An introduction to Topological Object Data Analysis

Adam Dixon, Victor Patrangenaru and Chen Shen

Abstract. Summary and analysis are important foundations in Statistics, but typical methods may prove ineffective at providing thorough summaries of complex object data. Topological data analysis (TDA) (also called topological object data analysis (TODA) when applied to object data) provides additional topological summaries, such as the persistence diagram and persistence landscape, that can be useful in distinguishing distributions based on data sets. The main tool is persistent homology, which tracks the births and deaths of various homology classes as one steps through a filtered simplicial complex that covers the sample. The persistence diagrams and landscapes can also be used to provide confidence sets for "significant" features and two-sample tests between groups. An example of application is provided via analyzing mammogram images for patients with benign and malignant masses.

M.S.C. 2020: 55N31, 62H35, 62R30, 62R40.
Key words: Topological object data analysis; persistent homology; persistence diagram; persistence landscape.

1 Introduction

Summary and analysis, which often complement each other, are two cornerstones of Statistics. In summarization, the goal is to reduce a presumably large or complex data set into several measurements that describe the data in some way. For example, if the data is quantitative, then its sample mean describes its location, and its sample standard deviation describes its dispersion. However, with complex data (and even some not-so-complex data), measurements of location and dispersion may provide a scant description.

As an example, consider the data sets in Figure 1. Some typical summary statistics used to describe this data might be the sample mean vectors and sample covariance matrices displayed in Table 1.

Given the sample mean vectors and sample covariance matrices do not differ much in value, it would be reasonable to think these two data sets are similar to one another.

BSG Proceedings 28. The International Conference "Differential Geometry - Dynamical Systems" DGDS-2021, 27-29 August 2020, Mangalia-Romania, pp. 30-44.

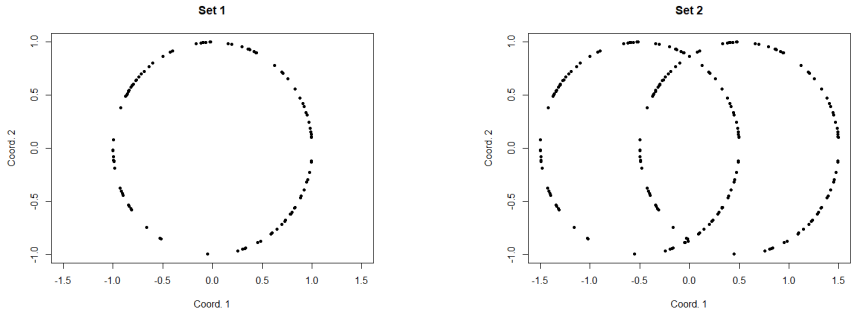

Figure 1: Circles ($n = 100$)

Table 1: Circle data summary statistics

Set	Sample Mean	Sample Cov.
1	$(0.0341 \quad 0.1038)^T$	$\begin{pmatrix} 0.5597 & -0.0768 \\ -0.0768 & 0.4384 \end{pmatrix}$
2	$(0.0341 \quad 0.1038)^T$	$\begin{pmatrix} 0.8081 & -0.0764 \\ -0.0764 & 0.4362 \end{pmatrix}$

But even a cursory viewing of Figure 1 would suggest they ought to be distinguished from one another.

What additional measurements can further aid in distinguishing these two sets apart? One idea is to use each sample's *topology*, specifically the topology of the support from which the data was sampled. As seen later, the topological measurements considered here are the *Betti numbers of the homology groups,* that is, counts of various dimensional "holes."

For example, the *point clouds* in Figure 1 can be considered samples of one circle and a union of two circles, respectively. The support of Set 1 has one connected component (0-dimensional hole) and one loop (1-dimensional hole), whereas the support of Set 2 has one connected component but three loops. Therefore, the two supports (and hence the samples) can be distinguished through their homological features.

