arXiv:1206.4581v2 [cs.CG] 17 Jan 2014

ROBUST STATISTICS, HYPOTHESIS TESTING, AND CONFIDENCE INTERVALS FOR PERSISTENT HOMOLOGY ON METRIC MEASURE SPACES

ANDREW J. BLUMBERG, ITAMAR GAL, MICHAEL A. MANDELL, AND MATTHEW PANCIA

Abstract. We study distributions of persistent homology barcodes associated to taking subsamples of a fixed size from metric measure spaces. We show that such distributions provide robust invariants of metric measure spaces, and illustrate their use in hypothesis testing and providing confidence intervals for topological data analysis.

1. Introduction

Topological data analysis assigns homological invariants to data presented as a finite metric space (a "point cloud"). If we imagine this data as measurements sampled from some abstract universal space X, the structure of that space is a metric measure space, having a notion both of distance between points and a notion of probability for the sampling. A standard homological approach to studying the samples is to assign a simplicial complex and compute its homology. The construction of the associated simplicial complex for a point cloud depends on a choice of scale parameter. The insight of "persistence" is that one should study homological invariants that encode change across scales; the correct scale parameter is a priori unknown. As such, a first approach to studying the homology of X from the samples is to simply compute the persistent homology $\mathrm{PH}_*(\tilde{X})$ of simplicial complexes associated to the sampled point cloud $\tilde{X}$.

We can gain some perspective from imagining that we could make measurements on X directly and interpret these measurements in terms of random sample points. With this in mind, we immediately notice some defects with homology and persistent homology as invariants of X. While the homology of X captures information about the global topology of the metric space, the probability space structure plays no role. This has bearing even if we assume X is a compact Riemannian manifold and the probability measure is the volume measure for the metric: handles which are small represent subsets of low probability but contribute to the homology in the same way as large handles. In this particular kind of example, persistent homology can identify this type of phenomenon (by encoding the scales at which homological features exist); however, in a practical context, the metric on the sample may be ad hoc (e.g., [10]) and less closely related to the probability measure. In this case, we could have handles that are medium size with respect to the metric but still low probability with respect to the measure. Homology and persistent homology

Date: November 26, 2013.
The authors were supported in part by DARPA YFA award N66001-10-1-4043.

have no mechanism for distinguishing low probability features from high probability features. A closely related issue is the effect of small amounts of noise (e.g., a situation in which a fraction of the samples are corrupted). A small proportion of bad samples can arbitrarily change the persistent homology. These two kinds of phenomena are linked, insofar as decisions about whether low probability features are noise or not is part of data analysis.

The disconnect with the underlying probability measure presents a significant problem when trying to adapt persistent homology to the setting of *hypothesis testing* and *confidence intervals*. Hypothesis testing involves making quantitative statements about the probability that the persistent homology computed from a sampling from a metric measure space is consistent with (or refutes) a hypothesis about the actual persistent homology. Confidence intervals provide a language to understand the variability in estimates introduced by the process of sampling. Because low probability features and a small proportion of bad samples can have a large effect on persistent homology computations, the persistent homology groups make poor test statistics for hypothesis testing and confidence intervals. To obtain useable test statistics, we need to develop invariants that better reflect the underlying measure and are less sensitive to large perturbation. To be precise about this, we use the statistical notion of robustness.

A statistical estimator is *robust* when its value cannot be arbitrarily perturbed by a constant proportion of bad samples. For instance, the sample mean is not robust, as a single extremely large sample value can dominate the result. On the other hand, the sample median is robust. As we discuss in Section 4, persistent homology is not robust. A small number of bad samples can cause large changes in the persistent homology, essentially as a reflection of the phenomenon of large metric low probability handles (including spurious ones).

In order to handle this, we adopt a standard statistical perspective, namely that the distribution of an estimator on some fixed finite number of samples is an appropriate way to grapple with such behavior. To make this precise, we need to be able to talk about probability distributions on homological invariants.

Using the idea of an underlying metric measure space X, formally the process of sampling amounts to considering random variables on the probability space $X^n = X \times \cdots \times X$ equipped with the product probability measure. The k-th persistent homology of a size n sample is a random variable on X^n taking values in the set $\mathcal{B}$ of finite *barcodes* [47], where a barcode is essentially a multiset of intervals of the form $[a, b)$. The set $\mathcal{B}$ of barcodes is equipped with a metric $d_{\mathcal{B}}$, the bottleneck metric [19], and we show in Section 3 that it is separable and that its completion $\overline{\mathcal{B}}$ is also a space of barcodes. Then $\overline{\mathcal{B}}$ is Polish, i.e., complete and separable, which makes it amenable to probability theory (see also [37] for similar results). In particular, various metrics on the set of distributions on $\overline{\mathcal{B}}$ metrize weak convergence, including the Prohorov metric d_{Pr} and the Wasserstein metric d_W. We consider the following probability distribution on barcode space $\overline{\mathcal{B}}$ (restated in Section 5 as Definition 5.1).

Definition 1.1. For a metric measure space (X, ∂_X, μ_X) and fixed $n, k \in \mathbb{N}$, define Φ_k^n to be the empirical measure induced by the kth n-sample persistent homology, i.e.,

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

the probability distribution on the set of barcodes $\overline{\mathcal{B}}$ induced by pushforward along PH_k from the product measure μ_X^n on X^n.

In other words, Φ_k^n is the probability measure on the space of barcodes where the probability of a subset A is the probability that a size n sample from X has k-th persistent homology landing in A. Note that the pushforward makes sense since PH_k is a continuous and hence Borel measurable function; see Section 5 for a discussion.

Although complicated, $\Phi_k^n(X)$ is a continuous invariant of X in the following sense. The moduli space of metric measure spaces admits a metric (in fact several) that combine the ideas of the Gromov-Hausdorff distance on compact metric spaces and weak convergence of probability measures [43]. We follow [30], and use the *Gromov-Prohorov metric*, d_{GPr}. We prove the following theorem in Section 5 (where it is restated as Theorem 5.2).

Theorem 1.2. *Let (X, ∂_X, μ_X) and $(X', \partial_{X'}, \mu_{X'})$ be compact metric measure spaces. Then we have the following inequality relating the Prohorov and Gromov-Prohorov metrics:*

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

This inequality becomes increasingly tight as the right-hand side approaches 0; we discuss precise estimates in Section 5. As we explain there, the fact that the bound increases with n is expected behavior: n should be thought of as a scale parameter, and increasing n yields a more sensitive invariant. The main import of Theorem 1.2 is that for fixed n, empirical approximations to $\Phi_k^n(X, \partial_X, \mu_X)$ computed from subsets $S \in X$ are asymptotically convergent as the number of samples increase.

Theorem 1.2 therefore validates computing Φ_k^n in practice using empirical approximations, where we are given a large finite sample S which we regard as drawn from X. Making S a metric measure space via the subspace metric from X and the empirical measure, we can compute $\Phi_k^n(S)$ as an approximation to $\Phi_k^n(X)$. This procedure is justified by the fact that as the sample size increases, the empirical metric converges (in d_{GPr}) to X; see Corollary 5.5. In particular, this justifies a resampling procedure to approximate Φ_k^n by subsampling from a large sample of size N. (We can also approximate using more sophisticated resampling methodology, a topic we study in future work.)

Moreover, as a consequence of the continuity implied by the previous theorem, we can use Φ_k^n to develop robust statistics: If we change X by adjusting the metric arbitrarily on ϵ probability mass to produce X', then the Gromov-Prohorov distance satisfies $d_{GPr}(X, X') \leq \epsilon$.

A difficulty with applying Φ_k^n is that it can be hard to interpret or summarize the information contained in a distribution of barcodes, unlike distributions of numbers for which there are various moments (e.g., the mean and the variance) which provide concise summaries of the distribution. One approach is to develop "topological summarizations" of distributions of barcodes; a version of this using Frechet means is explored in [45]. Another possibility is to embed the space of barcodes in a more tractable function space [5]. In this paper, we instead consider cruder invariants which take values in $\mathbb{R}$. One such invariant is the distance with respect to a reference distribution on barcodes $\mathcal{P}$, chosen to represent a hypothesis about the persistent homology of X.

Definition 1.3. Let (X, ∂_X, μ_X) be a compact metric measure space and let $\mathcal{P}$ be a fixed reference distribution on $\overline{\mathcal{B}}$. Fix $k, n \in \mathbb{N}$. Define the homological distance

on X relative to $\mathcal{P}$ to be

$$\mathrm{HD}_k^n((X,\partial_X,\mu_X),\mathcal{P}) = d_{Pr}(\Phi_k^n(X,\partial_X,\mu_X),\mathcal{P}).$$

We also consider a robust statistic MHD_k^n related to HD_k^n without first computing the distribution Φ_k^n. To construct MHD_k^n, we start with a reference barcode and compute the median distance to the barcodes of subsamples.

Definition 1.4. Let (X,∂_X,μ_X) be a compact metric measure space and fix a reference barcode $B \in \overline{\mathcal{B}}$. Fix $k,m \in \mathbb{N}$. Let $\mathcal{D}$ denote the distribution on $\mathbb{R}$ induced by applying $d_{\mathcal{B}}(B,-)$ to the barcode distribution $\Phi_k^n(X,\partial_X,\mu_X)$. Define the median homological distance relative to B to be

$$\mathrm{MHD}_k^n((X,\partial_X,\mu_X),B) = \mathrm{median}(\mathcal{D}).$$

Remark 1.5. The appearance of reference barcodes and distributions in the invariants above raises the question of where one obtains these quantities. As we illustrate below in Section 7, a common source of reference point is simply an a priori hypothesis about the data that we wish to test. The Frechet mean [45] of a collection of samples provides a more principled approach to producing such reference points.

The use of the median rather than the mean in the preceding definition ensures that we compute a robust statistic, in the following sense.

Definition 1.6. Let f be a function from finite metric spaces to a metric space (B,d). We say that f is robust with robustness coefficient $r > 0$ if for any non-empty finite metric space (X,∂), there exists a bound δ such that for any isometric embedding of X into a finite metric space (X',∂'), $|X'|/|X| < 1+r$ implies $d(f(X,\partial),f(X',\partial')) < \delta$, where $|X|$ denotes the number of elements of X.

For example, under the analogous definition on finite multi-subsets of $\mathbb{R}$ (in place of finite metric spaces), median defines a function to $\mathbb{R}$ that is robust with robustness coefficient $1-\epsilon$ for any ϵ since expanding a multi-subset X to a larger one X' with fewer than twice as many elements will not change the median by more than the diameter of X. Similarly, for a finite metric space X, expanding X to X', the proportion of n-element samples of X' which are samples of X is $(|X|/|X'|)^n$; when this number is more than $1/2$, the median value of any function f on the set of n-element samples of X' is then bounded by the values of f on n-element samples of X. Since $(N/(N+rN))^n > 1/2$ for $r < 2^{1/n}-1$, any such function f will be robust with robustness coefficient r satisfying this bound, and in particular for $r = (\ln 2)/n$.

Theorem 1.7. *For any $n,k,\mathcal{P}$, the function $\mathrm{MHD}_k^n(-,\mathcal{P})$ from finite metric spaces (with the uniform probability measure) to $\mathbb{R}$ is robust with robustness coefficient $> (\ln 2)/n$.*

The function Φ_k^n from finite metric spaces to distributions on $\overline{\mathcal{B}}$ and the function HD_k^n from finite metric spaces to $\mathbb{R}$ are robust for any robustness coefficient for trivial reasons since the Gromov-Prohorov metric is bounded. However, for these functions we can give explicit uniform estimates for how much these functions change when expanding X to X' just based on $|X'|/|X|$. We introduce the following notion of uniform robustness which is strictly stronger than the notion of robustness.

Definition 1.8. Let f be a function from finite metric spaces to a metric space (B,d). We say that f is uniformly robust with robustness coefficient $r > 0$ and

estimate bound δ if for any non-empty finite metric space (X, ∂) and any isometric embedding of (X, ∂) into a finite metric space (X', ∂'), $|X'|/|X| < 1 + r$ implies $d(f(X, \partial), f(X', \partial')) < \delta$.

Uniform robustness gives a uniform estimate on the change in the function from expanding the finite metric space. For example, the median function does not satisfy the analogous notion of uniform robustness for functions on finite multi-subsets of $\mathbb{R}$. We show in Section 5 that Φ_k^n and HD_k^n satisfy this stronger notion of uniform robustness.

Theorem 1.9. *For fixed n, k, Φ_k^n is uniformly robust with robustness coefficient r and estimate bound $nr/(1+r)$ for any r. For fixed $n, k, \mathcal{P}$, $\mathrm{HD}_k^n(-, \mathcal{P})$ is uniformly robust with robustness coefficient r and estimate bound $nr/(1+r)$ for any r.*

As with Φ_k^n itself, the law of large numbers and the convergence implied by Theorem 1.2 tells us that given a sufficiently large finite sample $S \subset M$, we can approximate HD_k^n and MHD_k^n of the metric measure space M in a robust fashion from the persistent homology computations on S. (See Corollaries 5.5, 6.3, and 6.6 below.)

In light of the results on robustness and asymptotic convergence, HD_k^n, MHD_k^n, and Φ_k^n (as well as various distributional invariants associated to Φ_k^n) provide good test statistics for hypothesis testing. Furthermore, one of the benefits of the numerical statistics HD_k^n and MHD_k^n is that we can use standard techniques to obtain confidence intervals, which provide a means for understanding the reliability of analyses of data sets. We discuss hypothesis testing and the construction of confidence intervals in Section 6, and explore examples in Sections 7 and 8. In this paper we primarily focus on analytic methods and Monte Carlo simulation for obtaining confidence intervals; however, these statistics are well-suited for the construction of resampling confidence intervals. In a follow-up paper [4] we establish the asymptotic consistency of the bootstrap for HD_k^n and MHD_k^n.

We regard this paper as a step towards providing a foundation for the integration of standard statistical methodology into computational algebraic topology. Our goal is to provide tools for practical use in topological data analysis.

Related work. We have developed an approach to using statistical tools to study persistent homological invariants for metric measure spaces accessed through finite samples. There are a number of related approaches to studying the statistical properties of persistent homological estimators; we quickly survey this work.

Bubenik [5] develops statistical inference via an embedding into function spaces called "persistence landscapes", and with various co-authors in [18, 6] studies an approach using Morse theory (and hence taking advantage of the ambient metric space for smoothing). The work of Harer, Mileyko, and Mukerjee in [37] parallels the development in Section 3 and introduces probability measures on barcode space, and these ideas are developed further (with Turner) in the context of Frechet means as ways of summarizing barcode distributions in [45].

In another direction, there has been a fair amount of work on the topological features of random simplicial complexes and noise due to Kahle [33, 34] as well as Adler, Bobrowski, Borman, Subag, and Weinberger [1, 2, 3]. This work is essential for understanding what persistent homological "null hypotheses" look like, and adapted to our setting should inform our statistical inference procedures.

Finally, there has also been a lot of excellent work arising on studying robustness in the context of understanding distances to measures for point clouds. This approach was introduced by Chazal, Cohen-Steiner, and Merigot in [13], and was further developed by Caillerie, Chazal, Dedecker, and Michel in [9]. The basic idea is that the distribution of distances to a point cloud is a robust invariant of the point cloud; indeed, this is closely related to the $n = 2$ case of our central invariant. Since preservation of explicit distances is a goal of this approach, it is more closely related to rigid geometric inference (and manifold learning) than purely topological inference, as in our homological approach.

Acknowledgments. The authors would like to thank Gunnar Carlsson and Michael Lesnick for useful comments, Rachel Ward for comments on a prior draft, and Olena Blumberg for help with background and for assistance with the analysis of the tightness of the main theorem. We would also like to thank the Institute for Mathematics and its Applications for hospitality while revising this paper.

Outline. The paper is organized as follows. In Section 2, we provide a rapid review of the necessary background on simplicial complexes, persistent homology, and metric measure spaces. In Section 3, we study the space of barcodes, establishing foundations needed to work with distributions of barcodes. In Section 4, we discuss the robustness of persistent homology. In Section 5, we study the properties of Φ_k^n, MHD_k^n, and HD_k^n and prove Theorem 1.2. We discuss hypothesis testing and confidence intervals in Section 6, which we illustrate with synthetic examples in Section 7. Section 8 applies these ideas to the analysis of the natural images data in [10].

