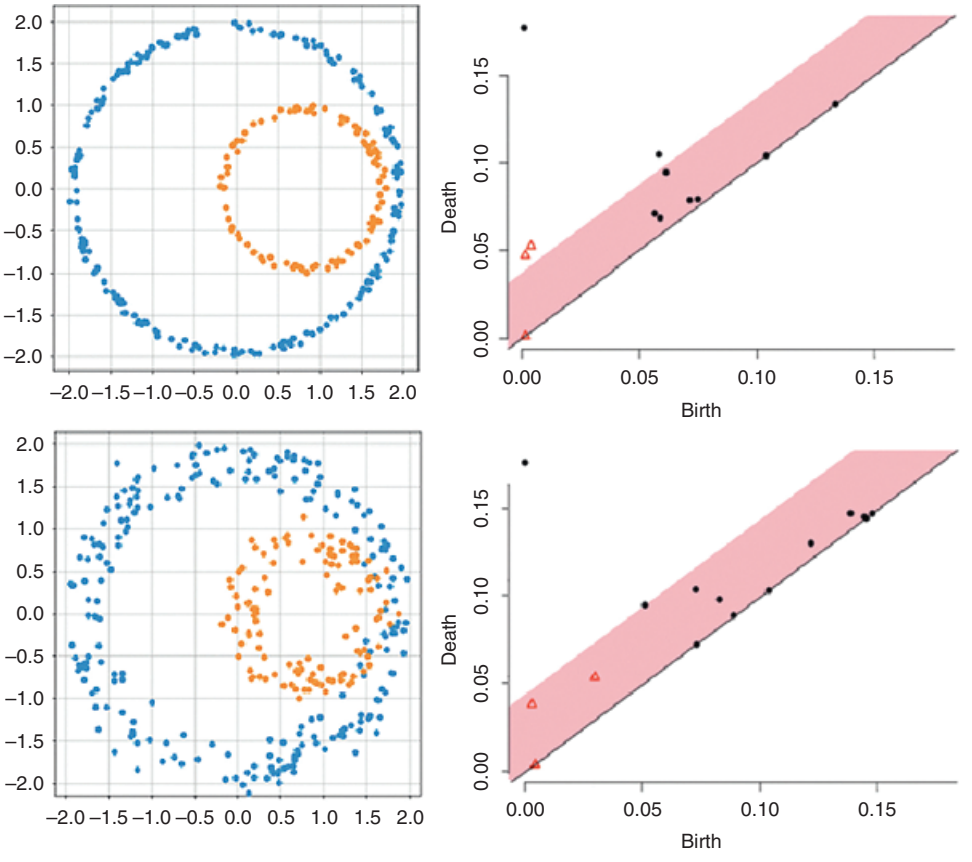

Figure 3.18 In the top panel, the 95% confidence interval contains the two bars for both H_0 and H_1 (dots represent H_0, triangles H_1); this correctly separates signal from noise. However, in the bottom panel, the 95% confidence interval suggests that all of the H_1 bars are noise.

As we can see in Figure 3.18, the confidence intervals computed in this fashion are fairly conservative.

For data embedded in Euclidean space, density estimation can also be used to eliminate outliers by smoothing to remove regions of low density. For example, this was performed manually in the famous example of the Klein bottle in visual image data [95], and is a standard data analysis tool [251]. Specifically, the persistent homology associated to the level set filtration of a density estimator is robust in the presence of outliers.

Let $X \subseteq \mathbb{R}^n$ denote the set of all points that might be returned by sampling, including both data points and noise points, i.e.,

$$X = X' \cup Z, \quad \text{where} \quad Z \cap X' = \emptyset,$$

where we regard X' as real data and Z as noise. Assume that the distribution on X we have experimental access to is

$$\Psi = \epsilon\theta + (1 - \epsilon)\mu,$$

for $0 \leq \epsilon \leq 1$, where μ is supported on X' and is the distribution we wish to estimate. We make no assumptions about θ.

Denote by $\mathcal{P}_\rho$ the persistence diagram associated to the level set filtration of the standard density estimator of equation (3.1), for fixed width parameter h, applied to empirical samples from a distribution ρ. The following lemma is now a simple calculation [170].

Lemma 3.5.8. *Let $X \subseteq \mathbb{R}^n$ be a subspace with probability density $\Psi = \epsilon\theta+(1-\epsilon)\mu$. Then*

$$d_B(\mathcal{P}_\Psi, \mathcal{P}_\mu) \leq C\epsilon,$$

where C is a constant that depends on h.

This result implies that when ϵ is small and h is chosen appropriately, $\mathcal{P}_\Psi$ is a good approximation to $\mathcal{P}_\mu$ no matter what θ is, in particular, no matter how far away from X' the points of Z may be. Simple experiments in low dimensions validate this result [170].

Although this result is very encouraging, the general problems with density filtering remain – namely, choosing the width parameter requires either knowledge of the feature scale of the underlying data or a lot of experimentation, and density filtering is really only tractable for data embedded in Euclidean space or comparatively simple manifolds (see Figure 3.19). (Nearest neighbor density estimators do not perform well for realistic numbers of sample points.)

We believe that density filtering could be an ideal application of multidimensional persistence.

3.5.2 Estimating Persistent Homology by Resampling

Resampling is a standard technique for estimating confidence intervals around an empirical estimate of some quantity by generating many new finite subsamples from the given finite sample. Given n data points $X = \{x_1, x_2, \ldots, x_n\}$, there are two distinct possibilities for resampling estimators.

1. Subsampling involves estimating confidence intervals from empirical quantiles computed from subsamples $\{S_i\}$ of size $k < n$ generated by drawing *without* replacement from the empirical distribution on X (e.g., [412]).

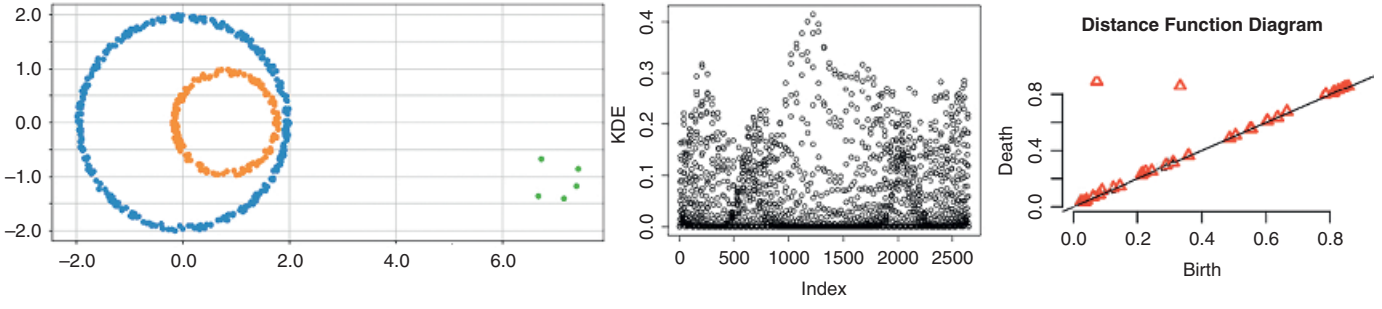

Figure 3.19 Using a density estimator provides a robust computation of the persistent homology.