The issue, of course, is that the support of a general sample is likely assumed to be unknown. Thus, it may not be reasonable to assume a particular type of support as was done for the Circles data. Rather, one can use the data itself to estimate the homology of the support through *persistent homology.* The following collection of tools and methods that use persistent homology are grouped under the umbrella term *topological data analysis* (TDA), which for this work's purposes will be referred to as *topological object data analysis* (TODA) when applied to object data.

These TODA methods can be applied beyond summarization to analysis. For example, confidence sets for *persistence diagrams*, a representation of the persistent homology classes, can be used to distinguish between "significant" and "non-significant" features, and *persistence landscapes*, another representation for the persistence diagram, can be used to test for differences between two groups. Some real-world applications include analyzing the Cosmic Web [21, 23], detecting curvature in certain

geometric constructions [3], and distinguishing between two leaves through sets of images [18] or between patients with or without glioblastoma multiforme through CT images [20].

In the following, Section 2 covers the basics of homology, including simplicial and singular homology (Section 2.1), the various simplicial complex constructions used in TODA (Section 2.2), and persistent homology, the main tool (Section 2.3). Section 3 details the construction of confidence sets for persistence diagrams, and Section 4 covers two-sample hypothesis testing with persistence landscapes. To conclude, Section 5 provides an example application by using TODA to summarize and analyze mammogram images corresponding to patients with benign and malignant masses.

2 Homology

For homology theory and more background on Algebraic Topology, the reader is referred to Hatcher[15]. In what follows we present some basic definitions and results necessary for a comprehension of the persistent homology and its connection to Statistics via Bubenik's landscapes (see [2]).

2.1 A Review of simplicial and singular homology

Given $v_0, v_1, \ldots, v_m \in \mathbb{R}^d$ that are affinely independent (i.e., $v_1 - v_0, v_2 - v_0, \ldots, v_m - v_0$ are linearly independent), an *m-simplex* $\sigma = [v_0\ v_1\ \cdots\ v_m]$ is the convex hull of $\{v_0, v_1, \ldots, v_m\}$ along with the order of the vertices' appearances. In particular, the *standard m-simplex* Δ^m is given by

$$\Delta^m = \Big\{(t_0, t_1, \ldots, t_m) \in \mathbb{R}^{m+1} : t_i \geq 0\ \forall\ i \text{ and } \sum_{i=0}^{m} t_i = 1\Big\}$$

A *proper face* of an m-simplex σ is the $(m-1)$-simplex $[v_0\ v_1\ \cdots\ \hat{v}_i\ \cdots\ v_m]$ formed by removing one of the vertices and preserving the order of those remaining. A *simplicial complex* K is a collection of simplices such that

- if $\sigma \in K$, then every proper face of σ is also in K; and
- if $\sigma_1, \sigma_2 \in K$ and $\sigma_1 \cap \sigma_2 \neq \varnothing$, then $\sigma_1 \cap \sigma_2 = \sigma_1$, $\sigma_1 \cap \sigma_2 = \sigma_2$ or $\sigma_1 \cap \sigma_2$ is a proper face of σ_1 and σ_2.

Given a simplicial complex K, the m-simplices in K form a basis for the free Abelian group $C_m(K)$ of *m-chains*, which are the finite formal sums of the m-simplices in K with integer coefficients. The *boundary operator* $\partial_m : C_m(K) \to C_{m-1}(K)$ is a group homomorphism that maps each m-chain to its boundary $(m-1)$-chain in $C_{m-1}(K)$. As ∂_m is completely characterized by its operation on the m-simplices that generate $C_m(K)$, one has for $[v_0\ v_1\ \cdots\ v_m] \in K$

$$\partial_m\big([v_0\ v_1\ \cdots\ v_m]\big) = \sum_{i=0}^{m} (-1)^m [v_0\ v_1\ \cdots\ \hat{v}_i\ \cdots\ v_m]$$