2. Background

In this section we provide background for the framework for topological data analysis we study in this paper. We focus on an approach which accesses the ambient metric measure space (X, ∂_X, μ_X) only through finite samples, i.e., point clouds.

2.1. Simplicial complexes associated to point clouds.

A standard approach in computational algebraic topology proceeds by assigning a simplicial complex (which usually also depends on a scale parameter ϵ) to a finite metric space (X, ∂). Recall that a simplicial complex is a combinatorial model of a topological space, defined as a collection of nonempty finite sets $\mathcal{Z}$ such that for any set $Z \in \mathcal{Z}$, every nonempty subset of Z is also in $\mathcal{Z}$. Associated to such a simplicial complex is the "geometric realization", which is formed by gluing standard simplices of dimension $|Z| - 1$ via the subset relations. (The standard n-simplex has $n + 1$ vertexes.) The most basic and widely used construction of a simplicial complex associated to a point cloud is the Vietoris-Rips complex:

Definition 2.1. For $\epsilon \in \mathbb{R}$, $\epsilon \geq 0$, the Vietoris-Rips complex $\mathrm{VR}_\epsilon(X)$ is the simplicial complex with vertex set X such that $[v_0, v_1, \ldots, v_n]$ is an n-simplex when for each pair v_i, v_j, the distance $\partial(v_i, v_j) \leq \epsilon$.

The Vietoris-Rips complex is determined by its 1-skeleton. The construction is functorial in the sense that for a continuous map $f\colon X \to Y$ with Lipshitz constant

κ and for $\epsilon \leq \epsilon'$, there is a commutative diagram

$$\begin{tikzcd} \mathrm{VR}_\epsilon(X) \arrow[r] \arrow[d] & \mathrm{VR}_{\kappa\epsilon}(Y) \arrow[d] \\ \mathrm{VR}_{\epsilon'}(X) \arrow[r] & \mathrm{VR}_{\kappa\epsilon'}(Y). \end{tikzcd} \tag{2.2}$$

The Vietoris-Rips complex is easy to compute, in the sense that is straightforward to determine when a simplex is in the complex. More closely related to classical constructions in algebraic topology is the Cech complex.

Definition 2.3. For $\epsilon \in \mathbb{R}$, $\epsilon \geq 0$, the Cech complex $C_\epsilon(X)$ is the simplicial complex with vertex set X such that $[v_0, v_1, \ldots, v_n]$ is an n-simplex when the intersection

$$\bigcap_{0 \leq i \leq n} B_{\frac{\epsilon}{2}}(v_i)$$

is non-empty, where here $B_r(x)$ denotes the r-ball around x.

The Cech complex has analogous functoriality properties to the Vietoris-Rips complex. The Cech complex associated to a cover of a paracompact topological space satisfies the nerve lemma: if the cover consists of contractible spaces such all finite intersections are contractible or empty, the resulting simplicial complex is homotopy equivalent to the original space.

Remark 2.4. Both the Vietoris-Rips complex and the Cech complex can be unmanageably large; e.g., for a set of points $Y = \{y_1, y_2, \ldots, y_n\}$ such that $\partial(y_i, y_j) \leq \epsilon$, every subset of Y specifies a simplex of the Vietoris-Rips complex. As a consequence, it is often very useful to define complexes with the vertices restricted to a small set of landmark points; the weak witness complex is perhaps the best example of such a simplicial complex [42]. We discuss this construction further in Section 8, as it is important in the applications.

The theory we develop in this paper is relatively insensitive to the specific details of the construction of a simplicial complex associated to a finite metric space (and scale parameter). For reasons that will become evident when we discuss persistence in Subsection 2.3 below, the main thing we require is a procedure for assigning a complex to $((M, \partial), \epsilon)$ that is functorial in the vertical maps of diagram (2.2) for $\kappa = 1$.

2.2. **Homological invariants of point clouds.** In light of the previous subsection, given a metric space (X, ∂), one defines the homology at the feature scale ϵ to be the homology of a simplicial complex associated to (X, ∂); e.g., $H_*(\mathrm{VR}_\epsilon(X))$ or $H_*(C_\epsilon(X))$. This latter definition is supported by the following essential consistency result, which is in line with the general philosophy that we are studying an underlying continuous geometric object via finite sets of samples.

Theorem 2.5 (Niyogi-Smale-Weinberger [39])**.** *Let (M, ∂) be a compact Riemannian manifold equipped with an isometric embedding $\gamma\colon M \to \mathbb{R}^n$, and let $X \subset M$ be a finite independent identically-distributed sample drawn according to the volume measure on M. Then for any $p \in (0,1)$, there are constants δ (which depends on the curvature of M and the embedding γ) and $N_{\delta,p}$ such that if $\epsilon < \delta$ and $|X| > N_{\delta,p}$ then the probability that $H_*(C_\epsilon(X)) \cong H_*(M)$ is an isomorphism is $> p$.*

In fact, Niyogi, Smale, and Weinberger prove an effective version of the previous result, in the sense that there are explicit numerical bounds dependent on p and a "condition number" which incorporates data about the curvature of M and the twisting of the embedding γ.

Work by Latschev provides an equivalent result for $\mathrm{VR}_\epsilon(X)$, with somewhat worse bounds, defined in terms of the injectivity radius of M [35]. Alternatively, one can show that in the limit $\mathrm{VR}_\epsilon(X)$ captures the homotopy type of the underlying manifold using the fact that there are inclusions

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

While reassuring, an unsatisfactory aspect of the preceding results is the dependence on a priori knowledge of the feature scale ϵ and the details of the intrinsic curvature of M and the nature of the embedding. A convenient way to handle the fact that it is often hard to know a good choice of ϵ at the outset is to consider multi-scale homological invariants that encode the way homology changes as ϵ varies. This leads us to the notion of persistent homology.

2.3. **Persistent homology.** Persistent homology arose more or less simultaneously and independently in work of Robins [41], Frosini and Ferri and collaborators [28, 8], and Edelsbrunner and collaborators [26]. See the excellent survey of Edelsbrunner and Harer [25] for a more expansive discussion of the history and development of these ideas. The efficient algorithms and the algebraic presentation we apply herein is due to [26] and [47].

Given a diagram of simplicial complexes indexed on $\mathbb{R}$, i.e., a complex X_s for each $s \in \mathbb{R}$ and maps $X_s \to X_{s'}$ for $s \leq s'$, there are natural maps $H_*(X_s) \to H_*(X_{s'})$ induced by functoriality.

We say that a class $\alpha \in H_p(X_i)$ is *born* at time i if it is not in the image of $H_k(X_j)$ for $j < i$, and we say a class $\alpha \in H_k(X_i)$ *dies* at time i if the image of α is 0 in $H_k(X_j)$ for $j \geq i$. This information about the homology can be packaged up into an algebraic object:

Definition 2.6. Let $\{X_i\}$ be a diagram of simplicial complexes indexed on $\mathbb{R}$. The pth persistent kth homology group of X_i is defined to be

$$H_{k,p}(X_i) = Z_k^i/(B_k^{i+p} \cap Z_k^i),$$

where Z and B denote the cycle and boundary groups respectively. Alternatively, $H_{k,p}(X_i)$ is the image of the natural map

$$H_k(X_i) \to H_k(X_{i+p}).$$

Barcodes provide a convenient reformulation of information from persistent homology. Although we will work over a field and in the presence of suitable finiteness hypotheses which are satisfied in our motivating examples, recent work makes it clear that this restriction could be weakened [14, 7]. We assume that the values $H_*(X_i)$ change only at a countable discrete subset of $\mathbb{R}$, so that by reindexing we have a direct system

$$X_0 \to X_1 \to \cdots \to X_n \to \cdots,$$

the direct system of simplicial complexes stabilizes at a finite stage and all homology groups are finitely-generated. Then a basic classification result of Zomorodian-Carlsson [47] describes the persistent homology in terms of a barcode, a multiset of non-empty intervals of the form $[a, b) \subset \mathbb{R}$. An interval in the barcode indicates

the birth and death of a specific homological feature. For reasons we explain below, the barcodes appearing in our context will always have finite length intervals.

The Vietoris-Rips (or Cech) complexes associated to a point cloud (X, ∂_X) fit into this context by looking at a sequence of varying values of ϵ:

$$\mathrm{VR}_{\epsilon_1}(X) \to \mathrm{VR}_{\epsilon_2}(X) \to \cdots .$$

We can do this in several ways, for example, using the fact that the Vietoris-Rips complex changes only at discrete points $\{\epsilon_i\}$ and stabilizes for sufficiently large ϵ, or just choosing and fixing a finite sequence ϵ_i independently of X. The theory we present below makes sense for either of these choices, and we use the following notation.

Notation 2.7. Let (X, ∂_X) be a finite metric space. For $k \in \mathbb{N}$, denote the persistent homology of X as

$$\mathrm{PH}_k((X, \partial_X)) = \mathrm{PH}_{k,p}(\{VR_{\epsilon_{(-)}}(X)\})$$

for some chosen sequence $0 < \epsilon_1 < \epsilon_2 < \cdots$ and $p \geq 0$.

More generally, we can make analogous definitions for any functor

$$\Psi \colon \mathcal{M} \times \mathbb{R}_{>0} \to \mathrm{sComp},$$

where $\mathcal{M}$ is the category of finite metric spaces and metric maps and sComp denotes the category of simplicial complexes. We will call such a Ψ "good" when the homology changes for only finitely values in $\mathbb{R}$. In this case, we can choose the directed system of values of ϵ_i to contain these transition values.

We note that for large values of the parameter ϵ, $VR_\epsilon(X)$ will be contractible. Therefore, if we use the reduced homology group in dimension 0, we get $H_k(VR_\epsilon) = 0$ for all k for large ϵ. The barcodes associated to these persistent homologies therefore have only finite length bars. For convenience in computation, we typically cut off ϵ at a moderately high value before this breakdown occurs. The result is a truncation of the barcode to the cut-off point.

2.4. **Gromov-Hausdorff stability and the bottleneck metric.** By work of Gromov, the set of isometry classes of compact metric spaces admits a useful metric structure, the Gromov-Hausdorff metric. For a pair of finite metric spaces (X_1, ∂_1) and (X_2, ∂_2), the Gromov-Hausdorff distance is defined as follows: For a compact metric space (Z, ∂) and closed subsets $A, B \subset Z$, the Hausdorff distance is defined to be

$$d_H^Z(A, B) = \max(\sup_{a \in A} \inf_{b \in B} \partial(a, b), \sup_{b \in B} \inf_{a \in A} \partial(a, b)).$$

One then defines the Gromov-Hausdorff distance between X_1 and X_2 to be

$$d_{GH}(X_1, X_2) = \inf_{Z, \gamma_1, \gamma_2} d_H^Z(X_1, X_2),$$

where here $\gamma_1 \colon X_1 \to Z$ and $\gamma_2 \colon X_2 \to Z$ are isometric embeddings.

Since the topological invariants we are studying ultimately arise from finite metric spaces, a natural question to consider is the degree to which point clouds that are close in the Gromov-Hausdorff metric have similar homological invariants. This question does not in general have a good answer in the setting of the homology of the point cloud, but in the context of persistent homology, Chazal, et al. [16, 3.1] provide a seminal theorem in this direction that we review as Theorem 2.9 below.

The statement of Theorem 2.9 involves a metric on the set of barcodes called the bottleneck distance and defined as follows. Recall that a barcode $\{I_\alpha\}$ is a multiset of non-empty intervals. Given two non-empty intervals $I_1 = [a_1, b_1)$ and $I_2 = [a_2, b_2)$, define the distance between them to be

$$d_\infty(I_1, I_2) = ||(a_1, b_1) - (a_2, b_2)||_\infty = \max(|a_1 - a_2|, |b_1 - b_2|).$$

We also make the convention

$$d_\infty([a, b), \emptyset) = |b - a|/2$$

for $b > a$ and $d_\infty(\emptyset, \emptyset) = 0$. For the purposes of the following definition, we define a matching between two barcodes $B_1 = \{I_\alpha\}$ and $B_2 = \{J_\beta\}$ to be a multi-subset C of the underlying set of

$$(B_1 \cup \{\emptyset\}) \times (B_2 \cup \{\emptyset\})$$

such that C does not contain $(\emptyset, \emptyset)$ and each element I_α of B_1 occurs as the first coordinate of an element of C exactly the number of times (counted with multiplicity) of its multiplicity in B_1, and likewise for every element of B_2. We get a more intuitive but less convenient description of a matching using the decomposition of $(B_1 \cup \{\emptyset\}) \times (B_2 \cup \{\emptyset\})$ into its evident four pieces: The basic data of C consists of multi-subsets $A_1 \subset B_1$ and $A_2 \subset B_2$ together with a bijection (properly accounting for multiplicities) $\gamma\colon A_1 \to A_2$; C is then the (disjoint) union of the graph of γ viewed as a multi-subset of $B_1 \times B_2$, the multi-subset $(B_1 - A_1) \times \{\emptyset\}$ of $B_1 \times \{\emptyset\}$, and the multi-subset $\{\emptyset\} \times (B_2 - A_2)$ of $\{\emptyset\} \times B_2$. With this terminology, we can define the bottleneck distance.

Definition 2.8. The bottleneck distance between barcodes $B_1 = \{I_\alpha\}$ and $B_2 = \{J_\beta\}$ is

$$d_{\mathcal{B}}(B_1, B_2) = \inf_C \sup_{(I,J)\in C} d_\infty(I, J),$$

where C varies over all matchings between B_1 and B_2.

Although expressed slightly differently, this agrees with the bottleneck metric as defined in [19, §3.1] and [16, §2.2]. On the set of barcodes $\mathcal{B}$ with finitely many finite length intervals, $d_{\mathcal{B}}$ is obviously a metric. More generally, for any $p > 0$, one can consider the ℓ^p version of this metric,

$$d_{B,p}(B_1, B_2) = \inf_C \Big(\sum_{(I,J)\in C} d_\infty(I, J)^p \Big)^{1/p}.$$

For simplicity, we focus on $d_{\mathcal{B}}$ in this paper, but analogues of our main theorems apply to these variant metrics as well.

We have the following essential stability theorem:

Theorem 2.9 (Chazal, et. al. [16, 3.1])**.** *For each k, we have the bound*

$$d_{\mathcal{B}}(\mathrm{PH}_k(X), \mathrm{PH}_k(Y)) \leq d_{GH}(X, Y).$$

Note that truncating barcodes (i.e., truncating each persistent interval) is a Lipshitz map $\mathcal{B} \to \mathcal{B}$ with Lipshitz constant 1, so the bound above still holds when we use a large parameter cut-off in defining PH_k.

Remark 2.10. The space of barcodes admits other metrics that are finer than the bottleneck metric for which versions of the stability theorem also hold; these can be useful in practical situations. Notably, the papers [37, 22, 45] study and apply a

family of Wasserstein (mass transportation) metrics on barcode space. We believe that our results can be extended to this setting.

2.5. **Metric measure spaces and the Gromov-Prohorov distance.** To establish more robust convergence results, we work with suitable metrics on the set of compact metric measure spaces. Specifically, following [30, 36, 43] we use the idea of the Gromov-Hausdorff metric to extend certain standard metrics on distributions (on a fixed metric measure space) to a metric on the set of all compact metric measure spaces.

A basic metric of this kind is the Gromov-Prohorov metric [30]. (For the following formulas, see Section 5 of [30] and its references.) This metric is defined in terms of the standard Prohorov metric d_{Pr} (metrizing weak convergence of probability distributions on separable metric spaces). First, recall that for measures μ_1 and μ_2 on a metric space Z, the Prohorov metric is defined as

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

where $A \subset Z$ varies over all closed sets and $B_\epsilon(A)$ is the set of points z such that $d_Z(z, a) < \epsilon$ for some $a \in A$. Then the Gromov-Prohorov metric is defined as

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

where the inf is computed over all isometric embeddings $\phi_X \colon X \to Z$ and $\phi_Y \colon Y \to Z$ into a target metric space (Z, ∂_Z).

It is very convenient to reformulate both the Gromov-Hausdorff and Gromov-Prohorov distances in terms of relations. For sets X and Y, a relation $R \subset X \times Y$ is a correspondence if for each $x \in X$ there exists at least one $y \in Y$ such that $(x, y) \in R$ and for each $y' \in Y$ there exists at least one $x' \in X$ such that $(x', y') \in R$. For a relation R on metric spaces (X, ∂_X) and (Y, ∂_Y), we define the distortion as

$$\operatorname{dis}(R) = \sup_{(x,y),(x',y') \in R} |\partial_X(x, x') - \partial_Y(y, y')|.$$

The Gromov-Hausdorff distance can be expressed as

$$d_{GH}((X, \partial_X), (Y, \partial_Y)) = \frac{1}{2} \inf_R \operatorname{dis}(R),$$

where we are taking the infimum over all correspondences $R \subset X \times Y$.