2. The bootstrap involves estimating confidence intervals from empirical quantiles computed from subsamples $\{S_i\}$ of size $k < n$ generated by drawing *with* replacement from the empirical distribution on X (e.g., [54]).

We now discuss the use of these ideas to estimate persistent homology from finite samples. We start with the first case above, subsampling. Results in this regime are asymptotic and so stated in terms of the convergence of both n and k to ∞. We first work with the hypotheses of Proposition 3.5.5.

Remark 3.5.9. In the following discussion, to talk about asymptotic convergence we use "big-O" and "little-o" notation.

1. To say that a sequence $\{x_n\}$ is $o(f(n))$ means that for every $k \in \mathbb{R}$, there exists an $N \in \mathbb{N}$ such that for all $m > N$, $x_m < kf(m)$.
2. To say that a sequence $\{x_n\}$ is $O(f(n))$ means that there exists a constant $k \in \mathbb{R}$ and $N \in \mathbb{N}$ such that for all $m > N$, $x_m < kf(m)$.

Roughly speaking, the sequence is $o(f(n))$ if it grows strictly more slowly than the function f whereas the sequence is $O(f(n))$ if it grows at most as fast as a constant times $f(n)$.

Let b_n denote a sequence such that

$$b_n \to \infty \qquad \text{and} \qquad b_n = o\left(\frac{n}{\log n}\right).$$

Let $N = \binom{n}{b_n}$, and denote by $\{S_i\}$ the collection of all N subsamples of size b_n from the given sample $\{x_1, x_2, \ldots, x_n\}$. Set

$$L_n(t) = \frac{1}{N} \sum_{j=1}^{N} I(d_H(S_i, S) > t),$$

where I is the indicator function and d_H is the Hausdorff metric. For a given $\alpha \in (0, 1)$, let

$$c_n = 2L_n^{-1}(\alpha).$$

The arguments of [412] then imply convergence of the subsamples to the underlying metric space in Hausdorff measure and hence the following theorem providing confidence regions.

Theorem 3.5.10. *Under the hypotheses of Proposition 3.5.5, for large n (and $\rho > 0$), we have*

$$\Pr(d_B(\mathcal{P}, \widehat{\mathcal{P}_n}) > c_n) \leq \alpha + O\left(\left(\frac{b_n}{n}\right)^{\frac{1}{4}}\right).$$

We can also apply the bootstrap; in this situation, the best results come from considering the context of level set estimation from the density estimator. We work with the hypotheses of Section 3.5.1.

Then we have the following theorem.

Theorem 3.5.11. *Under the hypotheses of Theorem 3.5.7, we have that*

$$\lim_{n\to\infty} \Pr\left(d_B(\mathcal{P}_h, \widehat{\mathcal{P}_h}) > \frac{q_\alpha}{\sqrt{n}}\right) \leq \alpha.$$

Here q_α is the $1-\alpha$ quantile and is described below. The estimated confidence interval is then of width $\frac{2q_\alpha}{\sqrt{n}}$.

We can estimate the value q_α as

$$\hat{q}_\alpha = \inf_q \left(\frac{1}{N}\sum_{i=1}^{N} I(\sqrt{n}\|\widehat{p_h^i} - \widehat{p_h}\|_\infty \geq q) \leq \alpha\right),$$

where $\widehat{p_h^i}$ is the empirical probability density of the ith bootstrap subsample and $\|(-)\|_\infty$ denotes the L_∞ norm. Figure 3.20 has an example of confidence regions produced in this fashion; again, notice that these regions are quite conservative.

Remark 3.5.12. It is also possible to show that resampling methods and the bootstrap can be applied directly in barcode space; this is more challenging technically due to the complexity of the metric geometry of $\mathcal{B}$. The issue is that establishing the asymptotic consistency of the bootstrap depends on obtaining control on the complexity of the class of functions used to describe empirical processes. For example, in $\mathbb{R}$, one uses the indicator functions supported on intervals $(-\infty, t]$. In barcode space, bounding the complexity of natural function classes is difficult and requires imposing further finiteness restrictions on the allowable barcodes.

3.6 Summarizing Persistence Diagrams

The results of Section 3.3 and Section 3.4 imply that it is possible in some circumstances to reliably estimate the persistent homology of a geometric object from samples. However, as we have emphasized, it remains difficult to directly apply the estimated barcode to inference. In view of this, a compelling approach is to study associated features produced by a choice of measurable map

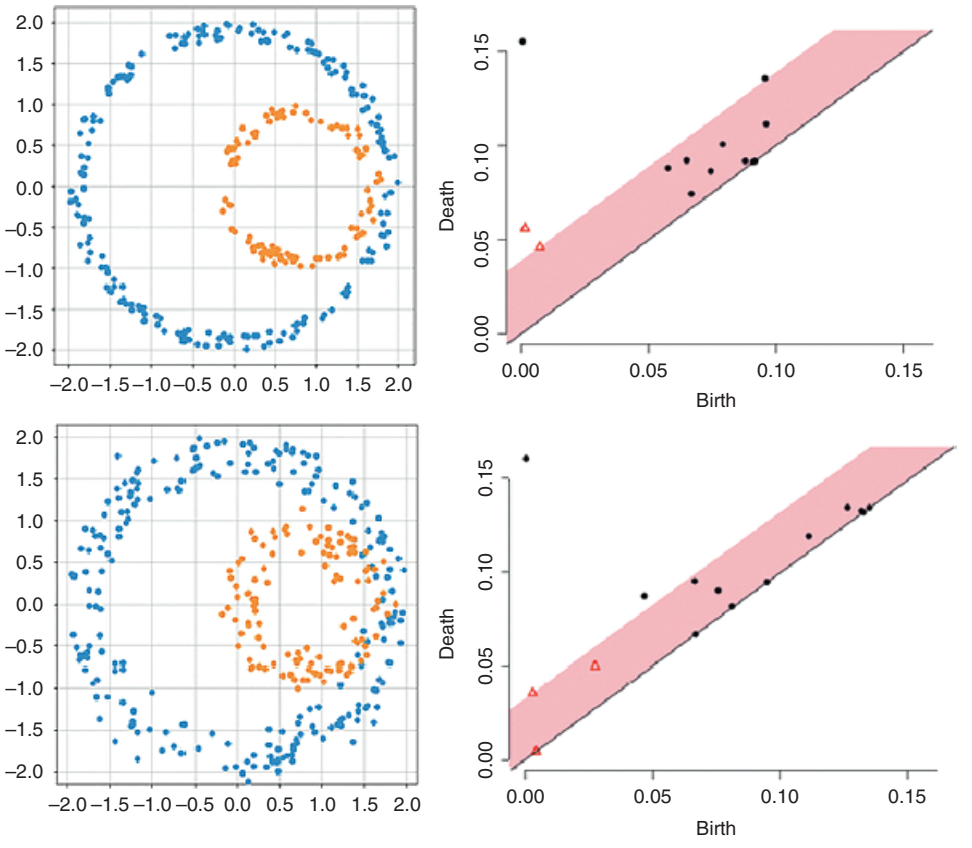

Figure 3.20 In the top panel, the 95% confidence interval clearly contains one bar and has a second at the edge for both H_0 and H_1; this correctly separates signal from noise. However, in the bottom panel, the 95% confidence interval suggests that all of the H_1 bars are noise whereas both the H_0 bars appear significant.

$$\theta\colon \mathcal{B} \to \mathbb{R}^n.$$

More generally, we might consider a measurable map

$$\theta\colon \mathcal{B} \to V,$$

for a vector space V which has a compatible topology (e.g., induced by the norm metric); V is regarded as equipped with the Borel σ-algebra. Then a distribution ρ on barcode space induces a pushforward distribution $\theta_*\rho$ on $\mathbb{R}^n$ or V.