The elements of $\operatorname{Ker}\partial_m$ are called *cycles*, and the elements of $\operatorname{Im}\partial_{m+1}$ are called *boundaries*. It can be shown that $\partial_m\partial_{m+1} = 0$, which implies $\operatorname{Im}\partial_{m+1} \subseteq \operatorname{Ker}\partial_m$. Consequently, the *m-th simplicial homology group* $H_m(K)$ is defined to be

$$H_m(K) = \operatorname{Ker}\partial_m / \operatorname{Im}\partial_{m+1}$$

The various simplicial homology groups of K $\{H_0(K), H_1(K), \ldots\}$ detect the "holes" of various dimensions in K, and the *Betti numbers* $\beta_k = \operatorname{rank}(H_k(K))$ count the number of the k-dimensional holes in K, specifically. In fact, if the chain groups are generated using coefficients in $\mathbb{Z}_2$ instead of $\mathbb{Z}$, then the simplicial homology groups $H_k(K)$ have a vector space structure, which implies $\beta_k = \dim\big(H_k(K)\big)$.

Thus far, homology groups have only been considered for simplicial complexes embedded in Euclidean space, but they can be extended to any topological space through *singular homology*. If (M, τ_M) is a topological space, then the *singular m-simplices* of X are the continuous maps $\sigma : \Delta^m \to X$. The *singular m-chains* $S_m(X)$ are the finite formal sums of singular m-simplices in X with integer coefficients. The boundary operator $\partial_m : S_m(X) \to S_{m-1}(X)$ is defined similarly to its simplicial counterpart, namely if σ is a singular m-simplex, then

$$\partial_m(\sigma) = \sum_{i=0}^{m} (-1)^i \sigma \circ \iota_i$$

where ι_i is the inclusion of Δ^{m-1} in Δ^m as the ith face with the ordering of the vertices preserved. Once again, one has $\partial_m\partial_{m+1} = 0$ for all m, which suggests the singular homology groups are also defined similarly to the simplicial homology groups – that is, $H_m(X) = \operatorname{Ker}\partial_m / \operatorname{Im}\partial_{m+1}$. The definition of the singular Betti numbers follows as well.

Let (N, τ_N) be another topological space, and let $\mathcal{C}(M, N)$ be the space of continuous maps between M and N. For $f_0, f_1 \in \mathcal{C}(M, N)$, a *homotopy* between them is a continuous function $F : M \times [0, 1] \to N$ such that $F(x, 0) = f_0(x)$ and $F(x, 1) = f_1(x)$ for all $x \in M$. If such a map exists, then the two functions are said to be *homotopic*. A continuous map $f : M \to N$ is a *homotopy equivalence* if there exists a continuous map $g : N \to M$ such that $f \circ g$ is homotopic to Id_N and $g \circ f$ is homotopic to Id_M. The spaces M and N are said to be *homotopy equivalent* or of the same *homotopy type* if there exists a homotopy equivalence between them, and a space is called *contractible* if it is homotopy equivalent to a point.

If $f \in \mathcal{C}(M, N)$, then f induces a homomorphism $f_* : H_m(M) \to H_m(N)$. If f is a homotopy equivalence, then the following occurs (shown as Corollary 2.11 in [15]).

Theorem 2.1. *The maps $f_* : H_m(M) \to H_m(N)$ induced by a homotopy equivalence $f : M \to N$ are isomorphisms for all m.*

This is a desirable result since rarely in a TODA setting is the support of interest known; rather, what is available is a sample of points on or near it. Therefore, if one uses the sample to construct a space S with the same homotopy type as the unknown support, then one can directly study the unknown support's homology groups through the known, constructed space.

Another helpful fact about simplicial and singular homology is the following (a special case of Theorem 2.27 in [15]).

Theorem 2.2. *If M is a triangulable space, then the mth simplicial and singular homology groups are isomorphic for all m.*

In particular, if M is a simplicial complex, then its simplicial and singular homology groups coincide. Furthemore, if the simplicial and singular homology groups are isomorphic, then their Betti numbers must be the same.

2.2 The Čech and Vietoris-Rips complexes

The results presented thus far presume the space M is fully known, but this is often not the case, particularly in statistical settings. Rather, what is available is a sample S_n, typically a point cloud, on or near M, but this presents an issue: what homological information about M can be gained from S_n?