Similarly, we can reformulate the Prohorov metric as follows. Given two measures μ_1 and μ_2 on a metric space X, let a coupling of μ_1 and μ_2 be a measure ψ on $X \times X$ (with the product metric) such that $\psi(X \times -) = \mu_2$ and $\psi(- \times X) = \mu_1$. Then we have

$$d_{Pr}(\mu_1, \mu_2) = \inf_\psi \inf\{\epsilon > 0 \mid \psi\{(x, x') \in X \times X \mid \partial_X(x, x') \geq \epsilon\} \leq \epsilon\}.$$

This characterization of the Prohorov metric turns out to be useful when working with the Gromov-Prohorov metric in light of the (trivial) observation that if $d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) < \epsilon$ then there exists a metric space Z and embeddings $\iota_1 \colon X \to Z$ and $\iota_2 \colon Y \to Z$ such that $d_{Pr}((\iota_1)_*\mu_X, (\iota_2)_*\mu_Y) < \epsilon$.

3. Probability measures on the space of barcodes

This section introduces the spaces of barcodes $\mathcal{B}_N$ and $\overline{\mathcal{B}}$ used in the distributional invariants Φ_k^n of Definition 1.1. These spaces are complete and separable under the bottleneck metric. This implies in particular that the Prohorov metric

on the set of probability measures in $\mathcal{B}_N$ or $\overline{\mathcal{B}}$ metrizes convergence in probability, which justifies the perspective in the stability theorem 1.2 and the definition of the invariants $\mathrm{HD}_k^n(-,\mathcal{P})$ in Definition 1.3.

A barcode is by definition a multi-set of intervals, in our case of the form $[a,b)$ for $0 \leq a < b < \infty$. The set $\mathcal{I}$ of all intervals of this form is of course in bijective correspondence with a subset of $\mathbb{R}^2$. A multi-set A of intervals is a multi-subset of $\mathcal{I}$, which concretely is a function from $\mathcal{I}$ to the natural numbers $\mathbb{N} = \{0,1,2,3,\dots\}$ which counts the number of multiples of each interval in A. We denote by $|A|$ the cardinality of A, which we define as the sum of the values of the function $\mathcal{I} \to \mathbb{N}$ specified by A (if finite, or countably or uncountably infinite, if not). The space $\mathcal{B}$ of barcodes of the introduction is the set of multi-sets of intervals A such that $|A| < \infty$. We have the following important subsets of $\mathcal{B}$.

Definition 3.1. For $N \geq 0$, let $\mathcal{B}_N$ denote the set of multi-sets of intervals (in $\mathcal{I}$) A with $|A| \leq N$.

The main result on $\mathcal{B}_N$ is the following theorem, proved below. (Similar results can also be found in [37].)

Theorem 3.2. *For each $N \geq 0$, $\mathcal{B}_N$ is complete and separable under the bottleneck metric.*

Since the homology H_k (with any coefficients) of any complex with n vertices can have rank at most $\binom{n}{k+1}$, our persistent homology barcodes will always land in one of the $\mathcal{B}_N$, with N depending just on the size of the samples. As we let the size of the samples increase, N may increase, and so it is convenient to have a target independent of the number of samples. The space $\mathcal{B} = \bigcup \mathcal{B}_N$ is clearly not complete under the bottleneck metric (consider a sequence of barcodes $\{X_n\}$ such that X_n is produced from X_{n-1} by adding a bar $(0, \frac{1}{n})$), so we introduce the following space of barcodes $\overline{\mathcal{B}}$.

Definition 3.3. Let $\overline{\mathcal{B}}$ be the space of multi-sets A of intervals (in $\mathcal{I}$) with the property that for every $\epsilon > 0$, the multi-subset of A of those intervals of length more than ϵ has finite cardinality.

Clearly barcodes in $\overline{\mathcal{B}}$ have at most countable cardinality, and the bottleneck metric extends to a pseudo-metric $d_{\mathcal{B}} \colon \overline{\mathcal{B}} \times \overline{\mathcal{B}} \to \mathbb{R}$. The following lemma shows it is a metric.

Lemma 3.4. *For $X, Y \in \overline{\mathcal{B}}$, $d_{\mathcal{B}}(X,Y) = 0$ only if $X = Y$.*

Proof. Let $X, Y \in \overline{\mathcal{B}}$ with $d_{\mathcal{B}}(X,Y) = 0$ and assume without loss of generality that X is not in $\mathcal{B}_N$ for any N. Then the possible distinct lengths of intervals in X or Y form a countable set $\ell_0 > \ell_1 > \cdots$. Let X_i and Y_i denote the multisubsets of X and Y consisting of the intervals of length exactly ℓ_i. Let $\epsilon_0 < (\ell_0 - \ell_1)/2$ and in general let

$$\epsilon_i < \min(\epsilon_0, \dots, \epsilon_{i-1}, (\ell_i - \ell_{i+1})/2)$$

(with each $\epsilon_i > 0$). For any n and any $0 < \epsilon < \epsilon_n$, any matching C of X and Y with

$$d_C(X,Y) = \sup_{(I,J) \in C} d_\infty(I,J) < \epsilon$$

must induce a bijection between X_i and Y_i for all $i \leq n$; moreover, if C_i denotes the restriction of C to a matching of X_i and Y_i, $d_{C_i}(X_i, Y_i) < \epsilon$. Letting ϵ go to zero, we see that $X_i = Y_i$ for all i and that $X = Y$. $\square$

Lemma 3.4 implies that $d_{\mathcal{B}}$ extends to a metric on $\overline{\mathcal{B}}$. We prove the following theorem.

Theorem 3.5. *$\overline{\mathcal{B}}$ is the completion of $\mathcal{B} = \bigcup \mathcal{B}_N$ in the bottleneck metric. In particular $\overline{\mathcal{B}}$ is complete and separable in the bottleneck metric.*

Proof of Theorems 3.2 and 3.5. The multi-sets of intervals with rational endpoints provides a countable dense subset for $\mathcal{B}_N$. To see that $\mathcal{B}$ is dense in $\overline{\mathcal{B}}$, given A in $\overline{\mathcal{B}}$ and $\epsilon > 0$, let A_ϵ be the multi-subset of A of those intervals of length $> \epsilon$. Then by definition of $\overline{\mathcal{B}}$, A_ϵ is in $\mathcal{B}$, and by definition of the bottleneck metric, using the matching coming from the inclusion of A_ϵ in A, we have that

$$d_{\mathcal{B}}(A, A_\epsilon) \leq \epsilon/2 < \epsilon.$$

It just remains to prove completeness of $\mathcal{B}_N$ and $\overline{\mathcal{B}}$. For this, given a Cauchy sequence $\langle X_n \rangle$ in $\mathcal{B}$ it suffices to show that X_n converges to an element X in $\overline{\mathcal{B}}$ and that X is in $\mathcal{B}_N$ if all the X_n are in $\mathcal{B}_N$.

Let $\langle X_n \rangle$ be a Cauchy sequence in $\mathcal{B}$. By passing to a subsequence if necessary, we can assume without loss of generality that for $n, m > k$, $d_{\mathcal{B}}(X_m, X_n) < 2^{-(k+2)}$. For each n, we have $d_{\mathcal{B}}(X_n, X_{n+1}) < 2^{-(n+1)}$; choose a matching C_n such that $d_\infty(I, J) < 2^{-(n+1)}$ for all $(I, J) \in C_n$. For each n, define a finite sequence of intervals $I_1^n, \ldots, I_{k_n}^n$ inductively as follows. Let $k_0 = 0$. Let k_1 be the cardinality of the multi-subset of X_1 consisting of those intervals of length > 1, and let $I_1^1, \ldots, I_{k_1}^1$ be an enumeration of those intervals. By induction, $I_1^n, \ldots, I_{k_n}^n$ is an enumeration of the intervals in X_n of length $> 2^{-n+1}$ such that for $i \leq k_{n-1}$, the intervals I_i^{n-1} and I_i^n correspond under the matching C_{n-1}. For the inductive step, we note that if I_i^n corresponds to J under C_n, then $d_\infty(I_i^n, J) < 2^{-(n+1)}$, so the length $||J||$ of J is bigger than $||I_i^n|| - 2^{-n}$, and

$$||J|| > 2^{-n+1} - 2^{-n} = 2^{-n} = 2^{-(n+1)+1}.$$

Thus, we can choose I_i^{n+1} to be the corresponding interval J for $i \leq k_n$, and we can choose the remaining intervals of length $> 2^{-(n+1)+1}$ in an arbitrary order. Write $I_i^n = [a_i^n, b_i^n)$ and let

$$a_i = \lim_{n\to\infty} a_i^n, \qquad b_i = \lim_{n\to\infty} b_i^n.$$

Since $|a_i^n - a_i^{n+1}| < 2^{-(n+1)}$ and $|b_i^n - b_i^{n+1}| < 2^{-(n+1)}$, we have

$$|a_i^n - a_i| \leq 2^{-n}, \qquad |b_i^n - b_i| \leq 2^{-n}.$$

Let X be the multi-subset of $\mathcal{I}$ consisting of the intervals $I_i = [a_i, b_i)$ for all i (or for all $i \leq \max k_n$ if $\{k_n\}$ is bounded).

First, we claim that X is in $\overline{\mathcal{B}}$. Given $\epsilon > 0$, choose N large enough that $2^{-N+2} < \epsilon$. Then for $i > k_N$, the interval I_i first appears in X_{n_i} for some $n_i > N$. Looking at the matchings $C_N, \ldots, C_{n_i - 1}$, we get a composite matching C_{N,n_i} between X_N and X_{n_i}. Since each C_n satisfied the bound $2^{-(n+1)}$, the matching C_{N,n_i} must satisfy the bound

$$\sum_{n=N}^{n_i-1} 2^{-(n+1)} = 2^{-N} - 2^{-n_i}.$$

Since all intervals of length $> 2^{-N+1}$ in X_N appear as an I_j^N, we must have that the length of $I_i^{n_i}$ in X_{n_i} must be less than

$$2^{-N+1} + 2(2^{-N} - 2^{-n_i}) = 2^{-N+2} - 2^{-n_i+1}.$$

Since each endpoint in I_i differs from the endpoint of $I_i^{n_i}$ by at most 2^{-n_i}, the length of I_i can be at most

$$2^{-N+2} - 2^{-n_i+1} + 2 \cdot 2^{-n_i} = 2^{-N+2} < \epsilon.$$

Thus, the cardinality of the multi-subset of X of those intervals of length $> \epsilon$ is at most k_N.

Next we claim that $\langle X_n \rangle$ converges to X. We have a matching of X_n with X given by matching the intervals $I_1^n, \ldots, I_{k_n}^n$ in X_n with the corresponding intervals $I_1, \ldots, I_{k_n}$ in X. Our estimates above for $|a_i^n - a_i|$ and $|b_i^n - b_i|$ show that $d_\infty(I_i^n, I_i) \leq 2^{-n}$. By construction, each leftover interval in X_n has length $\leq 2^{-n+1}$ and the previous paragraph shows that each leftover interval in X has length $< 2^{-n+2}$. Thus, $d_{\mathcal{B}}(X_n, X) < 2^{-n+1}$.

Finally we note that if each X_n is in $\mathcal{B}_N$ for fixed N, then each $k_n \leq N$ and so X is in $\mathcal{B}_N$. □

4. Failure of robustness

Inevitably physical measurements will result in bad samples. As a consequence, we are interested in invariants which have limited sensitivity to a small proportion of arbitrarily bad samples. Many standard invariants not only have high sensitivity to a small proportion of bad samples, but in fact have high sensitivity to a small number of bad samples. We do not claim particular novelty for the general nature of the results of this section, as these issues have been folklore for some time. However, we do not know any place in the literature where precise statements are written down. We use the following terminology to describe the instability of these invariants.

Definition 4.1. A function f from the set of finite metric spaces to $\mathbb{R}$ is *fragile* if it not robust (in the sense of Definition 1.6) for any robustness coefficient $r > 0$.

In some cases, an even stronger kind of sensitivity holds.

Definition 4.2. A function f from the set of finite metric spaces to $\mathbb{R}$ is *extremely fragile* if there exists a constant k such that for every non-empty finite metric space X and constant N there exists a metric space X' and an isometry $X \to X'$ such that $|X'| \leq |X| + k$ and $|f(X') - f(X)| > N$.

Informally, extremely fragile in Definition 4.2 means that adding a small constant number of points to any metric space can arbitrarily change the value of the invariant. In particular, an extremely fragile function is fragile, but extremely fragile is much more unstable than just failing to be robust (note the quantifier on the space X). As we indicated in the introduction, Gromov-Hausdorff distance is fragile; here we show it is extremely fragile.

Proposition 4.3. *Let (Z, d_Z) be a non-empty finite metric space. The function $d_{GH}(Z, -)$ is extremely fragile.*

Proof. Given $N > 0$, consider the space X' which is defined as a set to be the disjoint union of X with a new point w, and made a metric space by setting

$$\begin{aligned} d(w, x) &= \alpha, & x &\in X, \\ d(x_1, x_2) &= d_X(x_1, x_2), & x_1, x_2 &\in X, \end{aligned}$$

where $\alpha > \operatorname{diam}(Z) + 2d_{GH}(Z,X) + 2N$. We claim

$$|d_{GH}(Z,X) - d_{GH}(Z,X')| > N.$$

Given any metric space (Y, d_Y) and isometries $f\colon X' \to Y$, $g\colon Z \to Y$, we need to show that $d_Y(f(X'), g(Z)) > N + d_{GH}(Z,X)$. We have two cases. First, if no point z of Z has $d_Y(g(z), f(w)) \leq N + d_{GH}(Z,X)$, then we have $d_Y(f(X'), g(Z)) > N + d_{GH}(Z,X)$. On the other hand, if some point z of Z has $d_Y(g(z), f(w)) < N + d_{GH}(Z,X)$, then every point z in Z satisfies $d_Y(g(z), f(w)) \leq N + \operatorname{diam}(Z) + d_{GH}(Z,X)$. Choosing some x in X, we see that for every z in Z, $d_Y(f(x), g(z)) \geq \alpha - (N + \operatorname{diam}(Z) + d_{GH}(Z,X))$. It follows that

$$d_Y(f(X'), g(Z)) \geq \alpha - (N + \operatorname{diam}(Z) + d_{GH}(Z,X)) > N + d_{GH}(Z,X). \qquad \square$$

The homology and persistent homology of a point cloud turns out to be a somewhat less sensitive invariant. Nonetheless, a similar kind of problem can occur. It is instructive to consider the case of H_0 or PH_0. By adding ℓ points far from the original metric space X, one can change either H_0 or PH_0 by rank ℓ. The further the distance of the points, the longer the additional bars in the barcode and we see for example that the distance $d_{\mathcal{B}}(B, -)$ in the bottleneck metric from any fixed barcode B is a extremely fragile function. (If we are truncating the barcodes, $d_{\mathcal{B}}$ is bounded by the length of the interval we are considering, so technically is robust, but not in a meaningful way.) We can also consider the rank of H_0 or of PH_0 in a range; here the distortion of the function depends on the number of points, but we see that the function is fragile.

For H_k and PH_k, $k \geq 0$, the same basic idea obtains: we add small spheres sufficiently far from the core of the points in order to adjust the required homology. We work this out explicitly for PH_1.

Definition 4.4. For each integer $k > 0$ and real $\ell > 0$, let the metric circle $S^1_{k,\ell}$ denote the metric space with k points $\{x_i\}$ such that

$$d(x_i, x_j) = \ell\left(\min(|i-j|, |k-i-j|)\right).$$

For $\epsilon < \ell$, the Vietoris-Rips complex associated to $S^1_{k,\ell}$ is just a collection of disconnected points. It is clear that as long as $k \geq 4$, when $\ell \leq \epsilon < 2\ell$, $|R_\epsilon(S^1_{k,\ell})|$ has the homotopy type of (and is in fact homeomorphic to) a circle. In fact, we can say something more precise:

Lemma 4.5. *For*

$$\ell \leq \epsilon < \left\lceil \frac{k}{3} \right\rceil \ell,$$

the rank of $H_1(R_\epsilon(S^1_{k,\ell}))$ *is at least* 1.