This methodology has two substantial concrete benefits.

1. Many standard techniques in classical statistics apply essentially immediately to the distribution $\theta_*\rho$ on $\mathbb{R}^n$ or V. For example, summary statistics for $\theta_*\rho$, while not necessarily corresponding to any barcode, are now easy to compute

and work with. Consistency and convergence rates for empirical estimates can be quickly derived.
2. The resulting statistics can be used as input to visualization techniques or also as features for machine learning, e.g., classification and clustering algorithms. A particular advantage here is that such features can be combined with other sources of information or statistics produced from the raw data.

We can summarize the benefits of this simplification approach in terms of the following meta-theorem.

Theorem 3.6.1 (Meta-theorem of real projections from barcode space)**.** *For any reasonable real-valued test statistic of barcodes, i.e., a suitable map $\mathcal{B} \to \mathbb{R}^n$, all the standard theorems and techniques of statistics and machine learning can be applied to the pushforward of any distribution on $\mathcal{B}$.*

There is infinite variety in the choice of feature maps to apply; in the remainder of this section, we discuss some representative examples.

3.6.1 Tractable Features from Persistence Diagrams

We begin by considering two simple and generic approaches for embedding arbitrary metric spaces in $\mathbb{R}^m$: the distance distribution and landmark embeddings. Both of these are easy to apply to distributions of barcodes, and yield distributions on Euclidean space. Then, for example, the mean of the pushforward distribution is a useful summary statistic.

The distance distribution is simply the induced distribution produced by computing distances between points; the next definition makes sense since the metric is always a measurable map on a metric measure space. See Figure 3.21 for a simple example.

Definition 3.6.2. Let (X, μ_X, ∂_X) be a metric measure space. The *distance distribution* on $\mathbb{R}$ is defined to be the pushforward $(\partial_X)_* \mu_X^{\otimes 2}$ of the distribution $\mu_X^{\otimes 2}$ on $X \times X$ along the function $\partial_X \colon X \times X \to \mathbb{R}$.

There are various elaborations of this example; for instance, one could consider distributions of $k \times k$ distance matrices induced by samples of size k. (Distance matrices as summaries of barcodes were studied in [97].)

Remark 3.6.3. In fact, a famous result of Gromov implies that a metric measure space is uniquely characterized by such distance matrix distributions for all k, in

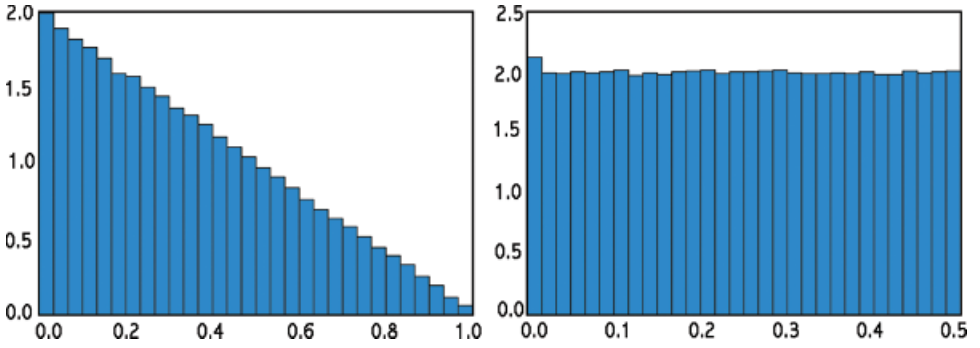

Figure 3.21 Left: The distance distribution for 1000 points sampled uniformly from $[0, 1]$. Right: The distance distribution for 1000 points sampled uniformly from S^1.

the sense that two metric measure spaces (X, ∂_X, μ_X) and (Y, ∂_Y, μ_Y) are isomorphic if and only if the distance distributions coincide as k goes to ∞ [212].

Another possibility is to consider distances to a fixed collection of points. Choose k landmark points $\{\ell_1, \ldots, \ell_k\}$; these can be selected arbitrarily, or as points of interest based on domain knowledge, or via a randomized algorithm biased to choose a point far from the existing landmarks, etc.

Definition 3.6.4. Let (X, μ_X, ∂_X) be a metric measure space and take a finite subset $\{\ell_1, \ldots, \ell_k\} \subset X$. Then the *landmark embedding distribution* on $\mathbb{R}^k$ is the pushforward of μ_X along the function $X \to \mathbb{R}^k$ specified by the formula

$$x \mapsto (\partial_X(x, \ell_1), \partial_X(x, \ell_2), \ldots, \partial_X(x, \ell_k)) .$$

Remark 3.6.5. The selection of landmarks introduces many new statistical problems. For instance, the choice of k introduces a rough notion combining dimension and feature scale; the larger the dimension and the smaller the feature scale, the more landmark points one needs. Moreover, questions of stability of the results in the face of shifts in landmark points immediately arise. Currently, there are not many theoretical results in this regime (e.g., recall the discussion of the properties of the weak witness complex in Section 2.7). On the other hand, standard statistical tools (e.g., empirical confidence intervals for quantities computed from these distributions) can be applied to handle such issues.

The landmark distribution is a first guess at how to embed a metric space in Euclidean space. There is in fact an enormous literature on the problem of

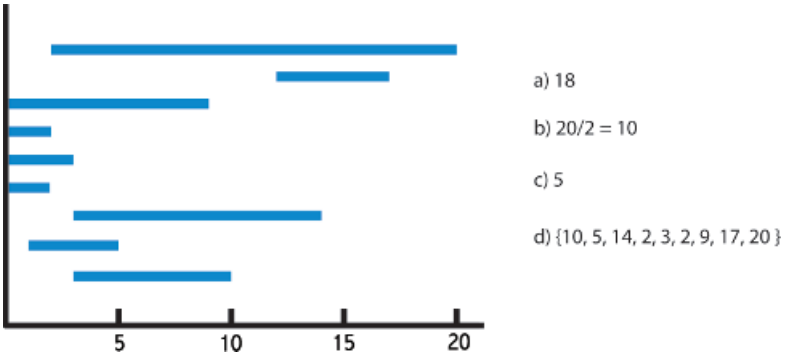

Figure 3.22 (a) The length of the longest bar, (b) the ratio of the endpoints of the longest bar, (c) the number of bars of length over 4, and (d) the righthand endpoints of each bar.

efficiently embedding a finite metric space in Euclidean space in a way which minimizes distortion (e.g., see [168] for a celebrated and essentially optimal result); although there has not been much investigation so far of these techniques in TDA (although see [455] for work that employs methods from this literature) we expect that this will be a useful avenue of research.