Theorem 2.1 offers a solution. If one can use S_n to construct a space of the same homotopy type as M, then studying the homology of S directly reveals the homology of M.

As presented in [5], the *nerve* $N(\mathcal{U})$ of an open cover $\mathcal{U} = \{U_\alpha : \alpha \in A\}$ of M is the (abstract) simplicial complex defined by

1. the vertices of $N(\mathcal{U})$ are the U_α, and
2. $[U_{i_0}, \ldots, U_{i_k}] \in N(\mathcal{U})$ if and only if $\bigcap_{j=0}^{k} U_{i_j} \neq \varnothing$

Theorem 2.3 (Nerve Theorem [5])**.** *Let $\mathcal{U} = \{U_\alpha : \alpha \in A\}$ be an open cover of a paracompact topological space M. If any nonempty intersection of finitely many sets in $\mathcal{U}$ is contractible, then M and $N(\mathcal{U})$ are homotopy equivalent. In particular, their homology groups are isomorphic.*

If M is a smooth submanifold of $\mathbb{R}^d$ and $S_n = \{x_1, \ldots, x_n\} \subset M$ is a sample, one can construct an open cover $\{B(x_i, r) : x_i \in S\}$ of M using the open balls $B(x_i, r)$ and a suitable choice of radius $r > 0$, and the nerve of this open cover is called the Čech complex $\check{C}_r(S_n)$ of radius r. A careful choice of r that satisfies the assumptions of the Nerve Theorem would then guarantee the singular homology groups of $\check{C}_r(S_n)$ are isomorphic to the singular homology groups of M: in particular, the Betti numbers would be the same. Moreover, since $\check{C}_r(S_n)$ is a simplicial complex, one need only use the *simplicial* homology groups.

Using Čech complexes involves a number of issues, first among them an appropriate choice of radius r which is discussed more in Section 2.3. Another problem is their computational cost, which can be excessive since one has to consider all intersections of the sets in the open cover.

An alternative simplicial complex which is computationally more efficient is the Vietoris-Rips complex $\mathcal{R}_r(S_n)$, which contains an m-simplex $[x_{i_0}, \ldots, x_{i_m}]$ if $d(x_{i_s}, x_{i_t}) \leq r$ for $0 \leq s, t \leq m$. Rather than computing the intersections of the open cover, one needs only the pairwise distances $d(x_i, x_j)$ for $x_i, x_j \in S_n$.

Note, however, that the Nerve Theorem cannot be applied to the Vietoris-Rips complex since it is not the nerve of the open cover. As such, a natural question is how well does the homology of the Vietoris-Rips complex capture that of M?

Theorem 2.4.

$$\check{C}_r(S) \subset R_r(S) \subset \check{C}_{2r}(S)$$

In a sense, one can view the Vietoris-Rips complex as an *approximation* to the Čech complex, and as such, it is presumed the homology groups of the Vietoris-Rips complex approximate the homology groups of the Čech complex for suitable radii r.

2.3 Persistent homology

As alluded in the previous section, one issue with using the Čech and Vietoris-Rips complexes is finding suitable choices for the radius r, which depends on typically unknown features of the underlying manifold [5]. Rather than select one particular radius, one can take a multi-scale approach using *persistent homology*.

Suppose $S_n = \{x_1, \ldots, x_n\}$ is a sample from a smooth submanifold M of a compact metric space, and consider the *filtered Vietoris-Rips complex* $\mathcal{R}_{r_0}(S_n) \subset \mathcal{R}_{r_1}(S_n) \subset \cdots \subset \mathcal{R}_{r_k}(S_n)$ for $r_0 < r_1 < \cdots < r_k$. The inclusion maps $\mathcal{R}_{r_i}(S_n) \hookrightarrow \mathcal{R}_{r_j}(S_n)$ for $i < j$ induce linear maps $H_k(\mathcal{R}_{r_i}(S_n)) \to H_k(\mathcal{R}_{r_j}(S_n))$ between the kth simplicial homology groups $H_k(\mathcal{R}_{r_i}(S_n))$ and $H_k(\mathcal{R}_{r_j}(S_n))$. The images $\mathrm{Im}(H_k(\mathcal{R}_{r_i}(S_n)) \to H_k(\mathcal{R}_{r_j}(S_n)))$ are called the *kth persistent homology groups*, and $\beta_j^i = \mathrm{rank}(H_k(\mathcal{R}_{r_i}(S_n)) \to H_k(\mathcal{R}_{r_j}(S_n)))$ are called the *persistent Betti numbers*.