Proof. Consider the map f from $R_\epsilon(S^1_{k,\ell})$ to the unit disk D^2 in $\mathbb{R}^2$ that sends x_i to $(\cos(2\pi\frac{i}{n}), \sin(2\pi\frac{i}{n}))$ and is linear on each simplex. The condition $\epsilon < \lceil \frac{k}{3} \rceil \ell$ precisely ensures that whenever $\{x_{i_1}, \ldots, x_{i_n}\}$ forms a simplex σ in the Vietoris-Rips complex, the image vertices $f(x_{i_1}), \ldots, f(x_{i_n})$ lie on an arc of angle $< \frac{2}{3}\pi$ on the unit circle, and so $f(\sigma)$ in particular lies in an open half plane through the origin. It follows that the origin $(0,0)$ is not in the image of any simplex, and f defines a map from $R_\epsilon(S^1_{k,\ell})$ to the punctured disk $D^2 - \{(0.0)\}$. Since $\ell \leq \epsilon$, we have the 1-cycle

$$[x_1, x_2] + \cdots + [x_{k-1}, x_k] + [x_k, x_1]$$

of $R_\epsilon(S^1_{k,\ell})$ which maps to a 1-cycle in $D^2 - \{(0,0)\}$ representing the generator of $H_1(D^2 - \{0,0\})$. □

The length ℓ and number $k \geq 4$ is arbitrary, so again, we conclude that functions like $d_{\mathcal{B}}(B, \mathrm{PH}_1(-))$ are extremely fragile. Results for higher dimensions (using similar standard equidistributed models of n-spheres) are completely analogous.

Proposition 4.6. *Let B be a barcode. The functions $d_{\mathcal{B}}(B, \mathrm{PH}_k(-))$ from finite metric spaces to $\mathbb{R}$ are extremely fragile.*

In terms of rank, the lemma shows that we can increase the rank of first persistent homology group of a metric space X on an interval $[a,b]$ by m simply by adding extra points. One can also typically reduce persistent homology intervals by adding points "in the center" of the representing cycle. It is somewhat more complicated to precisely analyze the situation, so we give a representative example: Suppose the cycle is represented by a collection of points $\{x_i\}$ such that the maximum distance $d(x_i, x_j) \leq \delta$. Then adding a point which is a distance δ from each of the other points reduces the lifetime of that cycle to δ. In any case, the results of the lemma are sufficient to prove the following proposition.

Proposition 4.7. *The function that takes a finite metric space to the rank of* PH_k *on a fixed interval* $[a,b]$ *is fragile.*

These computations suggest a problem with the stability of the usual invariants of computational topology. A small number of bad samples can lead to arbitrary changes in these invariants.

5. The main definition and theorem

Fix a good functorial assignment of a simplicial complex to a finite metric space and a scale parameter ϵ. Recall that we write PH_k of a finite metric space to denote the persistent homology of the associated direct system of complexes. Motivated by the concerns of the preceding section, we define Φ^n_k as the distribution of barcodes induced by samples of size n. The basic idea motivating Φ^n_k is that in order to obtain robust invariants, given a sample budget of N samples from (X, ∂_X, μ_X), instead of computing a single estimator from the N samples it is preferable to look at *the distribution* of estimators produced by blocks of samples of size $n \ll N$. Note that this is closely related to the idea behind bootstrap resampling. It is also a more sophisticated version of computing a trimmed mean (i.e., a mean in which extremal samples are thrown out) — rather than removing extremal samples, we simply subsample at a rate such that the likelihood of seeing a bad sample is low. Ideally, this approach retains the information contained in those samples while also estimating the "true" value.

Definition 5.1. For a metric measure space (X, ∂_X, μ_X) and fixed $n, k \in \mathbb{N}$, define the kth n-sample persistent homology as

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

the probability distribution on $\overline{\mathcal{B}}$ induced by pushforward along PH_k from the product measure μ_X^n on X^n.

This definition makes sense because PH_k is a continuous function and the measures on the domain and codomain are both Borel. Indeed, the stability theorem of

Chazal, et. al. [16, 3.1] (Theorem 2.9 above) and the fact that the Gromov Hausdorff metric is less than or equal to the product metric in X^n implies that PH_k is Lipschitz with Lipschitz constant at most 1.

In order to apply Φ_k^n, we need to know two things. First, that for fixed n and k the approximation to Φ_k^n computed by choosing samples from the empirical measure on a large sample space S drawn from (X, ∂_X, μ_X) converges in probability to the actual value (as $|S|$ goes to infinity). Second, that for fixed n and k the approximation to Φ_k^n obtained by computing the empirical measure from ℓ blocks of n samples converges in probability to the actual value (as ℓ goes to infinity). The latter follows from the weak law of large numbers for the empirical process. The goal of this section is to prove the following theorem, which establishes the former asymptotic consistency. For this (and in the remainder of the section), we assume that we are computing PH using the Vietoris-Rips complex.

Theorem 5.2. *Let (X, ∂_X, μ_X) and (Y, ∂_Y, μ_Y) be compact metric measure spaces. Then we have the following inequality:*

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(Y, \partial_Y, \mu_Y)) \leq n\, d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)).$$

Proof. Assume that $d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) < \epsilon$. Then we know that there exist embeddings $\iota_1 \colon X \to Z$ and $\iota_2 \colon Y \to Z$ into a metric space Z and a coupling $\hat{\mu}$ between $(\iota_1)_*\mu_X$ and $(\iota_2)_*\mu_Y$ such that the probability mass of the set of pairs (z, z') under $\hat{\mu}$ such that $\partial_Z(z, z') \geq \epsilon$ is less than ϵ.

We can regard the restriction of $\hat{\mu}^{\otimes n}$ to the full measure subspace $(X \times Y)^n$ of $(Z \times Z)^n$ as a probability measure on $X^n \times Y^n$. This then induces a coupling between $(\mathrm{PH}_k)_*(\mu_X^{\otimes n})$ and $(\mathrm{PH}_k)_*(\mu_Y^{\otimes n})$ on $\overline{\mathcal{B}}$, which we now study. Consider n samples $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ from $Z \times Z$ drawn according to the product distribution $\hat{\mu}^{\otimes n}$. Now consider the probability that

$$\alpha = \sup_{1 \leq i,j \leq n} |\partial_X(x_i, x_j) - \partial_Y(y_i, y_j)| \geq 2\epsilon.$$

The triangle inequality implies that

$$|\partial_X(x_i, x_j) - \partial_Y(y_i, y_j)| = |\partial_Z(x_i, x_j) - \partial_Z(y_i, y_j)| \leq \partial_Z(x_i, y_i) + \partial_Z(x_j, y_j).$$

Therefore, the union bound implies that the probability that $\alpha \geq 2\epsilon$ is bounded by

$$\Pr(\exists i \mid \partial_Z(x_i, y_i) \geq \epsilon) \leq 1 - (1 - \epsilon)^n < n\epsilon$$

Next, define a relation R that matches x_i and y_i. By definition, the distortion of this relation is $\operatorname{dis} R = \alpha$, and so

$$d_{GH}(\{x_i\}, \{y_i\}) \leq \frac{1}{2}\alpha.$$

By the stability theorem of Chazal, et. al. [16, 3.1] (Theorem 2.9 above), this implies that the probability that

$$d_{\mathcal{B}}(\mathrm{PH}_k(\{x_i\}), \mathrm{PH}_k(\{y_i\})) \geq \epsilon$$

is bounded by $n\epsilon$. This further implies that the probability that

$$d_{\mathcal{B}}(\mathrm{PH}_k(\{x_i\}), \mathrm{PH}_k(\{y_i\})) \geq n\epsilon$$

is also bounded by $n\epsilon$. Therefore, we can conclude that

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(Y, \partial_Y, \mu_Y)) \leq n\epsilon. \qquad \square$$

We note the dependence on n in the statement of the bound in Theorem 5.2. As n increases, the quantity Φ_k^n becomes a finer approximation of the persistent homology of the support of X. Specifically, more points per sample means that Φ_k^n is increasingly sensitive to small features of X. In this light, it is not surprising that the bound becomes weaker for larger n.

Next we discuss the tightness of the bound in Theorem 5.2. Clearly, this bound is vacuous when $d_{GPr}((X,\partial_X,\mu_X),(Y,\partial_Y,\mu_Y)) > \frac{1}{n}$ since the Prohorov metric is bounded by 1, but we show that it becomes tight as $d_{GPr}((X,\partial_X,\mu_X),(Y,\partial_Y,\mu_Y))$ approaches zero. Reviewing the argument, starting from the hypothesis that $d_{GPr}((X,\partial_X,\mu_X),(Y,\partial_Y,\mu_Y)) = \epsilon$, we used the union bound to obtain a bound of $n\epsilon$. The exact bound in question is $1-(1-\epsilon)^n$. The leading term in the expansion of this quantity is $n\epsilon$, and so as $\epsilon \to 0$ the bound in the theorem becomes increasingly tight. When ϵ is close to $\frac{1}{n}$, using more terms in the expansion yields better bounds (for example, when $\epsilon = \frac{1}{n}$, $1-(1-\epsilon)^n \leq .75$ and tends to $1-\frac{1}{e} \approx .632$ for large n).

The exact bound $1-(1-\epsilon)^n$ yields a tight estimate on $d_{GPr}(X^n,Y^n)$ (using the sup product metric), as we can see by the following example. Consider the case of two finite metric spaces $X = X_1 \cup X_2$ and $Y = Y_1 \cup Y_2$, where $|Y_1| = |X_1|$ and $|Y_2| = |X_2|$. Define d_X via $d_X(x_1,x_1') = \alpha$ for $x_1, x_1' \in X_1$, $d_X(x_2,x_2') = \beta$ for $x_2, x_2' \in X_2$, and $d_X(x_1,x_2) = \gamma$ for $x_1 \in X_1$ and $x_2 \in X_2$. Here γ should be substantially larger than α and β. We define d_Y analogously, using the same α and β but with γ' distinct from γ (and without loss of generality assume that $\gamma' > \gamma$). Consider the metric space Z formed from the disjoint union of X_1, X_2, and Y_2, and with the metric induced from d_X and d_Y except that $d_Z(x_2,y_2) = \gamma' - \gamma$. There are evident isometries $i\colon X \to Z$ and $j\colon Y \to Z$; it is easy to see that $d_{Pr}(i_*\mu_X, j_*\mu_Y) = \epsilon$ for $\epsilon = \frac{|X_2|}{|X_1|+|X_2|}$ and moreover that this pair of embeddings minimizes the Prohorov distance, so $d_{GPr}(X,Y) = \epsilon$. The induced embeddings $i^n\colon X^n \to Z^n$, $j^n\colon Y^n \to Z^n$ satisfy

$$d_{Pr}(i^n_*\mu_X^{\otimes n}, j^n_*\mu_Y^{\otimes n}) = 1-(1-\epsilon)^n$$

and a straight-forward combinatorics argument shows that this embedding also minimizes the Prohorov distance, so $d_{GPr}(X^n,Y^n) = 1-(1-\epsilon)^n$. (We thank Olena Blumberg for help with this example.)

By varying the parameters in the previous example, it is now clear that the bound on Φ_0^n is tight and we can achieve the upper bound with a variety of barcode lengths. Tightness for Φ_k^n for $k > 0$ is harder to analyze. Theorem 2.9 is expected to be tight for all k but nothing has yet appeared in the literature for $k > 0$. If the bound in Theorem 2.9 is tight, it is reasonable to expect the bound in Theorem 5.2 also to be tight; however, we do not know a rigorous argument.

Remark 5.3. For a related discussion involving finite distance matrices, see [21, §6, §7]. There the constant N (size of the matrix) is analogous to the parameter n above and enters into their formulas through the distance d_M, which depends on N.

We regard the bound as most useful for fixed n. Then a basic consequence of Theorem 5.2 is that consideration of large finite samples will suffice for computing Φ_k^n. For a finite metric space (X,∂_X), let μ and μ' denote two measures on X.

Then we have the following inequality [21, 5.4]

$$d_{GPr}((X,\partial_X,\mu),(X,\partial_X,\mu')) \leq 1 - \sum_{x\in X} \min(\mu(x),\mu'(x)), \tag{5.4}$$

which follows by choosing a coupling that has measure at least the minimum of the two measures on each point.

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

Proof. This result is a consequence of the fact that $\{S_i\}$ converges in probability to (X,∂_X,μ_X) in the Gromov-Prohorov metric (which can be checked directly using equation (5.4), as in [21, §5], or can be deduced from the analogous convergence result for the Gromov-Wasserstein distance [43, 3.5.(iii)] and the comparison between the Gromov-Prohorov distance and the Gromov-Wasserstein distance [30, 10.5]). □

Another consequence of Theorem 5.2 is that Φ_k^n provide robust descriptors for metric measure spaces (X,∂_X,μ_X). Specifically, observe that if we have finite metric spaces $(X,\partial_X) \subset (X',\partial_{X'})$ and a probability measure $\mu_{X'}$ on X' that restricts to μ_X on X (i.e., for $A \subset X$, $\mu_X(A) = \mu_{X'}(A)/\mu_{X'}(X)$), then equation (5.4) implies that

$$d_{Pr}(i_*\mu_X,\mu_{X'}) \leq 1 - \mu_{X'}(X).$$

Thus, when $X' \setminus X$ has probability $< \epsilon$,

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

In particular, when X and X' are finite metric spaces with the uniform measure, we get

$$d_{Pr}(\Phi_k^n(X,\partial_X,\mu_X),\Phi_k^n(X',\partial_X^{'},\mu_{X'})) \leq n(1-|X|/|X'|).$$

As an immediate consequence we obtain the following result.

Theorem 5.6. *For fixed n,k, Φ_k^n is uniformly robust with robustness coefficient r and estimate bound $nr/(1+r)$ for any r.*

Remark 5.7. It would be useful to prove analogues of the main theorem for other methods of assigning complexes; e.g., the witness complex (see Remark 2.4 and Section 8). We expect that the recent stability results of [15] will be useful in this connection.

6. Hypothesis testing, confidence intervals, and numerical invariants

In this section, we describe various ways to use Φ_k^n to perform statistical inference about the homological invariants of a point cloud. The basic goal is to provide quantitative ways of saying what observed barcodes or empirical barcode distributions "mean". We are predominantly interested in addressing two kinds of questions:

(1) Are two given empirical barcode distributions coming from the same underlying distribution?
(2) Is a particular empirical barcode distribution consistent with the hypothesis that the underlying distribution has k "long bars"?

We approach both of these questions from the perspective of classical hypothesis testing, likelihood scores, and confidence intervals; for example, see [23, §2] for a review. We discuss a variety of test statistics derived from Φ_k^n; thus, the use of these procedures are justified in practice by Theorem 5.2 (and specifically Corollary 5.5). Moreover, we are able to use Theorem 5.2 to show that many of the test statistics we describe are robust.

We begin by explaining the basic procedure for computing approximations to Φ_k^n. Corollary 5.5 justifies the treatment of Φ_k^n of the empirical measure on a sufficiently large sample $S \subset X$ of size N as a good approximation for $\Phi_k^n(X, \partial_X, \mu_X)$. (Note that the dependence on n in the bound in Theorem 5.2 implies that we will have to choose $n \ll N$ in order to expect reasonable results; see the discussion in the next section for some examples of how to choose n.) Next, we can estimate Φ_k^n on S empirically via Monte Carlo simulation, i.e., simply sampling blocks of n samples from S over and over again. The weak law of large numbers for empirical distributions guarantees that this estimate converges in probability as the number of such samples increases. Therefore, we have an asymptotically convergent scheme for numerically approximating Φ_k^n (and hence quantities derived from it). We now turn to questions of statistical inference.

6.1. **Hypothesis testing using Φ_k^n.** The most basic question we can pose is whether a given observed barcode B is more consistent with an underlying metric measure space (X, ∂_X, μ_X) or an alternate metric measure space $(X', \partial'_X, \mu'_X)$. A likelihood ratio provides a good test statistic to determine an answer to this question. Specifically, we can evaluate the likelihood of the hypotheses $\mathrm{Hyp}_k^n(X; B, \epsilon)$ and $\mathrm{Hyp}_k^n(X'; B, \epsilon)$ that B is within ϵ of a barcode drawn from (X, ∂_X, μ_X) and $(X', \partial_{X'}, \mu_{X'})$ respectively.