There are also many specific invariants of barcodes that provide values in $\mathbb{R}$ or $\mathbb{R}^n$; we provide some representative examples. (See Figure 3.22 for a specific example.) Note that a basic and important issue to consider for any such feature is whether it is stable with respect to perturbation in the barcode metric.

1. For a barcode B we can define

$$g_m(B) = |B(m)| - |B(m+1)| \qquad \text{and} \qquad h_m(B) = \frac{B(m)}{B(m+1)}$$

   where $B(k)$ denotes the kth largest interval in B.
2. Given a barcode B, we can consider the set of birth-times $\{x_i\}$ or the set of death-times $\{x_i\}$ to provide a map to $\mathbb{R}^n$, where n is a bound on the size of the barcodes we consider.
3. Given a barcode B, we can consider a map to $\mathbb{Z}$ given by the number of non-zero bars or the number of non-zero bars greater than some minimum length ϵ.
4. Given a barcode B, we can consider a map to $\mathbb{R}^n$ given by the set of lengths $\{y_i - x_i\}$ or the set of size ratios $\left\{\frac{x_i}{y_i}\right\}$; to make sense of this, we must again bound the size of the barcodes and also sort the bars by length.

In [49], an explicit embedding of persistence diagrams in high-dimensional Euclidean space is considered; the idea is to take a grid on the persistence diagram and count barcode points within it. Unfortunately, this is not stable for certain kinds of perturbations of the barcodes that are small in the bottleneck distance.

3.6.2 Kernel Methods for Barcodes

The idea of *kernel methods* for machine learning involves embedding the data points in some kind of infinite-dimensional vector space where standard machine learning techniques apply. The trick is that rather than working with the embedding directly, it turns out to be sufficient to understand the inner product between two points in the embedded space; this is the kernel function. We say a bit more about the specifics of this in Section 4.3.4. Here, we focus on explaining the construction and definition of kernels for barcodes based on approximating a persistence diagram with a sum of Gaussian functions [4, 425]. These sorts of approaches yield kernels that are stable in the bottleneck and p-Wasserstein metrics on barcodes and provide sensible feature vectors for machine learning.

In [425], the kernel at scale σ for persistence diagrams D_1 and D_2 is computed by the formula

$$k_\sigma(D_1, D_2) = \frac{1}{8\pi\sigma} \sum_{\substack{p\in D_1 \\ q\in D_2}} e^{\frac{-\partial(p,q)^2}{8\sigma}} - e^{\frac{-\partial(p,\bar{q})^2}{8\sigma}},$$

where $\bar{q}$ denotes the reflection across the line $x = y$. Roughly speaking, we can think of this as a approximation by positive and negative Gaussians. The basic idea is that a persistence diagram can be approximated in function space as a sum of Dirac δ-functions centered at the points. However, the resulting metric on functions does not incorporate information about the proximity to the diagonal (i.e., bars of zero length). So instead, the δ-functions are used to specify a diffusion equation with the diagonal providing boundary constraints; the resulting solutions are Gaussians.

In contrast, in [4] a closely related approach was studied which uses weighted positive Gaussians; the difference in weights permits more flexibility in focusing on different features in the barcodes, and the use of positive Gaussians in some circumstances can provide computational efficiency. Although this is not phrased as a kernel method per se (but simply as a vector-space valued summary), it can be applied to produce a kernel just as in [425].

Remark 3.6.6. We can regard the grid counting method of [49] as a discretization of the Gaussian kernel description.

3.6.3 Persistence Landscapes

Another systematic approach to producing features from persistence diagrams is provided by Bubenik's *persistence landscapes* [76]. Suppose that we are given

a barcode $\{[x_i, y_i)\}$, which we regard as a persistence diagram in $\mathbb{R}^2$. Changing coordinates via the transformation

$$[x, y) \mapsto \left[\frac{x+y}{2}, \frac{y-x}{2}\right),$$

we can equivalently represent a barcode as the multiset $\{[\frac{x_i+y_i}{2}, \frac{y_i-x_i}{2})\}$ in $\mathbb{R}^2$; we will assume that all persistence diagrams are represented in this format for the remainder of the section.

Next, define the piecewise-linear function

$$\Lambda_{(x,y)}(t) = \begin{cases} t - x, & t \in [x, \frac{x+y}{2}] \\ y - t, & t \in (\frac{x+y}{2}, y] \\ 0, & \text{otherwise.} \end{cases}$$

Definition 3.6.7. Let $B = \{[x_i, y_i)\}$ be a persistence diagram. The *persistence landscape* is the collection of functions $\lambda_B^k \colon \mathbb{R} \to \mathbb{R}$ for $k \in \mathbb{N}$, defined as

$$\lambda_B^k(t) = \lambda_B(k, t) = \text{kmax}_{[x_i,y_i)\in B}\, \Lambda_{[x_i,y_i)}(t),$$

where kmax denotes the kth largest value, defined to be 0 if the set in question contains fewer than k points. (We will often regard this collection as a single function $\Lambda \colon \mathbb{N} \times \mathbb{R} \to \mathbb{R}$.)

See Figure 3.23 for an example of a persistence landscape. One advantage of working with the persistence landscape is that for any fixed k this is a 1-Lipschitz function, and the set of all such functions is a $\mathbb{R}$-vector space with a metric induced by a norm that is complete and separable. As a consequence, one can easily define the mean landscape $\bar{\Lambda}$ for a collection of barcodes $\{B_i\}$, which is simply computed pointwise:

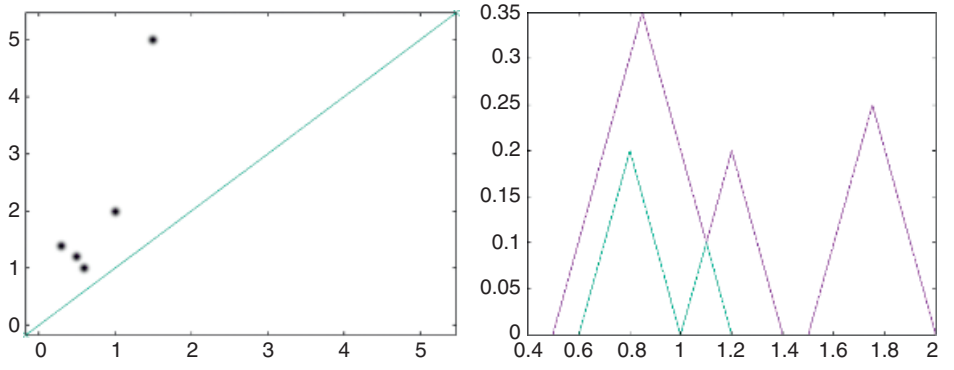

Figure 3.23 Left: A persistence diagram. Right: The associated persistence landscapes for $k = 1$ and $k = 2$.

$$\bar{\Lambda}_n = \frac{1}{n}\sum_{i=1}^{n} \lambda_{B_i}(k,t).$$

The mean landscape is the average value of the largest bar contained in k intervals. It is important to emphasize again that the mean landscape need not correspond to any particular barcode.