Of particular interest are those homology classes that *persist* as one steps through the filtration. When a class appears at filtration value r_b, it is said to be *born* at r_b, and when a class disappears at filtration value $r_d > r_b$, it is said to *die* at r_d. The *persistence* of that class is defined to be $r_d - r_b$. In general, classes that have long persistence are considered "topological signal," and classes that have short persistence are considered "topological noise."

3 Confidence sets for persistence diagrams

One avenue of statistical analysis in TODA is through persistence diagrams. Given a filtration $\mathcal{R}_{r_0}(S_n) \subset \mathcal{R}_{r_1}(S_n) \subset \cdots \subset \mathcal{R}_{r_k}(S_n)$ with persistent homology groups $H_k(\mathcal{R}_{r_i}(S_n)) \to H_k(\mathcal{R}_{r_j}(S_n))$, the *persistence diagram* $\mathcal{P}(S_n)$ is the multi-graph of points $(b_i, d_i) \in \mathbb{R}^2$ that represent the birth-death pairs of the persistent homology classes. For purposes that will be revealed later, the diagonal $\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq y\}$ with infinite multiplicity is included in the persistence diagram

As mentioned previously, homology classes that have longer persistences are often considered "topological signal," while those classes with short persistence are labeled "topological noise." This is equivalent to labeling classes represented in the persistence diagram as "significant" if they lie far enough away vertically from the diagonal and labeling those classes close to the diagonal as "non-significant."

Of course, one may ask what constitutes a "significant" feature, and in [14] Fasy et. al. provide one solution by constructing *confidence sets for persistence diagrams*. That is, supposing $S_n = \{x_1, \ldots, x_n\}$ is sampled independently from a distribution P on $\mathbb{R}^d$ whose support is M, one seeks to find a collection $\mathcal{C}_{n,\alpha}$ of persistence diagrams for $0 < \alpha < 1$ such that

$$\limsup_{n \to \infty} \mathbb{P}(\mathcal{P}(M) \in \mathcal{C}_{n,\alpha}) \geq 1 - \alpha \tag{3.1}$$

One may view the set of persistence diagrams as a metric space if one considers the

bottleneck distance $W_\infty(\mathcal{P}_1, \mathcal{P}_2)$ between persistence diagrams $\mathcal{P}_1$ and $\mathcal{P}_2$ given by

$$W_\infty(\mathcal{P}_1, \mathcal{P}_2) = \inf_\gamma \sup_{x \in \mathcal{P}_1} \|x - \gamma(y)\|_\infty \tag{3.2}$$

where γ is any bijection between $\mathcal{P}_1$ and $\mathcal{P}_2$ (here the diagonal with infinite multiplicity ensures the set of bijections is not empty) and $\|x\|_\infty = \max\{|x|, |y|\}$ for $(x, y) \in \mathbb{R}^2$. One then can then reformulate (3.1) as finding $c_{n,\alpha} \in \mathbb{R}$ such that

$$\limsup_{n \to \infty} \mathbb{P}(W_\infty(\mathcal{P}(S_n), \mathcal{P}(M)) > c_{n,\alpha}) \leq \alpha \tag{3.3}$$

Unfortunately, the bottleneck distance can be difficult to compute, but this difficulty can be overcome due to the fact the bottleneck distance is stable with respect to the Hausdorff distance. The *Hausdorff distance* $\mathcal{H}(A, B)$ between two compact subsets A and B of $\mathbb{R}^d$ is given by

$$\mathcal{H}(A, B) = \inf\{\epsilon > 0 : A \subset B \oplus \epsilon \text{ and } B \subset A \oplus \epsilon\} \tag{3.4}$$

where $X \oplus \epsilon = \bigcup_{x \in X} B(x, \epsilon)$. Furthemore, one can show (3.4) is equivalent to $\|d_A - d_B\|_\infty = \sup_{x \in \mathbb{R}^d} |d_A(x) - d_B(x)|$ with d_A and d_B the distance functions to A and B, respectively.