Given an observed barcode B (e.g., obtained by sampling n points from an unknown metric measure space (Z, ∂_Z, μ_Z)), we can compute the likelihood score

$$L_Y = L(Y, \partial_Y, \mu_Y) = \Pr(d_{\mathcal{B}}(B, \tilde{B}) < \epsilon \mid \tilde{B} \text{ drawn from } \Phi_k^n(Y, \partial_Y, \mu_Y))$$

for each hypothesis space (X, ∂_X, μ_X) and $(X', \partial_{X'}, \mu_{X'})$. The ratio $L_X/L_{X'}$ then provides a test statistic for comparing the two hypotheses. To determine how to interpret the test statistics (e.g., to compute p-values), we require knowledge of the distribution of the test statistic induced by assuming that B was drawn from $\Phi_k^n(X, \partial_X, \mu_X)$ and $\Phi_k^n(X', \partial_{X'}, \mu_{X'})$ respectively. These distributions can be approximated by Monte Carlo simulation, i.e., repeated sampling from the two distributions and computation of histograms.

More generally, for a metric measure space (X, ∂_X, μ_X) and a particular subset S of $\overline{\mathcal{B}}$, we can test the hypothesis that the distribution $\Phi_k^n(X)$ has mass $\geq \epsilon$ on S as follows. For any hypothetical distribution on $\overline{\mathcal{B}}$ with mass $\geq \epsilon$ on S, the probability of an empirical sample of size N having q or fewer elements in S is bounded above by the binomial cumulative distribution function

$$\mathrm{BD}(N, q, \epsilon) = \sum_{i=0}^{q} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}.$$

Then given an empirical approximation $\mathcal{E}$ to Φ_k^n obtained from N samples, we can test the hypothesis that Φ_k^n has mass $\geq \epsilon$ in S, by taking q to be the number of such elements in $\mathcal{E}$. When $\mathrm{BD}(N, q, \epsilon) < \alpha$, we can reject this hypothesis at the $1 - \alpha$ level.

6.2. **Distribution comparison test statistics.** Another kind of question we might ask is to determine whether to reject the hypothesis that two empirical distributions on barcode space (i.e., Φ_k^n computed based on two different large samples S and S') came from the same underlying distribution. In our setting we cannot assume very much about the class of possible distributions and so we are forced to rely on non-parametric methods. This imposes significant constraints — most asymptotic results on non-parametric tests for distribution comparison work only for distributions on $\mathbb{R}$. Thus, the first step is to project the data from barcode space into $\mathbb{R}$. The following definition is the first of several kinds of projections we discuss.

Definition 6.1. Let (X, ∂_X, μ_X) be a compact metric measure space. Fix $k, n \in \mathbb{N}$.

(1) Define the distance distribution $\mathcal{D}^2$ on $\mathbb{R}$ to be the distribution on $\mathbb{R}$ induced by applying $d_{\mathcal{B}}(-,-)$ to pairs (b_1, b_2) drawn from $\Phi_k^n(X, \partial_X, \mu_X)^{\otimes 2}$.
(2) Let B be a fixed barcode in $\overline{\mathcal{B}}$, and define $\mathcal{D}_B$ to be the distribution induced by applying $d_{\mathcal{B}}(B, -)$.

Since both $\mathcal{D}^2$ and $\mathcal{D}_B$ are continuous with respect to the Gromov-Prohorov metric [30, 6.6], Corollary 5.5 justifies working with empirical approximations to $\mathcal{D}^2$ and $\mathcal{D}_B$.

One application of these projections is simply a direct use of the two-sample Kolmogorov-Smirnov statistic [23, §6]. This test statistic gives a way to determine whether two observed empirical distributions were obtained from the same underlying distribution; the salient feature about this statistic is that for distributions on $\mathbb{R}$ the p-values of the test statistic are asymptotically independent of the underlying distribution as long as the samples are identically independently drawn.

To compute the Kolmogorov-Smirnov test statistic for two sets of samples $\mathcal{S}_1$ and $\mathcal{S}_2$, we first compute the empirical approximations $\mathcal{E}_1$ and $\mathcal{E}_2$ to the cumulative density functions,

$$E_i(t) = |\{x \in S_i \mid x \leq t\}|/|S_i|,$$

and use the test statistic $\sup_t |\mathcal{E}_1(t) - \mathcal{E}_2(t)|$. In practice, since $|S_i|$ is large we approximate $\mathcal{E}_i$ using Monte Carlo simulation. The distribution-independence of the statistic now implies that standard tables (e.g., in the appendix to [23]) or the built-in Matlab functions can then be used to compute p-values for deciding if the statistic allows us to reject the hypothesis that the distributions are the same.

One might similary consider the Mann-Whitney test or various other nonparametric techniques for testing the same hypotheses [23, §5]. For example, another way to handle this problem is to use a χ^2 test for discrete distributions. There are many ways to construct suitable distributions for this test; we present two natural choices here.

(1) Take histograms from $\mathcal{D}^2_{S_1}$ and $\mathcal{D}^2_{S_2}$ with identical fixed numbers of bins and bin widths.
(2) Fix a finite set $\{B_j\} \subset \overline{\mathcal{B}}$ of reference barcodes, where $1 \leq j \leq m$. These reference barcodes should be chosen without reference to the observed data. Next, for each barcode with nonzero probability measure in (the given empirical approximation to) Φ_k^n, assign the count to the nearest reference barcode.

The second method makes sense if we have a priori information about the expected shape of the barcode distributions.

Let $\mathcal{A}_i(j)$ denote the count either for bin j or for reference barcode B_j in sample i (for $i = 1, 2$). The test statistic in the χ^2 test for comparing S_1 and S_2 is then defined to be

$$\chi^2 = \sum_{j=1}^{m} \frac{(\mathcal{A}_1(j) - \mathcal{A}_2(j))^2}{\mathcal{A}_1(j) + \mathcal{A}_2(j)}.$$

As the notation suggest, asymptotically this has a χ^2 distribution with $m' - 1$ degrees of freedom (where m' is the number of reference barcodes with nonzero counts) [46, §17]. As such, we can again look up the p-values for this distribution in standard tables when performing hypothesis testing.

6.3. **Numerical summaries as test statistics.** Natural test statistics for studying hypotheses about empirical barcode distributions come from numerical summaries associated to Φ_k^n. For instance, a natural test statistic measures the distance to a fixed hypothesis distribution.

Definition 6.2. Let (X, ∂_X, μ_X) be a compact metric measure space and let $\mathcal{P}$ be a fixed reference distribution on $\overline{\mathcal{B}}$. Fix $k, n \in \mathbb{N}$. Define the homological distance on X relative to $\mathcal{P}$ to be

$$\mathrm{HD}_k^n((X, \partial_X, \mu_X), \mathcal{P}) = d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \mathcal{P}).$$

Corollary 5.5 again applies to show that large finite samples $S \subset X$ suffice to approximate HD_k^n. In fact, the convergence is better since we are working over $\mathbb{R}$ and the Glivenko-Cantelli theorem applies.

Lemma 6.3. *Let $S_1 \subset S_2 \subset \cdots \subset S_i \subset \cdots$ be a sequence of randomly drawn samples from (X, ∂_X, μ_X). We regard S_i as a metric measure space using the subspace metric and the empirical measure. Then for $\mathcal{P}$ a fixed reference distribution on $\overline{\mathcal{B}}$, $\mathrm{HD}_k^n(S_i, \mathcal{P})$ converges almost surely to $\mathrm{HD}_k^n((X, \partial_X, \mu_X), \mathcal{P})$.*

An immediate consequence of Theorem 5.2 is the following robustness result (paralleling Theorem 5.6).

Theorem 6.4. *For fixed $n, k, \mathcal{P}$, $\mathrm{HD}_k^n(-, \mathcal{P})$ is uniformly robust with robustness coefficient r and estimate bound $nr/(1 + r)$ for any r.*

Another source of tractable test statistics is the moments of the distributions introduced in Definition 6.1. A virtue of distributions on $\mathbb{R}$ is that they can be naturally summarized by moments; in contrast, moments for distributions on barcode space are hard to compute (for instance, see [45]). Even simply constructing meaningful centroids for a set of points in barcode space is challenging; for example, geodesics between close points are not unique, although the barcode metric space is a length space (it is straightforward to construct midpoints between any pair of barcodes). Because we have emphasized robust statistics, we work with the median or a trimmed mean and introduce the following test statistics:

Definition 6.5. Let (X, ∂_X, μ_X) be a compact metric measure space and fix a reference barcode $B \in \mathcal{B}$. Fix $k, n \in \mathbb{N}$. Define the median homological distance relative to B to be

$$\mathrm{MHD}_k^n((X, \partial_X, \mu_X), B) = \mathrm{median}(\mathcal{D}^2).$$

For $0 < \alpha < \frac{1}{2}$, define the α-trimmed mean homological distance to be

$$\widetilde{\mathrm{MHD}}_k^n((X, \partial_X, \mu_X), B) = \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} q(\mathcal{D}^2),$$

where q denotes the quantile function. (Roughly speaking, we discard the fraction α of the highest and lowest values and take the mean of the remainder.)

Again, Corollary 5.5 implies that consideration of large finite samples $S \subset X$ suffices to approximate these test statistics.

Lemma 6.6. *Let $S_1 \subset S_2 \subset \cdots \subset S_i \subset \cdots$ be a sequence of randomly drawn samples from (X, ∂_X, μ_X). We regard S_i as a metric measure space using the subspace metric and the empirical measure. Let $B \in \mathcal{B}$ be a fixed reference barcode.*

(1) *Assume that $\mathcal{D}^2((X, \partial_X, \mu_X), B)$ has a distribution function with a positive derivative at the median. Then $\mathrm{MHD}_k^n(S_i, B)$ almost surely converges to $\mathrm{MHD}_k^n((X, \partial_X, \mu_X), B)$.*
(2) *$\widetilde{\mathrm{MHD}}_k^n(S_i, B)$ almost surely converges to $\widetilde{\mathrm{MHD}}_k^n((X, \partial_X, \mu_X), B)$.*

Proof. As in the proof of Corollary 5.5, the fact that $\{S_i\}$ converges to (X, ∂_X, μ_X) in the Gromov-Prohorov metric implies that $\mathcal{D}(S_i)$ weakly converges to $\mathcal{D}$. Now the central limit theorem for the sample median (see for instance [40, III.4.24]) and the hypothesis about the derivative at the median implies the convergence of medians. Analogously, the central limit theorem for the trimmed mean [17, §4] (which holds without further assumption provided that $\alpha < \frac{1}{2}$) gives the second part of the result. □

The hypothesis on the median is the standard hypothesis for consistency of the central limit theorem (and the bootstrap estimator for) the sample median; it is known that this hypothesis is required [46, 5.11]. Although it is our experience that this hypothesis holds in practice, it can be difficult to rigorously verify for an unknown underlying distribution. For this reason, the use of the trimmed mean may be preferable in cases where constraint on the possible hypotheses is unavailable. As α approaches $\frac{1}{2}$, the trimmed mean converges to the median, and so choosing $\alpha = \frac{1}{2} - \epsilon$ for small ϵ yields a reasonable alternative to the median.

As discussed in the introduction, a counting argument yields the following robustness result.

Theorem 6.7. *For any n, k, B, the function $\mathrm{MHD}_k^n(-, B)$ from finite metric spaces (with the uniform probability measure) to $\mathbb{R}$ is robust with robustness coefficient $> (\ln 2)/n$.*

Proof. For a finite metric space X, expanding X to X', the proportion of n-element samples of X' which are samples of X is $(|X|/|X'|)^n$; when this number is more than $1/2$, the median value of any function f on the set of n-element samples of X' is then bounded by the values of f on n-element samples of X. Since $(N/(N+rN))^n > 1/2$ for $r < 2^{1/n} - 1$, any such function f will be robust with robustness coefficient r satisfying this bound, and in particular for $r = (\ln 2)/n$. □

In order to obtain the p-value cutoffs for performing hypothesis testing, we can again use Monte Carlo simulation to estimate the distribution of these estimators under different hypotheses. Another possibility is to use asymptotic estimates, which we discuss below in the context of confidence intervals.

We close the section by remarking that there are many other possible numerical invariants one might associate to Φ_k^n (and apply as test statistics). For instance, we define for a barcode B the quantity

$$g_m(B) = |B(m)| - |B(m+1)|,$$

where $B(i)$ denotes the ith largest interval in B. Then the quantity

$$g_m = \text{median}(g_m(\Phi_k^n(-)))$$

and the related quantity

$$g = \max(g_m)$$

are useful test statistics for determining if there is a group of "long bars" in the underlying distribution by checking for which (if any m) has a large value of g_m. For instance, when g is small this suggests that the underlying metric measure space is generated by a topological space with no homology in dimension k. And large values of g_m suggest that the underlying space has rank m homology in dimension k. Of course, in order to make precise statistical statements to replace "suggests", we have to use Monte Carlo simulation in order to compute p-values and confidence intervals.

6.4. **Confidence intervals.** For HD_k^n, the only way to produce p-values and confidence intervals is to use Monte Carlo simulation (to estimate the distribution of HD_k^n on finite approximations to Φ_k^n). A particular advantage of MHD_k^n is that we can define confidence intervals using the standard non-parametric techniques for determining confidence intervals for the median and trimmed mean [24, §7.1]. For the median, we use appropriate sample quantiles (order statistics) to determine the bounds for an interval which contains the actual median with confidence $1-\alpha$. These confidence intervals then immediately yield cutoffs for p-values for hypothesis testing. For example, a simple approximation can be obtained from the fact that order statistics asymptotically obey binomial distributions, which lead to the following definition using the normal approximation to the binomial distribution.

Definition 6.8. Let (X, ∂_X, μ_X) be a metric measure space and B a fixed barcode. Fix $0 \leq \alpha \leq 1$ and n, k. Given m samples from $\mathcal{D}_B$, let $\{s_m\}$ denote the samples sorted from smallest to largest. Let u_α denote the $\frac{\alpha}{2}$ significance threshold for a standard normal distribution. The $1-\alpha$ confidence interval for the sample median (i.e., MHD_k^n) is given by the interval

$$\left[s_{\lfloor \frac{m+1}{2} - \frac{1}{2}\sqrt{m}u_\alpha \rfloor}, s_{\lceil \frac{m+1}{2} + \frac{1}{2}\sqrt{m}u_\alpha \rceil}\right].$$

For the trimmed mean, the situation is similar: asymptotic confidence intervals can be obtained from the sample standard deviation and an explicit formula [44]. Since we find that the median converges in practice, we do not write out the formula here (as it involves a number of complicated auxiliary quantities) and refer the interested reader to the cited reference.

6.5. **The validity of asymptotic p-values.** In the preceding discussion, the p-values and confidence intervals for our tests are always computed either via Monte Carlo simulation (i.e., sampling to estimate the distribution of the test statistic) or using formulas derived from asymptotic results. The latter are substantially easier and less computationally intensive to apply. However, we may be concerned about whether sample sizes are large enough for the asymptotic p-value to be good

approximations of the exact p-value — this issue is a pervasive problem when applying such non-parametric tests based on asymptotic results (e.g., see [46, §1.3]).

A standard approach to mitigating such concerns is to perform Monte Carlo simulation of the distribution of these test statistics computed from representative models for Φ_k^n (e.g., synthetic distributions generated by various standard manifolds); this provides heuristic guidance about suitable sample sizes. We provide some example calculations of this form in Section 7 below. However, the careful analyst with access to adequate samples and computer resources may simply choose to rely on Monte Carlo simulation methods. When adequate samples are lacking, resampling methods also often provide a more reliable means to compute cutoffs than asymptotic results. For instance, standard results about the consistency of the bootstrap for the sample median and sample trimmed mean [17] allow us to compute p-value thresholds and cofidence intervals for MHD_k^n via bootstrap resampling.

7. Demonstration of hypothesis testing on synthetic examples

In this section, we provide numerical experiments on synthetic data sets to demonstrate the statistical inference procedures and robustness results described in the previous section. We study a pair of examples embedded in $\mathbb{R}^2$ (an annulus and a pair of nested circles) and three families of examples in $\mathbb{R}^3$ (spheres, tori, and uniform noise in a box). Although the examples embedded in $\mathbb{R}^2$ are essentially trivial, the simplicity of the expected results allows us to focus on the methodology. The examples in $\mathbb{R}^3$ are more realistic but correspondingly are more complicated to interpret.