In this context, there is both a law of large numbers and a central limit theorem; these say that the mean of the landscapes of samples converges to the mean of the underlying distribution, and explain how fast this convergence occurs. Moreover, the average persistence landscape weakly converges to a Gaussian process (with a known rate of convergence) [76]. Specifically, we have the following result.

Theorem 3.6.8. *Provided that the expectation is finite,*

$$\bar{\Lambda} \to E(\Lambda),$$

where $\bar{\Lambda}_n$ is the empirical mean of the first n sample landscapes and $E(-)$ denotes the expected value.

Theorem 3.6.9. *Provided that the expectation and variance are both finite, then*

$$\sqrt{n}[\bar{\Lambda} - E(\Lambda)]$$

converges to a Gaussian random variable with the same covariance structure as Λ. (Here recall that the covariance structure determines the width of each Gaussian in the random variable.)

The following corollary allows us to perform inference.

Corollary 3.6.10. *The random variable produced by applying any functional (i.e., function from the space of landscapes to $\mathbb{R}$) also satisfies the central limit theorem.*

Of course, a choice of a useful and informative functional depends on the data and is not always evident. A simple approach is to use an indicator function for t in an interval $[-B, B]$ and k bounded by K.

Remark 3.6.11. In fact, we can prove a uniform version of the central limit theorem and bound the rate of convergence. This implies in particular that the bootstrap is asymptotically consistent and so can be used to estimate confidence intervals for persistence landscapes; see [109] for results of this form involving the multiplier bootstrap.

Furthermore, landscapes satisfy an evident analogue of the stability theorem: the L_∞ distance between landscapes is bounded by the Gromov-Hausdorff distance between point clouds.

A natural application of persistence landscapes to robust inference was studied in [110], where they used the average persistence landscape of the samples in Φ_k^n as a summary; this has the advantage of being easy to compute and study. In analogy with Theorem 3.4.2, one can show that the average persistence landscape is Wasserstein stable. Moreover, explicit estimates of the bias of this estimator as a function of the number of sample points can be obtained. (See Figures 3.24 and 3.25 for examples of this approach.)

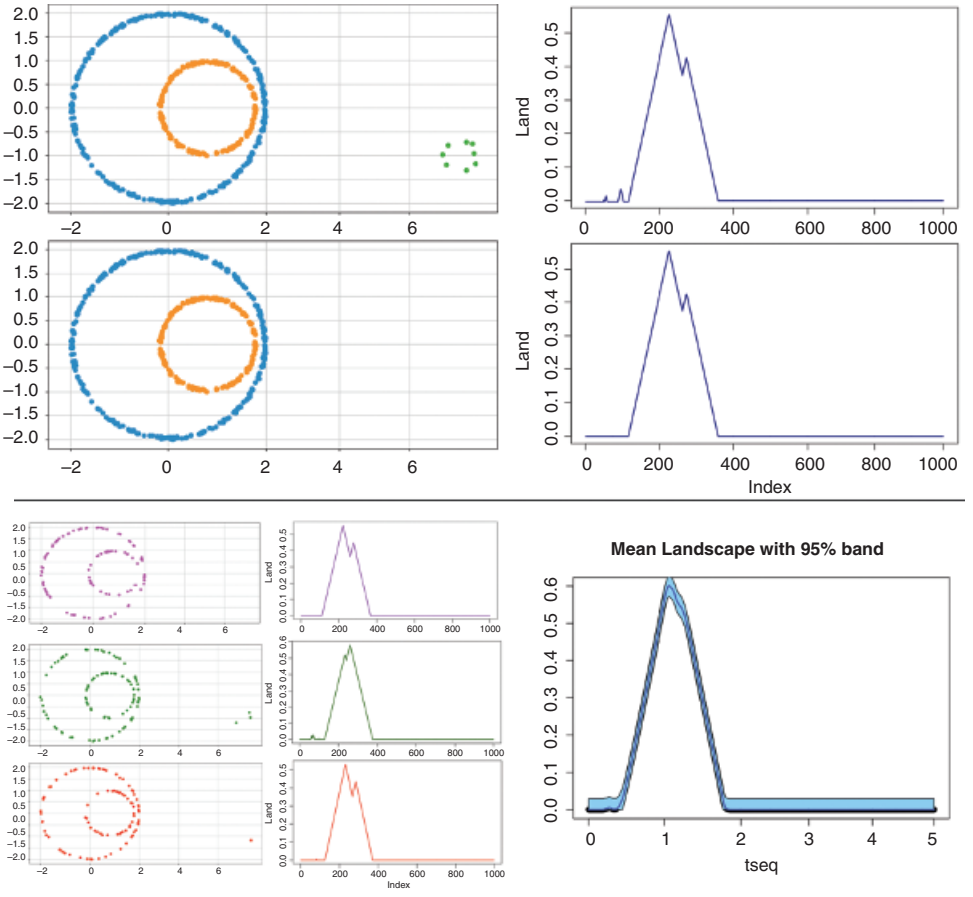

Figure 3.24 Since the landscape is a real-valued function, the pointwise average is easy to compute. The top two panels show the landscape for samples from two circles plus a noisy circle far away and two circles without the noisy circle. The bottom panels represent the effect of subsampling and averaging to remove the effect of the noisy circle.