Equipped with these tools, the stability result [6, 12, 14] is as follows.

Theorem 3.1 (Bottleneck Stability)**.** *Let M be a d-dimensional manifold embedded in a compact subset $\mathbb{X}$ of $\mathbb{R}^d$, and let $S_n \subset M$.*

If $\mathcal{P}(S_n)$ and $\mathcal{P}(M)$ are the persistence diagrams of S and M, then

$$W_\infty(\mathcal{P}(S_n), \mathcal{P}(M)) \leq \|d_{S_n} - d_M\| = \mathcal{H}(S_n, M)$$

In particular, (3.3) is satisfied if $c_{n,\alpha}$ is found such that

$$\limsup_{n \to \infty} \mathbb{P}(\mathcal{H}(S_n, M) > c_{n,\alpha}) \leq \alpha$$

In [14], Fasy et al also provide algorithms, two of which are *subsampling* and *kernel density estimation*, for computing the confidence sets. The subsampling method, which involves taking subsamples $S_{b,n}^j$ of size b from S_n, constructing a reference distribution for $\mathcal{H}(S_{b,n}^j, S_n)$, and determing the $1 - \alpha$ quantile, requires some strong assumptions about the sampling distribution P and no outliers. The kernel density method, which considers instead the persistence diagram of the super-level sets of the kernel-smoothed density, provides a robust alternative. Interested readers are encouraged to read [14] for more details.

4 Two-sample tests using persistence landscapes

An alternative, more statistics-friendly representation of a persistence diagram is the persistence landscape introduced by Bubenik [2]. Given a persistence diagram $\mathcal{P}$ with birth-death pairs $\{(b_i, d_i)\}$, consider the functions

$$f_{(b_i,d_i)}(t) = \begin{cases} t - b_i, & \text{if } b_i \leq t < \frac{b_i + d_i}{2} \\ d_i - t, & \text{if } \frac{b_i + d_i}{2} \leq t < d_i \\ 0, & \text{otherwise.} \end{cases}$$

For $k \geq 1$, the *kth persistence landscape function* of $\mathcal{P}$ is

$$\lambda_k(t) = \text{kmax}_{(b_i,d_i)\in\mathcal{P}} f_{(b_i,d_i)}(t)$$

The *persistence landscape* is the sequence of functions $\Lambda = \{\lambda_1, \lambda_2, \ldots\}$. One can alternatively view the persistence landscape as real-valued function on $\mathbb{N} \times \mathbb{R}$ given by

$$\lambda(k,t) = \lambda_k(t)$$

If $\Lambda_1, \ldots, \Lambda_n$ are the persistence landscapes for persistence diagrams $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n$, then the *sample mean persistence landscape* $\bar{\Lambda}_n(k,t)$ is defined pointwise by

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

Equipped with the p-norm $\|\cdot\|$ given by

$$\|\Lambda\|_p^p = \sum_{k=1}^{\infty} \|\lambda_k\|_p^p,$$

where $\|\lambda_k\|_p^p = \int_{-\infty}^{\infty} |\lambda_k(t)|^p dt$, one can view the persistence landscape of a random sample $S_n \subset M$ as a random object in a separable, Banach space $\mathcal{B}$.

For a random object $X \in \mathcal{B}$, its *Pettis integral* is $E(X) \in \mathcal{B}$ such that for all $f \in \mathcal{B}^*$, the dual space of $\mathcal{B}$, $E(f(X)) = f(E(X))$ with $E(f(X))$ being the expected value of the random variable $f(X)$. The *covariance structure* of X is $E[f(X) - E(f(X))][g(X) - E(g(X))]$ for $f, g \in \mathcal{B}^*$. Moreover, X is *Gaussian* if $f(X)$ is a Gaussian random variable with mean zero for all $f \in \mathcal{B}^*$.