All of our experiments rely on the following procedures for producing empirical approximations to Φ_k^n. We fix a Monte Carlo parameter K which is large (we discuss estimates of how large K needs to be below). We then have the following basic algorithm:

Algorithm 7.1. For a fixed metric measure space (X, ∂_X, μ_X).

(1) Uniformly select K subsamples of size n from μ_X.
(2) Compute the empirical approximation to Φ_k^n from the K subsamples.

To better represent the use of these procedures in practice, we have the following variant algorithm. Fix a subsample size N.

Algorithm 7.2. For a fixed metric measure space (X, ∂_X, μ_X).

(1) Uniformly sample N points from μ_X.
(2) Uniformly select K subsamples of size n from the empirical measure on the N samples.
(3) Compute the empirical approximation to Φ_k^n from the K subsamples.

To actually carry out these algorithms, we used the Perseus codebase [38] to compute persistent homology from a finite metric space, executed from within a series of Python and Cython scripts that ran our various experimental setups. In order to avoid combinatorial explosion in the number of simplices when the scale parameter results in complete graphs, we typically capped the maximum scale parameter (i.e., truncated each of the bars in the barcodes). The experiments were run on various stock Linux machines; no individual experiment took more than a few minutes to complete. Our random number generation was done using the GSL

library [29] to produce uniform and Gaussian samples, and rejection sampling to simulate all other distributions (as described below).

Synthetic Example 1: The annulus and the annulus plus diameter linkage. We first consider a simple example which illustrates the robustness of the distributional invariants. The underlying metric measure space A is an annulus of inner radius 0.8 and outer radius of 1.2 in $\mathbb{R}^2$ (see Figure 1), equipped with the subspace metric and the area measure. The underlying manifold of A is clearly homotopy equivalent to a circle.

We sample from the annulus via rejection sampling; we sample uniformly from the bounding box $[-1.2, 1.2] \times [-1.2, 1.2]$ and only keep points (x, y) such that $0.8 \leq \sqrt{x^2 + y^2} \leq 1.2$.

We began by examining the rate of convergence in Corollary 5.5. Specifically, for $k = 1$ and various n, we consider subsamples S_i in the annulus of increasing cardinality and study the convergence of various distributions derived from $\Phi_k^n(S_i)$, using Algorithms 7.1 and 7.2 as a base. We compute the distance distribution $\mathcal{D}^2$ from $\Phi_k^n(S_i)$ and $\Phi_k^n(A)$ as the cardinality N_i of S_i increases and n varies, using a barcode cutoff of 0.375. We then used both the Kolmogorov-Smirnov test and the χ^2 test on histograms to repeatedly compare the estimates computed from samples of cardinality N_i to each other and to the reference distribution from A. Fixing $K = 1000$, our results indicated that $|S_i| = 1000$ sufficed to approximate the distribution for $n \leq 100$; with these parameters, we were essentially never able to reject the null hypothesis that the two distributions were drawn from the same underlying distribution.

Next, we turn to an illustrative example of the behavior of Φ_k^n in the face of maliciously chosen noise points. We generated sets $\mathcal{S}_1$ of 1000 points by sampling uniformly (via rejection sampling) from the annulus. Using the Vietoris-Rips complex, computing the barcode for the first homology group (with cutoff of 0.375) yields a single long interval, displayed in Figure 2. (We repeated this procedure many times with different subsamples of size 1000; the picture displayed is wholly representative of the results, which varied only very slightly across the samples.)

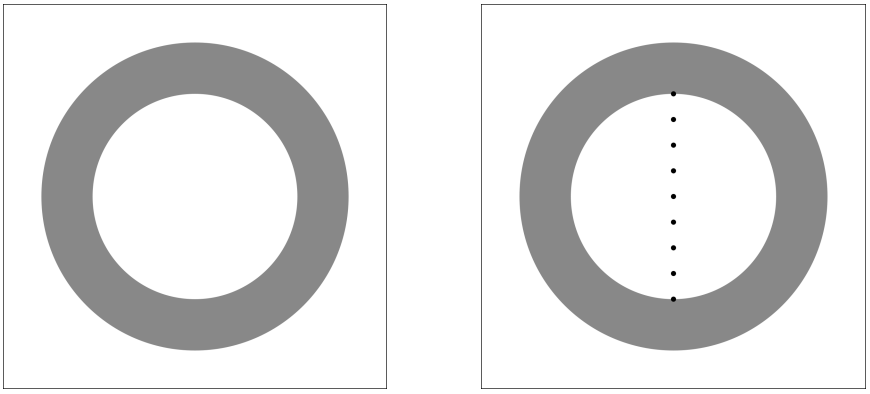

FIGURE 1. Annulus with inner radius 0.8 and outer radius 1.2 (left), and same annulus together with diameter linkage (right). (In experiments, points on diameter are chosen randomly.)

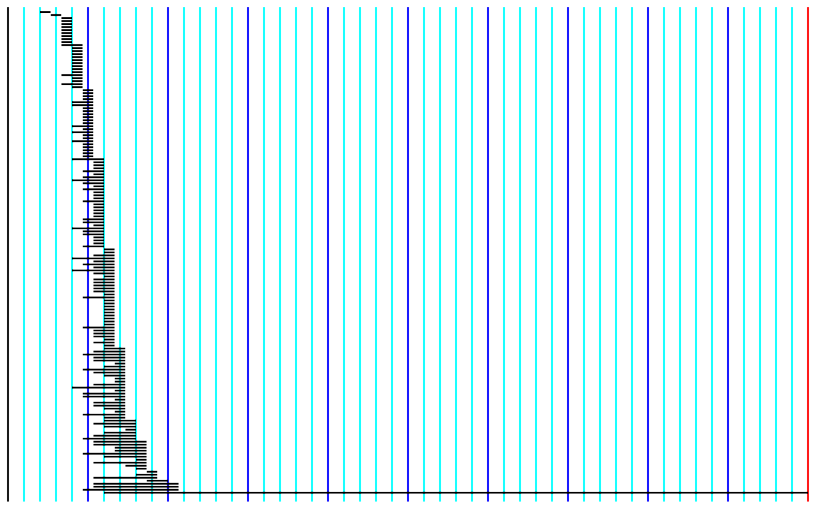

FIGURE 2. Barcode for annulus via the Vietoris-Rips complex with 1000 points, showing 1 long bar. Horizontal scale goes from 0 to 0.375. (Vertical scale is not meaningful.)

We then generated sets by drawing $\mathcal{S}_1$ as above and unioning with sets X drawn uniformly from the region $\{0\}\times[-0.8, 0.8] \subset \mathbb{R}^2$ to form sets $\mathcal{S}_2 = \mathcal{S}_1 \cup X$. When the added points are sufficiently numerous and well-distributed, the point cloud now appears to have been sampled from an underlying manifold homotopy equivalent to a figure 8 when the scale parameter is sufficiently large. (See Figure 1.) Computing the barcode for the first homology group now yields two long intervals, displayed in Figure 3. (Again, this barcode was stable over many repetitions.)

To test our methodology, we considered varying sizes for X (as a proportion of $|\mathcal{S}_1| = 1000$), and using Algorithm 7.2 we computed 1000 empirical approximations to $\Phi_1^{75}(\mathcal{S}_1)$ and $\Phi_1^{75}(\mathcal{S}_2)$, using the parameters $K = 1000$ and $n = 75$ and barcode cutoffs of 0.375.

We then ran the following tests:

(1) We compared the empirical distance distributions $\mathcal{D}^2$ for $\mathcal{S}_1$ and $\mathcal{S}_2$ using the Kolmogorov-Smirnov statistic.
(2) We computed histograms from $\mathcal{D}^2$ for $\mathcal{S}_1$ and $\mathcal{S}_2$ (with 25 bins equally spaced over the maximum bounding region) and compared using the χ^2 test.
(3) Fixing a reference barcode B_1 with a single long bar, we computed the distance distribution $\mathcal{D}_{B_1}$ for $\mathcal{S}_1$ and $\mathcal{S}_2$ and repeated the comparisons above, using the Kolmogorov-Smirnov and χ^2 statistic (after forming histograms).

The results of these tests are summarized in Figure 4. We see that whereas the first two tests detect differences even with relatively small amounts of malicious noise, the third test is less sensitive and only begins to suggest rejection of the null hypothesis around at 2.0% or 2.5% noise added. (Note that in the third test, the median of the distribution is precisely the statistic MHD_k^n.) On the one hand,

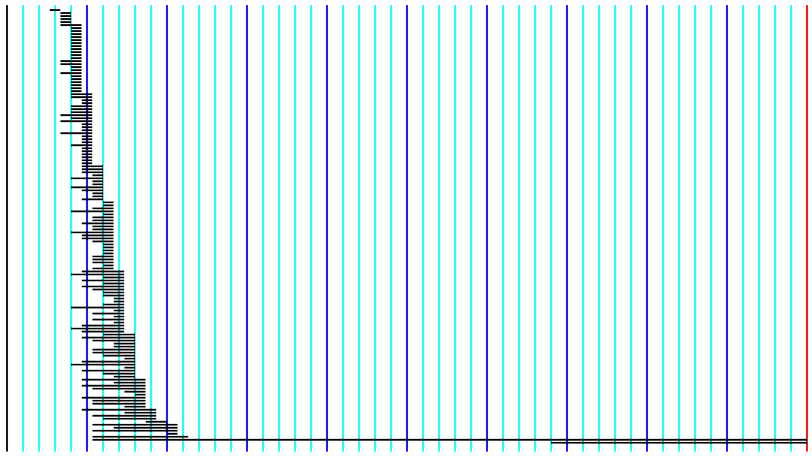

FIGURE 3. Barcode for the annulus plus diameter linkage via the Vietoris-Rips complex with 1000 points, showing 2 long bars. Horizontal scale goes from 0 to 0.375. (Vertical scale is not meaningful.)

Noise	χ^2 99%	χ^2 95%	χ^2 90%	KS 99%	KS 95%	KS 90%
0.0%	0.0	0.0	0.0	0.0	0.0	0.0
0.5%	0.05	0.05	0.05	0.2	0.2	0.2
1.0%	0.05	0.15	0.15	0.2	0.45	0.55
1.5%	0.15	0.2	0.35	0.25	0.4	0.65
2.0%	0.2	0.45	0.55	0.35	0.5	0.65
0.0%	0.0	0.0	0.0	0.0	0.0	0.0
0.5%	0.0	0.0	0.0	0.0	0.0	0.0
1.5%	0.0	0.0	0.05	0.0	0.05	0.1
2.0%	0.0	0.1	0.15	0.0	0.1	0.2
2.5%	0.1	0.15	0.2	0.35	0.55	0.65

FIGURE 4. Comparison tests for samples $\mathcal{S}_1$ and $\mathcal{S}_2 = \mathcal{S}_1 \cup X$, where $\mathcal{S}_1$ is a random sample of 1000 points from annulus A and X consists of a given proportion of random "noise" points along the diameter. Top: comparison tests for the $\mathcal{D}^2$ distribution (tests (1) and (2) in the text). Bottom: comparison tests for the $\mathcal{D}(B_1, -)$ distributions, where B_1 is the barcode with a single long bar (test (3) in the text).

these results provide context for interpreting the results of using the Kolmogorov-Smirnov and χ^2 statistics with more reasonable noise models (in other examples below). On the other hand, we see that using the third test we can extract robust topological information from the data.

Finally, for a different application of the χ^2 test to compare these distributions, we used k-means clustering to produce discrete distributions, as follows. Performing

k-means clustering on the empirical approximations to Φ_k^n for A indicated that the resulting distributions had nontrivial mass clustered in three regions: around a barcode B_0 with no long intervals, a barcode B_1 with one long interval, and a barcode B_2 with two long intervals. This led to the following test, which we repeated 1000 times.

(1) Fixing $K = 1000$, for $\mathcal{S}_1$ and $\mathcal{S}_2$, we counted the number of "long bars" (i.e., bars with length over a threshold of 0.125, which was determined by the k-means cluster centroids).
(2) We use the χ^2 test to determine if we can reject the hypothesis that the resulting histograms were drawn from the same distribution even at the 90% level.

The results were analogous to the more sensitive preceding experiments; at 1.0% noise added, we found that the χ^2 test *never* permitted rejection of the null hypothesis. (As an example, a sample distribution of masses on the centroid from a single run was 0.017, 0.983, and 0 for $\mathcal{S}_1$ and 0.020, 0.975, and 0.005 for $\mathcal{S}_2$.) On the other hand, at 2.0% noise added, we always rejected the null hypothesis. However, looking at the actual values, we see that even at 5.0% noise added, representative masses for $\mathcal{S}_2$ were 0.024, 0.827, and 0.149. We will see below how to use confidence intervals to extract precise inferences about the underlying homology from such data.

The example of the annulus also begins to illuminate a relationship between the distributional invariants and density filtering. Notice that the second interval at the bottom of Figure 3 starts somewhat later, reflecting a difference in average interpoint distance between the original samples and the additional points added. As a consequence, one might imagine that appropriate density filtering would also remove these points. On the one hand, in many cases density filtering is an excellent technique for concentrating on regions of interest. On the other hand, it is easy to construct examples where density filtering fails — for instance, we can build examples akin to the one studied here where the "connecting strip" has comparable density to the rest of the annulus simply by reducing the number of sampled points or by expanding the outer radius while keeping the number of sampled points fixed. In the former case our methods also degrade, but the latter produces results akin to the reported results above. More generally, studying distributional invariants (such as Φ_k^n) by definition allows us to integrate information from different density scales. In practice, we expect there to be a synergistic interaction between density filtering and the use of Φ_k^n; see Section 8 for an example of this interaction in practice.

Synthetic Example 2: Friendly circles. Next, we considered a somewhat more complicated example. The underlying metric measure space X is the subset of $\mathbb{R}^2$ specified as the union of the circle of radius 2 centered at $(0,0)$ and the circle of radius 1 centered at $(0.8,0)$, equipped with the intrinsic metric and the length measure. We sampled from X by choosing uniformly $\theta \in [0,2\pi]$ and assigning the indicated point to the first circle with probability $\frac{2}{3}$ and the second circle with probability $\frac{1}{3}$ (proportionally to their lengths). Convergence experiments analogous to those discussed in the previous example indicated that choosing subsamples of cardinality greater than roughly 500 resulted in good approximations to Φ_k^n.

Our experiments here are designed to indicate the robustness of our invariants to both Gaussian and uniform noise — the point of this example is that noise points

will introduce many classes in H_1 by linking the two circles where they are near one another. Once again, it is illuminating to simply begin with persistent homology computed from the entire subsample. We sampled 1500 points from X. We then consider two noise models:

(1) All points have ambient Gaussian noise added (i.e., we convolved with a Gaussian of mean 0 and covariance matrix $\sigma^2 I_2$ in $\mathbb{R}^2$). (See Figure 5.)
(2) A fraction of the points are replaced with uniform noise sampled from the bounding rectangle $[-2,2] \times [-2,2] \subset \mathbb{R}^2$. (See Figure 6.)

Computing the persistent homology from the Vietoris-Rips complex on these points, without noise, we saw the expected pair of long bars in the barcode for the first persistent homology group, computed using the Vietoris-Rips complex. With Gaussian noise, the results of computing the barcodes degraded as the width of the Gaussians increased; for example, when the width was $\sigma^2 = 0.1$ there were many long bars in the barcode. (We omit a graph of the barcode in the interest of space, as the phenomenon is similar to the uniform noise case.) And as uniform noise was added, the results of computing barcodes using the Vietoris-Rips complex degraded very rapidly, as we see in Figure 7 — there are many long bars. This is precisely what one would expect in light of the discussion in Section 4 and the geometry of the situation.

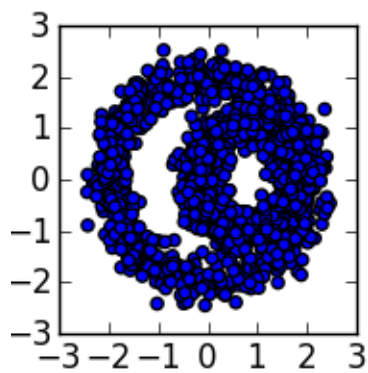

FIGURE 5. Two circles with Gaussian noise.