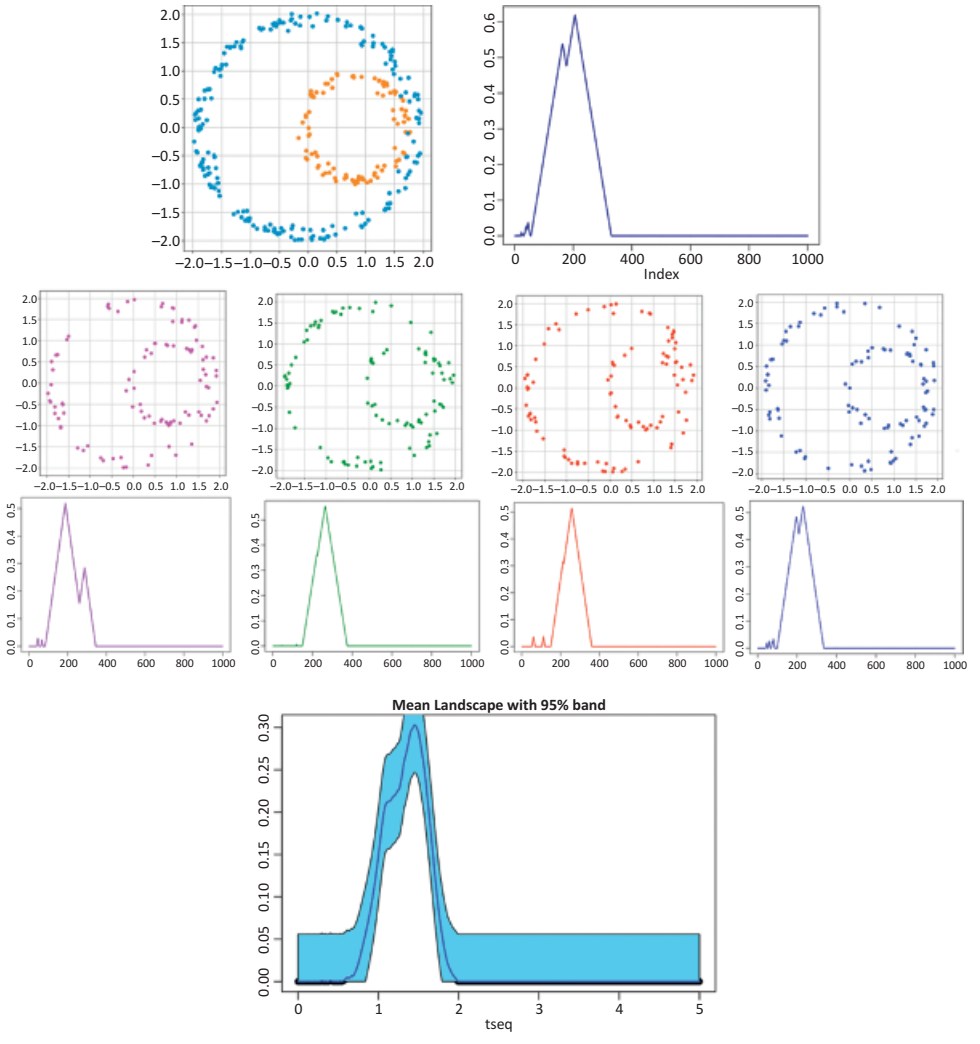

Figure 3.25 Subsampling and averaging is also effective with noisy data where the noise is concentrated around the underlying space.

3.6.4 Coordinates on Persistent Homology

A more principled source of real-valued invariants from barcodes comes from considerations from algebraic geometry. Adcock, Carlsson, and Carlsson introduced the idea of regarding subsets of barcode space as *algebraic varieties* and studying their coordinate rings [5]. *Coordinates* on a barcode just means a collection of functions from a space of barcodes to $\mathbb{R}$. In [5], the basic idea is to use symmetric polynomials in the start and endpoints of the bars, for barcodes with a fixed number

of bars. (The symmetry of the polynomials is a consequence of the fact that we do not care about the ordering of the bars within the barcode.)

Remark 3.6.12. This approach was extended to multidimensional persistence in [465].

Unfortunately, these coordinates are not stable with respect to perturbation of the barcode in the bottleneck or p-Wasserstein metric; this is clear, as very short bars with large start and endpoints can affect these polynomials dramatically. To fix this problem, Verovšek [283] (building on [94]) introduced ideas from *tropical geometry* to build stable coordinates. Tropical geometry studies a semiring structure on $\mathbb{R}$ where addition of x and y is computed by $\max(x, y)$ or $\min(x, y)$ and multiplication of x and y by $x + y$ (ordinary addition on real numbers). This is a semiring in the sense that we do not require every number to have an additive inverse.

The work of [283] showed that stable coordinates on barcode space could be obtained from rational functions (i.e., fractions) in "polynomial" expressions on the bar endpoints using the max-plus tropical structure. In [357], it is further shown that these coordinates provide sufficient statistics suitable for parametric inference; applications to reassortment in avian flu are discussed.

3.7 Stochastic Topology and the Expected Persistent Homology of Random Complexes

In the preceding sections, we have discussed techniques to produce stable persistent homology invariants of data despite the presence of noise. Another part of the statistical aspect of the story is to quantify the effect of idealized noise by describing the expected persistent homology of a "random complex." For example, such a description yields a family of strong null hypotheses. However, despite the mathematical interest and depth of theoretical work of this kind, in practice it is typically more suitable to use Monte Carlo simulation to find empirical estimates.

As a consequence, our discussion is brief and we refer the interested reader to the primary sources for precise theorem statements (see also Kahle's survey [281] and the article [61]).

In order to specify the problem, we need a model for generating random complexes. Recall that the Vietoris-Rips complex is completely determined by its 1-skeleton (see Definition 2.1.6), which is a graph. Therefore, processes that generate random graphs can also be regarded as producing random simplicial complexes.

The most familiar model of a random graph is the Erdös-Renyi model, which connects vertices with some fixed probability. However, although there is a

substantial literature on random simplicial complexes from this perspective (e.g., see [64] for a classic exposition), this is not a sensible model of random simplicial complexes in the geometric setting. The most relevant definition of a random complex from this perspective arises from the definition of a geometric random graph. (See [403] for an extensive treatment of the properties of geometric random graphs.)

Definition 3.7.1. Let (M, ∂_M, μ_M) be a metric measure space. Fix $\epsilon > 0$. A *geometric random graph* with k points is generated by sampling k points $\{x_i\}$ from M according to μ_M and forming the graph with k vertices and an edge (i, j) if $\partial_M(x_i, x_j) < \epsilon$.

Example 3.7.2. The most frequently studied example is the case when M is the unit cube $[0, 1]^n \subseteq \mathbb{R}^n$.

Definition 3.7.3. Let (M, ∂_M, μ_M) be a metric measure space. Fix $\epsilon > 0$. A *geometric random complex* with k points is generated by sampling k points $\{x_i\}$ from M according to μ_M and forming either the Vietoris-Rips or Čech complex associated to ϵ and the finite metric space $\{x_i\}$.

Although we have stated the definitions in full generality, most existing work studies distributions supported either on $\mathbb{R}^n$ or in a few cases on a smooth compact manifold embedded in $\mathbb{R}^n$ (e.g., see [62] for the latter).

Most current results (e.g., the work of Kahle) about geometric random complexes consider the expected ranks of the homology groups β_ℓ as simultaneously $\epsilon \to 0$ and $k \to \infty$. The results are controlled by $k\epsilon^n$:

1. in the *sub-critical* regime, $k\epsilon^n \to 0$,
2. in the *critical regime*, $k\epsilon^n$ goes to a constant, and
3. in the *super-critical regime*, $k\epsilon^n$ goes to ∞.

We now summarize what is known in these various settings.

1. **Sub-critical.** There are various results on the expected Betti numbers [282]. Here the situation is sometimes referred to as "dust," since there are many disconnected components and so the most important contribution is to H_0. This is the easiest non-trivial regime to analyze.
2. **Critical.** There is an enormous amount of non-trivial homology, and [543] provides detailed estimates on the expected rank of the homology for certain distributions on $\mathbb{R}^d$ and weak and strong laws of large numbers describing convergence.

3. **Super-critical.** The complex is asymptotically contractible and so there is no contribution to homology (and the analysis is basically trivial). This is analogous to the emergence of the "giant component" in the classical results on the behavior of random graphs.

A closely related but distinct perspective is provided by the work of Adler, Bobrowski, and Weinberger [7]. They consider distributions with infinite support on $\mathbb{R}^n$, and observe that sufficiently large samples separate into

- the "core," which is densely sampled and contractible, and
- the periphery, which "crackles" with homology.