The following Central Limit Theorems are due to Bubenik [2]. Here $L^p(S) = \mathcal{L}^p(S)/\sim$ where $f \sim g$ if $\|f - g\|_p = 0$, and $X_n \to_d X$ signifies a sequence of random objects $\{X_n : n \geq 1\}$ converges in distribution to a random object X.

Theorem 4.1 (Central Limit Theorem for Persistence Landscapes)**.** *Assume $2 \leq p < \infty$ and $\Lambda_1, \ldots, \Lambda_n$ are i.i.d. copies of a random persistence landscape Λ in $L^p(\mathbb{N}\times\mathbb{R})$.*

If $E(\|\Lambda\|) < \infty$ and $E(\|\Lambda\|^2) < \infty$, then

$$\sqrt{n}\big(\bar{\Lambda}_n - E(\Lambda)\big) \to_d V$$

where V is a Gaussian random object with the same covariance structure as Λ. Moreover, if $f \in L^q(\mathbb{N}\times\mathbb{R})$ with $\frac{1}{p} + \frac{1}{q} = 1$ and $Y = \|f\Lambda\|$, then

$$\sqrt{n}(\bar{Y}_n - E(Y)) \to_d \mathcal{N}(0, \text{Var}(Y))$$

Theorem 4.1 provides a key foundation for constructing two-sample hypothesis tests for persistence landscapes. Let $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_{n'}$ be independent samples of i.i.d random point clouds in $\mathbb{R}^d$, and let $\Lambda_1, \ldots, \Lambda_n$ and $\Lambda'_1, \ldots, \Lambda'_{n'}$ be their persistence landscapes, respectively.

For some functional f, let $Y_i = \|f\Lambda_i\|$ with common finite mean $\mu = E(Y)$ and $Y'_i = \|f\Lambda'_i\|$ with common finite mean $\mu' = E(Y')$, and consider the hypothesis $H_0 : \mu = \mu'$. If H_0 is true, then since $\sqrt{n}(\bar{Y}_n - \mu) \to_d \mathcal{N}(0, \text{Var}(Y))$ and $\sqrt{n}(\bar{Y'}_n - \mu) \to_d \mathcal{N}(0, \text{Var}(Y'))$, one has

$$Z = \frac{\bar{Y}_n - \bar{Y'}_{n'}}{\sqrt{S^2_Y/n + S^2_{Y'}/n'}} \to_d \mathcal{N}(0,1) \tag{4.1}$$

where the sample variance $S^2_Y/n + S^2_{Y'}/n'$ can be substituted for $\mathrm{Var}(Y)/n + \mathrm{Var}(Y')/n'$ by Slutsky's Theorem. From this result a p-value can be obtained.

One drawback of the previous test is that it compares homological information in one degree only. While it may be tempting to perform multiple tests for multiple degrees, this becomes a multiple comparisons problem which may increase the family-wise significance level. Alternatively, it would be more desirable to test the persistence landscapes of various degrees simultaneously.

One option is the two large sample Hotelling's T^2 test. For the first group sample $\{X_{1,1}, \ldots, X_{1,n_1}\}$, consider the vector $Y_i = \left(\|\Lambda_i^1\|, \ldots, \|\Lambda_i^p\|\right)$ in which $\Lambda_i^1, \ldots, \Lambda_i^p$ are the persistence landscapes up to degree p. Similarly, one can construct $Y'_1, \ldots, Y'_{n_2}$.

Let $Y = (Y_1, \ldots, Y_{n_1})^T$ and $Y' = (Y'_1, \ldots, Y'_{n_2})^T$. If $n_1 = n_2 = n$, matched pairs data, then let $D = Y - Y'$ and $\mu = E(D)$, and assuming $p \ll n$ consider

$$T_n^2 = n(\bar{D} - \mu)^T S^{-1} (\bar{D} - \mu)$$

where S is the sample covariance matrix of D. Asymptotically, $T_n^2 \to_d \chi_p^2$, which means one can compute a p-value.