Even with only 10 noise points, we see 3 bars, and with 90 noise points there are 12. (These results were stable across different samples; we report results for a representative run.)

In contrast, we computed Φ_1^{300} for the same point clouds (i.e., the two circles plus varying numbers of noise points), using $K = 1000$ samples of size 300 and a cutoff of 0.75. The resulting empirical distributions had essentially all of their weight concentrated around barcodes with a small number of long intervals (revealed once again by k-means clustering). For the points in the empirical estimate of Φ_1^{300} around we counted the number of "long bars" with length over the threshold of 0.25 (again determined from the k-means centroids). The results are summarized in Figure 8 below.

A glance at the table shows that the majority of the weight is clustered around a barcode with 2 long bars and that the data overwhelming supports a hypothesis

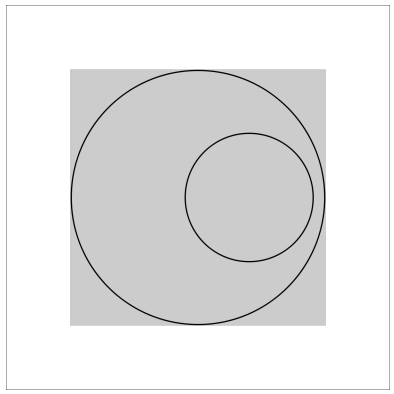

FIGURE 6. Two circles with uniform noise (indicated by gray box).

of ≤ 3 barcodes under all noise regimes. More precisely, the likelihood statistic of Section 6.1 allows us to evaluate the hypothesis H that the observed empirical approximation to Φ_k^n was drawn from an underlying barcode distribution with weight $\geq 5\%$ on barcodes with more than 3 long bars. In the strictest tests with 80 and 90 noise points, 31 out of 1000 samples were near barcodes with more than 3 long bars, and so we estimate that the probability of the distribution having $\geq 5\%$ of the mass at 4 or more barcodes as $\leq \mathrm{BD}(1000, 31, .05) < 0.22\%$. Put another way, we can reject the hypothesis that the actual distribution has more than 5% mass at 4 or more barcodes at the 99.7% level.

We also ran a similar experiment with Gaussian noise, looking at Φ_1^{300} and varying widths; the results are summarized in the Figure 9 below. As one would expect, sufficiently wide Gaussians cause the smaller circle to appear to be a (contractible) disk attached to the larger circle.

Spheres and tori in $\mathbb{R}^3$. We now turn to more realistic synthetic examples that are less easily summarized (and better represent the ambiguity present in the typical application of topological data analysis). We studied two standard geometric examples of smooth manifolds.

(1) Two-dimensional spheres of varying radii r, which we denote $S(r)$,
(2) Tori of inner radius r and outer radius R for varying parameter values which we denote $T(r, R)$ (see figure 10).

These examples have interestingly different characteristics; detecting the sphere's top homology class is relatively easy even in the face of noise, whereas noise can introduce many spurious homology classes in degree 1. In contrast, the torus $T(0.5, 1)$ is sufficiently different in the scale of its two axes that detecting the degree 2 homology class and both degree 1 homology classes is quite challenging.

There are various reasonable choices to make about how to sample from these objects. In our experiments, we use the intrinsic metric and sample using the area measure in each case:

(1) To draw a uniform point on the sphere using the area measure, we draw points z_1, z_2, z_3 from the standard normal distribution and consider the point $(\frac{z_1}{\sqrt{z_1^2+z_2^2+z_3^2}}, \frac{z_2}{\sqrt{z_1^2+z_2^2+z_3^2}}, \frac{z_3}{\sqrt{z_1^2+z_2^2+z_3^2}})$.

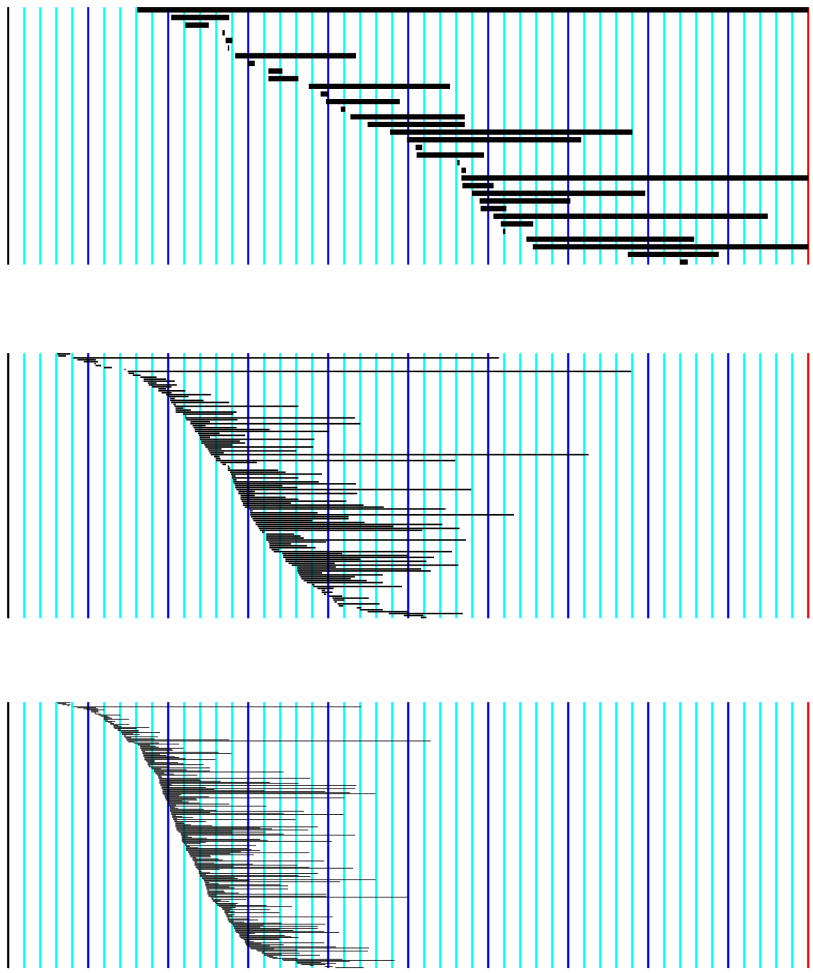

FIGURE 7. Barcode for two circles with 10, 50, and 90 noise points. Horizontal scale goes from 0 to 0.75. (Vertical scale is not meaningful.)

(2) To draw a uniform point on the torus using the area measure, we parametrize the torus as

$$(\theta, \psi) \mapsto ((R + r\cos(\theta))\cos(\psi), (R + r\cos(\theta))\sin(\psi), r\sin(\theta)),$$

for $0 \leq \theta, \psi \leq 2\pi$ and use the rejection sampling procedure explained in [20, 2.2]. (Note that drawing θ and ψ uniformly in $[0, 2\pi]$ does not work.)

We again work with two noise models, adding both Gaussian noise (by convolving with a mean 0 Gaussian with covariance matrix $\sigma^2 I_3$ in $\mathbb{R}^3$) to all points and replacing some of the points with uniform noise (obtained from uniform samples in $\mathbb{R}^3$ using the bounding box $[-2, 2]^3$) to the samples. We note that these two noise

Number of noise pts	0 bars	1 bars	2 bars	3 bars	4 bars	5 bars
0	0	303	696	1	0	0
10	0	305	589	106	0	0
20	0	278	590	132	0	0
30	0	285	594	119	2	0
40	1	259	584	149	6	1
50	0	289	553	154	4	0
60	0	254	591	146	7	2
70	0	277	564	154	5	0
80	1	229	543	196	29	2
90	0	229	533	207	28	3

FIGURE 8. Distribution summaries for Φ_1^{300} in "Friendly Circles" example for number of long bars occurring in 1000 tests with given number of noise points added.

σ^2	0 bars	1 bars	2 bars	3 bars	4 bars	5 bars
0.05	2	59	930	9	0	0
0.075	44	351	585	20	0	0
0.1	204	537	249	10	0	0

FIGURE 9. Distribution summaries for Φ_1^{300} in "Friendly Circles" example for number of long bars occurring in 1000 tests with Gaussian noise added of mean 0 and covariance $\sigma^2 I_2$.

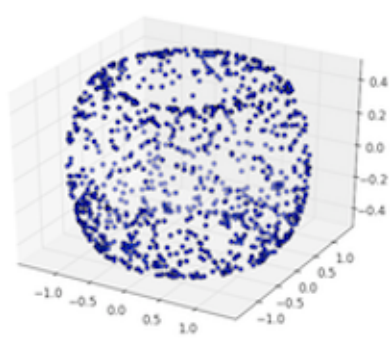

FIGURE 10. Torus $T(0.5, 1)$

models are somewhat different in character; the Gaussian noise affects all points, whereas the uniform noise corrupts some fraction of the total number of points.

Our first set of experiments studied the rate of convergence in Corollary 5.5; our methodology is the same as in the previous section, and we find that acceptable minimum cardinalities for $S \subset X$ in order for $\Phi_k^n(S)$ and $\Phi_k^n(X)$ (for varying n)

Shape	Noise	χ^2 99%	χ^2 95%	χ^2 90%	KS 99%	KS 95%	KS 90%
S(1)	1%	0.0	0.0	0.0	0.0	0.05	0.05
S(1)	5%	0.0	0.0	0.0	0.0	0.0	0.05
S(1)	10%	0.0	0.0	0.1	0.0	0.1	0.1
S(1)	20%	0.0	0.1	0.15	0.1	0.2	0.35
T(0.5,1)	1%	0.0	0.0	0.0	0.0	0.0	0.0
T(0.5,1)	5%	0.0	0.0	0.0	0.0	0.0	0.05
T(0.5,1)	10%	0.0	0.05	0.05	0.0	0.05	0.1
T(0.5,1)	20%	0.0	0.1	0.2	0.15	0.2	0.3

FIGURE 11. Comparisons tests for the sphere (for Betti 2) and torus (for Betti 1) for distributions with uniform noise replacing a fraction of the points compared against the noiseless distributions.

to be indistinguishable to the χ^2 and Kolmogorov-Smirnov tests are around 1000 for the sphere and 2000 for the torus. We fix $K = 1000$ throughout. We use these results as a guide when carrying out experiments analyzing the metric measure spaces in this region.

Next, in order to explore how the inference procedures described in Section 6 can be used in the context of our running examples in $\mathbb{R}^3$, we carry out the following different experiments, again using Algorithms 7.1 and 7.2 as the base.

(1) We use the Kolmogorov-Smirnov and χ^2 tests to study how the distribution $\mathcal{D}^2$ of distances induced from Φ changes as noise is added.
(2) We use estimates of MHD_k^n both to extract information about the salient topological features of the sphere and the torus and also to test the robustness of this invariant to added noise.

We began by looking at what the Kolmogorov-Smirnov and χ^2 tests tell us about the sphere and torus. Working with the uniform noise model, we used subsamples of 1000 points for the sphere and considered Φ_2^{150} as our base. For the torus, we used subsamples of 2000 points and Φ_1^{150} for our base. We replaced an increasing fraction of the points with noise and compared to the distribution from the underlying (noiseless) model, with the results summarized in Figure 11. Here the percentage in the table once again indicates the fraction of runs in which we could reject the null hypothesis of the same distribution at the indicated significance level. A clear conclusion to draw is that Φ_k^n is relatively insensitive to even large amounts of uniform noise. In contrast, when the corresponding experiments were run with the Gaussian noise model, we found that there was a threshold effect; for noise widths smaller than roughly $\sigma^2 = 0.05$, the distributions could not be distinguished by these tests, but for larger noise widths they basically always appeared to be distinct.

Although the previous experiments indicate the degree to which Φ_k^n is robust against noise, in practice it is more likely that we will want to extract information about easily expressed hypotheses concerning the rank of the homology groups of the underlying space. To this end, we consider MHD_k^n with regards to various reference barcodes; let $m[a,b)$ denote the barcode consisting of m copies of the interval $[a,b)$. We used Algorithms 7.1 and 7.2 to compute MHD_k^n and we used

the asymptotic estimates of Definition 6.8 to produce the confidence intervals. We chose subsets of size 1000 to subsample from.

We begin by considering results for uniform noise in a box, as a reference benchmark; the results are summarized in Figure 12. We then compute for the sphere; the results are summarized in Figure 13 below. Finally, we did the computations for the torus; the results are summarized in Figure 14 below. We obtained the reference barcodes by inspection of a single run; this procedure is a proxy for the kind of exploratory data analysis that we expect would generate the hypotheses to test using our test statistics. The confidence intervals in the table were generated by using 100 samples; the reported results are representative for these parameter settings. We also ran a number of experiments with Gaussian noise as well. In the interest of space, we report only the results on the torus, which are summarized in Figure 15, as these are representative.

Before we begin to discuss these results, a few observations about the data sets are in order. We expect that the sphere should be a relatively easy example; uniform noise is unlikely to interfere with the top-dimensional homology class. This expectation is borne out by simply computing the persistent homology using 1000 points — even with 10% uniform noise added, we see a single much longer bar. (We omit the picture of this.) In this situation, we regard our experiments as validating the use of MHD_k^n to make precise statistical statements about topological hypotheses. In contrast, the torus $T(0.5, 1)$ is a difficult test; the scale of the two one-cycles is different, and we need a large number of points in order to resolve them both. When running the persistent homology using all 1000 points, even tiny amounts of uniform noise cause substantial disruptions in the results, i.e., many long bars. (We again omit the picture of this.) As a consequence, in the presence of noise, working without the statistical methodology makes it basically impossible to draw conclusions about the data.

For the sphere, the measured results indicate that MHD_2^{150} does an excellent job of detecting the class in dimension 2. Specifically, until the noise reaches 20%, the confidence interval for the hypothesis $1[0.4, 0.55)$ is the closest to 0 and does not overlap with the other confidence intervals. When confidence intervals for different population quantities do not overlap, the difference between the two is statistically significant at the 99% level. We could also use Monte Carlo simulation to estimate the difference between the medians (for the two hypotheses), if a more refined test statistic explicitly comparing the hypotheses was desired. The measured results do not detect any classes in dimension 1, even with really substantial amounts of noise. (The results are comparable to the results for the box in dimension 1.)

For the torus, we begin by discussing the case of uniform noise. In dimension 1 we see that both 1 and 2 bar variants are close to the observed data. When we perform Monte Carlo simulation to obtain confidence intervals for the difference between the medians, the 95% intervals contain 0 — this suggests that we cannot distinguish between the two hypotheses with this test statistic. One interpretation of this result is that there are in fact a larger number of long bars, and indeed inspection of the barcode results reflect approximately 5 "long" bars. It is encouraging that our results are very robust in the face of large amounts of uniform noise, however. We can obtain better results by increasing the number of samples points; when using MHD_1^{500} and a subsample of size 10000, the medians and confidence intervals for 2 bars is substantially smaller than for 1 bar or 3 bars — the difference is now

k	m	Median	95% Confidence interval
1	0	0.09	[0.0875,0.0925]
1	1	0.08	[0.0775,0.0825]
1	2	0.075	[0.075,0.0775]
2	0	0.0175	[0.0125,0.025]
2	1	0.085	[0.075,0.1]
2	2	0.095	[0.085,0.115]

FIGURE 12. Confidence intervals for MHD_k^{150} applied to uniform noise in $[-2,2]\times[-2,2]$ for reference bar codes $m[0.40.55)$.

Shape	Noise	k	m	Median	95% Confidence interval
S(1)	0%	1	0	0.0925	[0.09,0.0975]
S(1)	0%	1	1	0.195	[0.195,0.2]
S(1)	0%	1	2	0.205	[0.205,0.21]
S(1)	5%	1	0	0.095	[0.0925,0.1]
S(1)	5%	1	1	0.175	[0.17,0.175]
S(1)	5%	1	2	0.185	[0.185,0.19]
S(1)	10%	1	0	0.0975	[0.095,0.1025]
S(1)	10%	1	1	0.15	[0.14,0.165]
S(1)	10%	1	2	0.18	[0.175,0.185]
S(1)	20%	1	0	0.0975	[0.0925,0.1025]
S(1)	20%	1	1	0.115	[0.105,0.12]
S(1)	20%	1	2	0.145	[0.135,0.155]
S(1)	0%	2	0	0.07	[0.065, 0.075]
S(1)	0%	2	1	0.02	[0.02, 0.025]
S(1)	0%	2	2	0.075	[0.075, 0.075]
S(1)	5%	2	0	0.065	[0.0625, 0.07]
S(1)	5%	2	1	0.025	[0.015, 0.03]
S(1)	5%	2	2	0.075	[0.075, 0.075]
S(1)	10%	2	0	0.06	[0.0575,0.065]
S(1)	10%	2	1	0.03	[0.025,0.035]
S(1)	10%	2	2	0.075	[0.075,0.08]
S(1)	20%	2	0	0.0525	[0.045,0.0575]
S(1)	20%	2	1	0.045	[0.04,0.065]
S(1)	20%	2	2	0.075	[0.075,0.075]

FIGURE 13. Confidence intervals for MHD_k^{150} for reference barcodes $m[0.4,0.55)$ applied to the sphere with a given percentage of points replaced with uniform noise.

statistically significant at the 99% level. (For reasons of space we omit reporting the specific tables.)