This perspective is a variation on the results summarized above, insofar as the core and periphery correspond to super-critical and critical regimes simultaneously arising due to variation in the density.

The conceptual frameworks of "core" and "crackle" provide two kinds of indications of the limits of certain approaches to topological data analysis:

- a large core will obscure the signal, and
- the crackle will generate spurious homology classes.

All of the work discussed so far has focused on understanding homology for complexes with specific ϵ; only very recently has there been work extending this to persistent homology [63]. Here, there is more similarity between the regimes, but the scale of events differs. (See Figure 3.26 for a representative example.)

Notably, in the critical regime the longest bar in the barcode appears to satisfy a "law of the iterated logarithm" describing its length, for certain distributions on a cube (notably the Poisson distribution) and both the Čech and Vietoris-Rips complexes. Such a bound gives a precise estimate for how fast the length increases as the number n of sample points increases; roughly $\left(\frac{\log n}{\log\log n}\right)^{\frac{1}{k}}$ for kth homology. (This phenomenon is also mentioned in passing in the Adler-Bobrowski-Weinberger work.)

3.8 Euler Characteristics in Topological Data Analysis

A reasonable conclusion to draw from the discussion of this section is that it is advantageous to use the simplest possible topological invariants, e.g., low-dimensional persistent homology. This perspective suggests consideration of the Euler characteristic as a potentially interesting topological invariant which is robust and easy to compute and yet rich enough to capture topological properties of the underlying space.

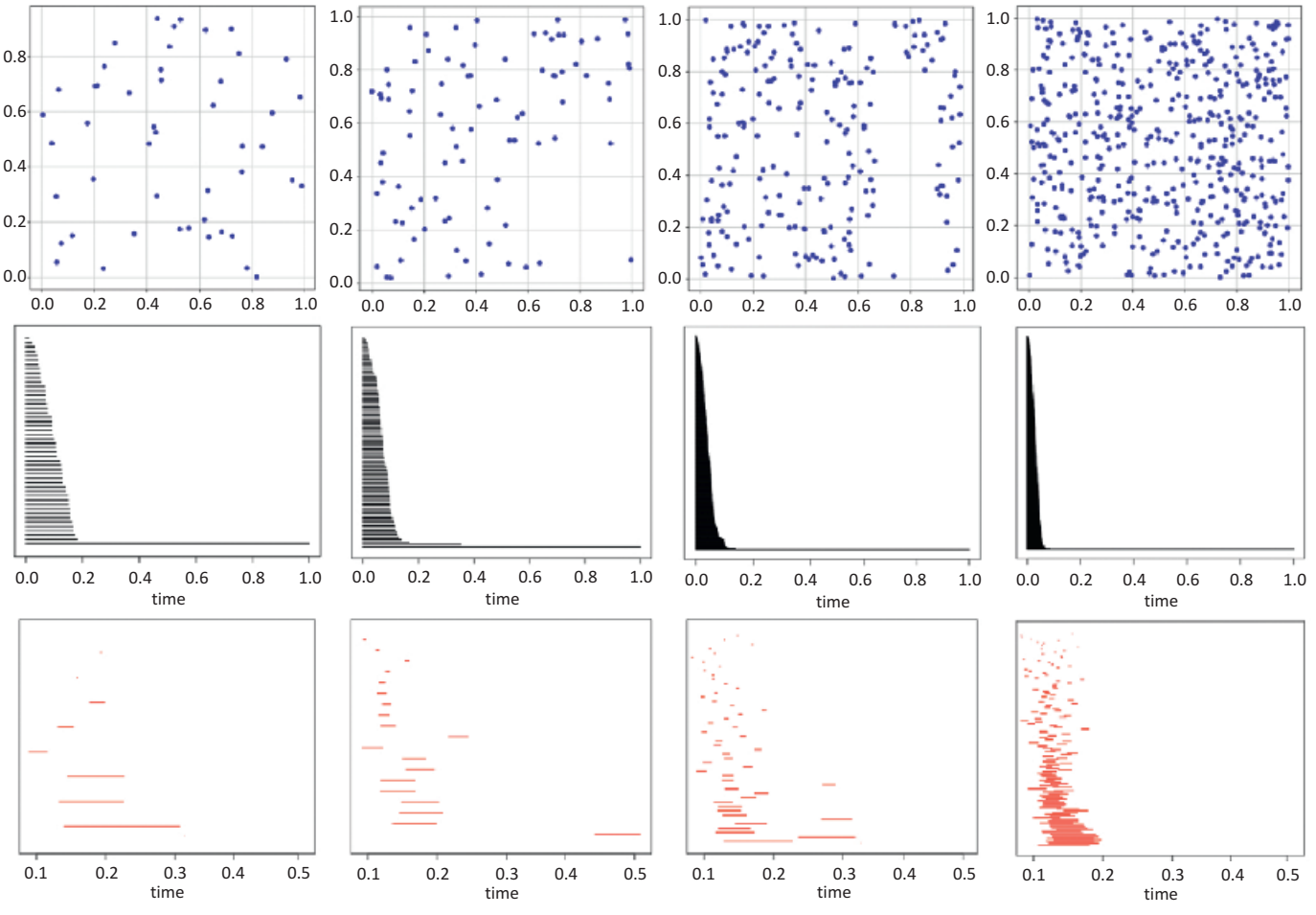

Figure 3.26 Persistent homology of points sampled uniformly from a unit square.

To further motivate this focus, Weinberger has pointed out that the Euler characteristic of a simplicial complex is *locally testable* [160]. Locally testable in this case means that the Euler characteristic can be computed from a small number of random samples from a simplicial complex, with high probability [204]. Specifically, fix $\epsilon > 0$. A tester for the Euler characteristic chooses $K(\epsilon)$ random vertices of the complex and has access to neighborhoods of size $D(\epsilon)$ around those vertices. The tester then returns a guess $\chi'(X)$ for the Euler characteristic such that

$$\Pr\left(\frac{\chi(X) - \chi'(X)}{|X_0|} \geq \epsilon\right) \leq \epsilon.$$

The existence of a tester is interesting because the functions K and D do not depend on the size of the complex but only on ϵ! Weinberger proposes that local testability is a good proxy for understanding when a topological invariant will be robust and reasonable to compute for small samples [533].

Remark 3.8.1. Although more generally rational homology groups are known to be locally testable [160], no such results are known for other coefficients.

There has been a great deal of study of the special case of the Euler characteristic of Gaussian random fields. Let M be a smooth compact manifold and f a Gaussian random field on M; then Adler and Taylor provide formulas describing the expected Euler characteristic of the "excursion sets" $f^{-1}(u, \infty)$. See [8] for an overview of this work, and [9] for an interpretation in terms of persistent homology. These kinds of results have had numerous applications in situations where smooth processes of this sort arise, notably imaging. However, application of these techniques in genomics is in its infancy, although searching for ways to apply them seems like a productive endeavor.

A potentially promising direction for problems related to genomics comes from the *smooth Euler characteristic transform*, a generalization of the persistent homology transform [127]. We again assume we are working with a finite simplicial complex M embedded in Euclidean space $\mathbb{R}^d$. For a given direction v, let a_v and b_v denote the minimum and maximum values of $x \cdot v$ over the points of M. The Euler characteristic curve in the direction v is now defined to be the function

$$[a_v, b_v] \to \mathbb{Z}$$

defined by $t \mapsto \chi(M(v)_t)$. Let $\bar{\chi}(M(v))$ denote the average value of the Euler characteristic curve in the direction v.

Definition 3.8.2. The *smoothed Euler characteristic curve* for the direction v is defined to be the function

$$F_v^M(y) = \int_{-\infty}^{y} (\chi(M(v)_x) - \bar{\chi}(M(v)))\,\mathrm{d}x.$$

Observe that by construction this is a smooth piecewise-linear function with compact support.

Definition 3.8.3. The *smooth Euler characteristic transform* is the function

$$\mathrm{SECT}\colon S^{d-1} \to L_2(\mathbb{R})$$

specified by

$$v \mapsto F_v^M.$$

Interestingly, when $d \leq 3$, the SECT can be shown to be injective; this is a sufficient statistic for describing the underlying distribution. Moreover, since the result is a function in L_2, just as in the case of the discussion of persistent landscapes, the SECT can be used as input to standard statistical models and resampling techniques can be used to obtain confidence intervals for predictors and summary statistics. This approach has been used to generate clinically meaningful conclusions from imaging data from glioblastoma tumors in [127].

3.9 Exploratory Data Analysis with Mapper

Because of the tremendous possible space of topological hypotheses, the framework of exploratory data analysis is very well suited for TDA. That is, rather than seeking to confirm specific hypotheses or test existing ideas about the data set, it is often much more sensible to simply attempt to find structure in the data.

The Mapper algorithm (as discussed in Section 2.8) is particularly well suited for this.

- The output of Mapper is a colored graph representing a multiscale clustering; it is often possible to visually interpret the results.
- As Mapper requires choices about bin sizes and filter functions, varying these allows us to explore structural properties of the data. For example, Mapper can account for the measure on the data by using a density estimator as the filter function.

Remark 3.9.1. Although Mapper output is not stable with regard to perturbation of these choices, in the exploratory paradigm this is not as substantial a problem as it might seem. One can use the same statistical tools normally used to assess

the stability of the results of clustering, i.e., cross-validation. There are different ways to do this, but all of them boil down to either partitioning or subsampling the data and then comparing clustering results by counting pairs which end up changing depending on whether they are in the same or different clusters. But perhaps more importantly, there is a strong sense in which instability is not as big an issue in genomics as one might expect. Exploratory analysis will typically be validated by further experiment. That is, in this kind of usage, predictions from TDA are confirmed by a follow-up experiment before being regarded as a reliable discovery. As such, the consequence of errors in inference due to instability is a wasted experiment; this is in stark contrast to applications in machine learning such as, for example, self-driving cars or clinical recommendations.

A common experimental application of Mapper is to explore various choices of filter function and other parameters in order to find clusterings of the data such that the clusters correlate strongly with other known properties of the data (e.g., clinically significant variables). More precisely, we have the following setup.

1. In addition to the data (X, ∂_X), filter function, and cover, we have an additional function $\theta\colon X \to \mathbb{R}$.
2. We extend θ to a function with domain the Mapper complex by defining θ on a point in the complex to be the average or median of the values of f along the corresponding data points.
3. We want to identify regions in the Mapper complex where θ is unusually large.

Now we can apply permutation tests (i.e., randomly relabeling the points and computing the values of the function θ) to determine the significance of an observed value. To be precise, we carry out the following.

1. We generate a distribution on values of θ by randomly shuffling the values of θ on X and recomputing the values on the points of the Mapper complex.
2. We then regard an actual value as significant if it is larger than 99% of the values produced in this fashion, for example. (The specific cutoff for significance is a parameter choice as usual.)

This procedure has been used in applications, for instance in the cell differentiation example we described previously in Example 2.8.3. However, note that as is usual with permutation tests, it can be expensive computationally to obtain confidence intervals as opposed to simply p-values. Also, the stability of this procedure does not yet have sound theoretical foundations in general, although in practice it appears to be stable with respect to cross-validation.

3.10 Summary

- This chapter provides tools with which we may formally discuss sampling from geometric objects. We adopt the working hypothesis that we have data randomly sampled from an underlying metric measure space (X, ∂_X, μ_X) (see Definition 3.2.10).
- In order to state probabilistic stability theorems, we need distances between distributions and more generally metric measure spaces. Toward this goal, we use the Gromov-Prohorov distance (see Definition 3.2.33) and the Gromov-Wasserstein distance (see Definition 3.2.34).
- We can study probability measures on barcode space; Section 3.3 provides a formal approach to probability theory on barcodes.
- Using metrics on distributions, Theorem 3.4.2 provides an analogue of the stability theorem of persistent homology (Theorem 2.4.10) in the context of metric measure spaces. Another version of a probabilistic stability theorem is given by Theorem 3.4.5.
- Section 3.5 provides a rigorous approach to this chapter's overarching goal of estimating persistent homology by taking sufficiently many samples from a space in order to recover the persistent homology of the support of the probability distribution.
- Summarizing distributions of barcodes turns out to be a challenging problem. One possibility is to consider techniques that involve extracting real-valued features from persistence diagrams.
- We may also approach this problem via *kernel methods* (see Section 3.6.2), *persistence landscapes* (see Section 3.6.3) or *coordinates* on a barcode (see Section 3.6.4); all of these methods map barcodes to a vector space where traditional statistical methods can be applied.
- In addition to the study of techniques to produce reliable persistent homology invariants despite the presence of noise, we are interested in considering the effect of idealized noise itself through the persistent homology of random complexes.
- Adaptation of the Euler characteristic is an attractive idea due to the advantages of using simple topological invariants.
- The Mapper algorithm (see Section 2.8) is a useful tool for exploratory data analysis. Section 3.9 outlines a procedure for the use of Mapper in applications.

The integration of topological data analysis with statistical methods is still in its infancy. As the discussion in the next part of the book makes clear, the kinds of techniques presented in this chapter have not yet made it into practice. Some of this is due to the lack of consensus about the best way to handle some of the issues that arise. But the lack of power of some of the tests (e.g., techniques for

estimating confidence intervals) combined with difficulties in producing topological summaries also provides a substantial impediment. We hope that the readers of this book will feel particularly motivated to work to develop standards for statistical practice in topological data analysis.

3.11 Suggestions for Further Reading

For background in probability theory, we recommend Billingsley's textbook [57]. For discussion of probability theory in non-positively curved metric measure spaces, Gromov's book [212] and Sturm's article [487] are very informative. However, in general, there are not yet any good survey articles or textbooks about probability theory in the context of topological data analysis; as an exception, Kahle's survey article on random complexes [281] is comprehensive. For a review of statistics, Wasserman's books [526, 527] provide good introductions, and Freedman's classic introduction to statistical modeling [184] teaches a healthy dose of skepticism about the power of statistical inference.