For the torus with Gaussian noise, the results admit a comparable analysis, with the exception of the fact that we see a substantial degradation as the width increases (and at noise of width $\sigma^2 = 0.1$ our procedures are basically useless).

Shape	Noise	k	m	Median	95% Confidence interval
T(0.5,1)	0%	1	0	0.1575	[0.1525,0.16]
T(0.5,1)	0%	1	1	0.0925	[0.09,0.0975]
T(0.5,1)	0%	1	2	0.1	[0.0975,0.1]
T(0.5,1)	5%	1	0	0.15	[0.145,0.1525]
T(0.5,1)	5%	1	1	0.095	[0.0925,0.0975]
T(0.5,1)	5%	1	2	0.0975	[0.095,0.1]
T(0.5,1)	10%	1	0	0.145	[0.14,0.15]
T(0.5,1)	10%	1	1	0.0975	[0.0925,0.1025]
T(0.5,1)	10%	1	2	0.0975	[0.095,0.1]
T(0.5,1)	20%	1	0	0.14	[0.1375,0.1425]
T(0.5,1)	20%	1	1	0.0925	[0.09,0.095]
T(0.5,1)	20%	1	2	0.0975	[0.0925,0.1]

FIGURE 14. Confidence intervals for MHD_k^{150} for reference barcodes $m[0.3, 0.55)$ applied to the torus with a given percentage of points replaced with uniform noise.

Shape	σ^2	k	m	Median	95% Confidence interval
T(0.5,1)	0.01	1	0	0.145	[0.1425,0.15]
T(0.5,1)	0.01	1	1	0.085	[0.0825,0.0875]
T(0.5,1)	0.01	1	2	0.115	[0.11,0.125]
T(0.5,1)	0.05	1	0	0.1275	[0.1225,0.135]
T(0.5,1)	0.05	1	1	0.0775	[0.075,0.0825]
T(0.5,1)	0.05	1	2	0.11	[0.105,0.115]
T(0.5,1)	0.1	1	0	0.095	[0.0925,0.1]
T(0.5,1)	0.1	1	1	0.0875	[0.085,0.0925]
T(0.5,1)	0.1	1	2	0.105	[0.105,0.11]

FIGURE 15. Confidence intervals for MHD_k^{150} for reference barcodes $m[0.3, 0.55)$ applied to the torus with Gaussian noise of mean 0 and covariance matrix $\sigma^2 I_3$.

8. Application: confidence intervals for the natural images dataset

One of the most prominent applications of persistent homology in topological data analysis is the study of the natural images dataset described in [27]. This data consists of 3×3 patches sampled from still photographs of "natural" scenes (i.e., pictures of rural areas without human artifacts). The results of Carlsson, Ishkhanov, de Silva, and Zomorodian [10] extract topological signals from this data set which can be interpreted in terms of collections of patches which are known to be meaningful based on the neurophysiology of the eye. The goal of this section is to apply our statistical methodology to validate the conclusions of their work.

8.1. **Setup.** We compute the confidence intervals based on MHD_k^n for a subset of patches from the natural images dataset as described in [10]. We briefly review the setup. The dataset consists of 15000 points in $\mathbb{R}^8$, generated as follows. From the

natural images, 3×3 patches (dimensions given in pixels) were sampled and the top 30% with the highest contrast were retained. These patches were then normalized twice, first by subtracting the mean intensity and then scaling so that the Euclidean norm is 1. The resulting dataset can be regarded as living on the surface of an S^7 embedded in $\mathbb{R}^8$. After performing density filtering (with a parameter value of $k = 15$; refer to [10] for details) and randomly selecting 15000 points, we are left with the dataset $\mathcal{M}(15, 30)$. At this density, one tends to see a barcode corresponding to 5 cycles in the H_1. In the Klein bottle model, these cycles are generated by three circles, intersecting pairwise at two points (which can be visualized as unit circles lying on the xy-plane, the yz-plane, and the xz-plane).

8.2. **Results.** We computed empirical approximations to $\Phi_1^{500}(\mathcal{M}(15, 30))$ using Algorithm 7.1, with $K = 1000$ and using a barcode cutoff of 2 (we used the value reported in [10] as the maximal filtration value). We found that (after applying k-means clustering, as above) the weight was distributed as 0.1% with one long bar, 1.1% with two long bars, 7.4% with three long bars, 34.2% with four long bars, and 57.2% with five long bars. (Here the threshold for a long bar was 1.) Analyzing likelihood test statistics as in Section 6.1, we find that the underlying distribution has at least 95% of its mass on two, three, or four bars at the the 99.7% confidence level.

We also analyze the results using MHD. We use as the hypothesis barcode the multi-set $5[0, 2) = \{[0, 2), [0, 2), [0, 2), [0, 2), [0, 2)\}$. We find using the nonparametric estimate from Definition 6.8 that the 95% confidence interval for $\mathrm{MHD}_1^{500}(\mathcal{M}(15, 30)$ is $[0.442, 0.476]$. The 99% confidence interval for $\mathrm{MHD}_1^{500}(\mathcal{M}(15, 30))$ is $[0.436, 0.481]$. These results represents high confidence for the data to be further than 0.442 but closer than 0.476 to the reference barcode. On the other hand, when we compute the confidence intervals using the reference barcode the empty set, we find that both endpoints for the 95% and 99% confidence intervals are the cutoff value of 2. We find the same results for hypothesis barcodes with ℓ bars $[0, 2)$ for $0 \leq \ell \leq 10, \ell \neq 5$. In particular, this means that the differences between the distance to the 5 bar hypothesis and any other is statistically significant at the 99% level.

We interpret these results to suggest that the hypothesis barcode is consistent with the underlying distribution amongst barcode distributions that put all of their mass on a single barcode. Of course, these results also suggest that when sampling at 500 points, we simply do not expect to see a distribution that is heavily concentrated around a single barcode. In the next subsection, we discuss the use of the witness complex, which does result in such a narrow distribution.

Remark 8.1. To validate the non-parametric estimate of the confidence interval, we also used bootstrap resampling to compute bootstrap confidence intervals. Although we do not justify or discuss further this procedure herein, we note that we observed the reassuring phenomenon that the bootstrap confidence intervals agreed closely with the non-parametric estimates for both the 95% confidence intervals and the 99% confidence intervals in each instance.

8.3. **Results with the witness complex.** Because of the size of the datasets involved, the analysis performed in [10] used the weak witness complex W rather than the Vietoris-Rips complex VR. The weak witness complex for a metric space (X, ∂) depends on a subset $X_0 \subset X$ of witnesses; the size of the complexes is controlled by $|X_0|$ and not $|X|$.

Definition 8.2. For $\epsilon \in \mathbb{R}$, $\epsilon \geq 0$ and witness set $X_0 \subset X$, the weak witness complex $\mathrm{W}_\epsilon(X, X_0)$ is the simplicial complex with vertex set X_0 such that $[v_0, v_1, \ldots, v_n]$ is an n-simplex when for each pair v_i, v_j, there exists a point $p \in X$ (a witness) such that the distances $\partial(v_i, p) \leq \epsilon$.

When working with the witness complex, we adapt our basic approach to study the induced distribution on barcodes which comes from fixing the point cloud and repeatedly sampling a fixed number of witnesses. The theoretical guarantees we obtained for the Vietoris-Rips complex in this paper do not apply directly; we intend to study the robustness and asymptotic behavior of this process in future work. Here, we report preliminary numerical results.

Specifically, we again computed empirical approximations to $\Phi_1^n(\mathcal{M}(15, 30))$ using Algorithm 7.1, with $K = 1000$ and using a barcode cutoff of 2. However, to produce the underlying complex, we use the n points for each subsample as the landmark points X_0 in the construction of the witness complex rather than as the vertices for the Vietoris-Rips complex.

We use as the hypothesis barcode the multi-set $5[0, 2)$ as above. We found using the non-parametric estimate of Definition 6.8 that the 95% confidence interval for $\mathrm{MHD}_1^{100}(\mathcal{M}(15, 30))$, is $[0.024, 0.027]$. The 99% confidence interval for $\mathrm{MHD}_1^{100}(\mathcal{M}(15, 30))$ was also $[0.024, 0.027]$. When we computed the 95% confidence interval for $\mathrm{MHD}_1^{150}(\mathcal{M}(15, 30)$ we obtained $[0.021, 0.023]$. The 99% confidence interval for $\mathrm{MHD}_1^{150}(\mathcal{M}(15, 30))$ was $[0.021, 0.024]$. This represents high confidence for the data to be further than 0.021 (for Φ_1^{150}) and 0.024 (for Φ_1^{100}) but closer than 0.024 (for Φ_1^{150}) and 0.027 (for Φ_1^{100}) to the reference barcode. We obtained essentially the same results MHD_1^{500} as for MHD_1^{150}. On the other hand, when using hypothesis barcodes with ℓ bars $[0, 2)$ for $0 \leq \ell \leq 10, \ell \neq 5$, the confidence intervals start and end at 2. Again, this means that the difference between the distances to the 5-bar hypothesis and the other hypotheses is statistically significant at the 99% level. We interpret these results to mean that the underlying distribution is essentially concentrated around the hypothesis barcode; the distance of 0.025 is essentially a consequence of noise.

Remark 8.3. In contrast, when we compute $\mathrm{MHD}_1^{25}(\mathcal{M}(15, 30))$ (using the same experimental procedure as above), we find the confidence interval is $[1.931, 1.939]$. When we compute $\mathrm{MHD}_1^{75}(\mathcal{M}(15, 30))$, we find that the confidence interval is $[1.859, 1.866]$. This represents high confidence that MHD_1^{25} and MHD_1^{75} are far from this reference barcode, which in light of the confidence intervals above for MHD_1^{150} and MHD_1^{500} appear to indicate that samples sizes 25 and 75 are too small.

References

[1] R. J. Adler, O. Bobrowski, and S. Weinberger. Crackle: The persistent homology of noise. arXiv:1301.1466, 2013.

[2] O. Bobrowski and R. J. Adler. Distance functions, critical points, and topology for some random complexes. arXiv:1107.4775, 2011.

[3] R.J. Adler, O. Bobrowski, M. S. Borman, E. Subag, and S. Weinberger. Persistent homology for random fields and complexes. arXiv:1003.1001, 2010.

[4] A. J. Blumberg, M. A. Mandell, Resampling methods for estimating persistent homology. In preparation.

[5] P. Bubenik. Statistical topology using persistence landscapes. arXiv:1207.6437, 2012.

[6] P. Bubenik, G. Carlsson, P. T. Kim, and Z.-M. Luo. Statistical topology via Morse theory persistence and nonparametric estimation. In *Algebraic methods in statistics and probability II*, Contemp. Math., 516:75–92. Amer. Math. Soc., Providence, RI, 2010.

[7] P. Bubenik and J. A. Scott. Categorication of persistent homology. arXiv:1205.3669, 2012

[8] F. Cagliari, M. Ferri and P. Pozzi. Size functions from the categorical viewpoint. *Acta Appl. Math.*, 67:225–235, 2001.

[9] C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Statist.* 5:1394–1423, 2011.

[10] G. Carlsson, T. Ishkhanov, V. de Silva., A. Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76 (1):1–12, 2008.

[11] G. Carlsson and F. Memoli. Characterization, stability, and convergence of hierarchical clustering methods. *Journal of machine learning research*, 11:1425–1470, 2009.

[12] G. Carlsson and V. De Silva. Zigzag persistence. *Foundations of computational mathematics*, 2009.

[13] F. Chazal, D. Cohen-Steiner, and Q. Merigot. Geometric inference for probability measures. *Found. Comp. Math.* 11(6):733–751, 2011.

[14] F. Chazal, V. De Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. arXiv:1207.3674, 2012.

[15] F. Chazal, V. De Silva, and S. Oudot. Persistence stability for geometric complexes. arXiv:1207.3885, 2013.

[16] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Memoli, S. Oudot. Gromov-Hausdorff stable signatures for shapes using persistence. *Comput. Graph. Forum*, 28(5):1393–1403, 2009.

[17] E. Gine and Z. Chen. Another approach to asymptotics and bootstrap of randomly trimmed means. *Ann. of the Institute of Stat. Math.* 56:771–790, 2004.

[18] M. K. Chung, P. Bubenik, and P. T. Kim. Persistence diagrams in cortical surface data. In *Information Processing in Medical Imaging (IPMI) 2009*, 5636:386–397, Lecture Notes in Computer Science, 2009.

[19] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Disc. and Comp. Geom.*, 37(1):103–120, 2007.

[20] P. Diaconis, S. Holmes, M. Shahshahani. Sampling from a manifold. arXiv:1206.6913, 2011.

[21] S. Gadgil and M. Krishnapur. Lipschitz correspondence between metric measure spaces and random distance matrices. arXiv:1110.6333, 2011.

[22] D Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have L_p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139, 2010.

[23] W.J. Conover. Practical nonparametric statistics. Wiley, 3rd edition, 1999.

[24] H. A. David and H. N. Nagaraja. Order statistics. Wiley, 3rd edition, 2003.

[25] H. Edelsbrunner and J. Harer. Persistent homology — a survey. in *Contemporary Mathematics*, 2008.

[26] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Disc. and Comp. Geom.*, 28:511-533, 2002.

[27] J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, 265:359–366, 1998.

[28] P. Frosini and C. Landi. Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis* 9:596–603, 1999.

[29] Free Software Foundation. `http://www.gnu.org/software/gsl/`, 2013.

[30] A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces. *Prob. Theo. Rel. Fields*, 145(1):285–322, 2009.

[31] J.A. Hartigan. Consistency of singe linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76(374):388–394, 1981.

[32] J.A. Hartigan. Statistical theory in clustering. *J. Classification*, 2(1):63–76, 1985.

[33] M. Kahle. Random geometric complexes. *Disc. and Comp. Geometry*, 45(3):553–573, 2011.

[34] M. Kahle and E. Meckes. Limit theorems for Betti numbers of random simplicial complexes. Preprint, arXiv:1009.4130, 2010.

[35] J. Latschev. Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold. *Archiv der Math.*, 77(6):522–528, 2001.

[36] F. Memoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 2011.

[37] Y. Mileyko, S. Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 2012.
[38] V. Nanda. Perseus software. `http://www.math.rutgers.edu/~vidit/perseus/index.html`, 2013.
[39] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Disc. and Comp. Geometry*, 39:419–441, 2008.
[40] D. Pollard. Convergence of stochastic processes. Springer-Verlag, 1984.
[41] V. Robins. Toward computing homology from finite approximations. *Topology Proceedings*, 24, 1999.
[42] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Proc. of Symp. on Point-Based Graph.*, 2004.
[43] K-T. Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196(1):65–131, 2006.
[44] M. Tsao and J. Zhou. A nonparametric confidence interval for the trimmed mean. J. of Nonparametric Stat. 14(6):665–673, 2002.
[45] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Frechet means for distributions of persistence diagrams. arXiv:1206.2790.
[46] A. V. van der Waart. Asymptotic statistics. Cambridge University Press, 1998.
[47] A. Zomorodian and G. Carlsson. Computing persistent homology. *Proc. of 20th ACM Symp. Comp. Geo.*, 2004.

Department of Mathematics, University of Texas at Austin, Austin, TX 78712
E-mail address: `blumberg@math.utexas.edu`

Department of Mathematics, University of Texas at Austin, Austin, TX 78712
E-mail address: `igal@math.utexas.edu`

Department of Mathematics, Indiana University Bloomington, IN 47405
E-mail address: `mmandell@indiana.edu`

Department of Mathematics, University of Texas at Austin, Austin, TX 78712
E-mail address: `mpancia@math.utexas.edu`