If $n_1 \neq n_2$, and assuming $p \ll n - 2$, where $n = n_1 + n_2$ is the total sample size, then consider the sample means $\bar{Y}, \bar{Y}'$ and sample covariance matrices S and S'. The unbiased pooled covariance matrix is

$$S_{\text{pooled}} = \frac{1}{n-2}[(n_1 - 1)S + (n_2 - 1)S']$$

and, under $H_0 : \mu = \mu'$,

$$\frac{n_1 n_2}{n}(\bar{Y} - \bar{Y}')^T S^{-1}_{\text{pooled}} (\bar{Y} - \bar{Y}') \to_d \chi_p^2$$

Thus, one can proceed as normal to compute a p-value.

In each of the prior schemes, one can also use the nonparametric bootstrap procedure if the sample sizes are small. Each iteration, one resamples the i.i.d. samples with replacement and computes both the persistence landscapes and relevant test statistics. Repeating this procedure a large number of times yields a bootstrap distribution for the test statistic from which one can derive a p-value.

5 Example: analyzing breast cancer images with persistence diagrams and landscapes

To show how persistence diagrams and persistence landscapes can be used to summarize data sets and perform hypothesis testing, a sample was taken from the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [11, 17, 16]. The sample contains five left-side, mediolateral oblique (MLO)

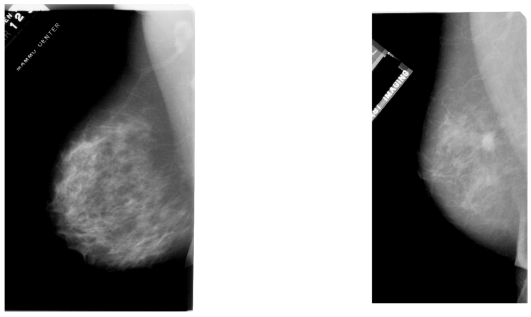

Figure 2: Benign (left) and malignant (right) left-side, MLO mammograms

mammograms that indicate a benign mass and five left-side, MLO mammograms that indicate a malignant mass. See Figure 2 for examples.

Each of the 10 DICOM images were segmented using a local, adaptive thresholding technique in Image Segmenter in MATLAB R2020b with an intensity threshold of 0.5. Example masks are shown in Figure 3.

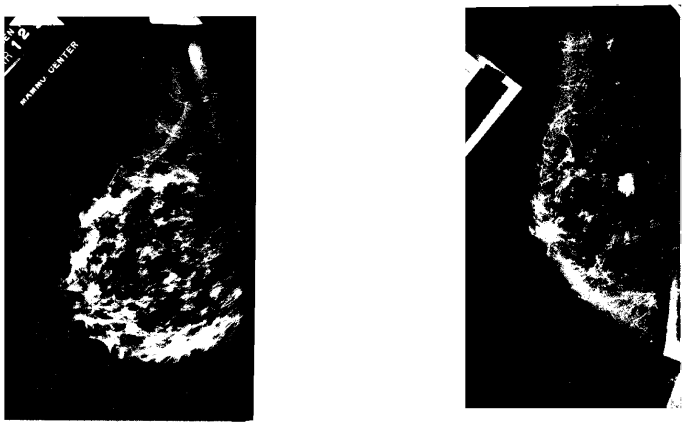

Figure 3: Benign (left) and malignant (right) segmented mammograms

In each of the ten masks, "intrinsic" coordinates [25] were assigned – that is, if a binary image is represented by a $m \times n$ matrix, then a pixel at position (i, j) in the matrix is assigned intrinsic coordinates $(x, y) = (j, i)$.

Using these coordinates, one can view the white pixels as forming a point cloud. After cleaning the images (removing artifacts such as labels and apparent outlying points) and standardizing (centering and scaling), 500 points were randomly selected from each point cloud to serve as the data in the analysis. Figure 4 shows examples of these point clouds.

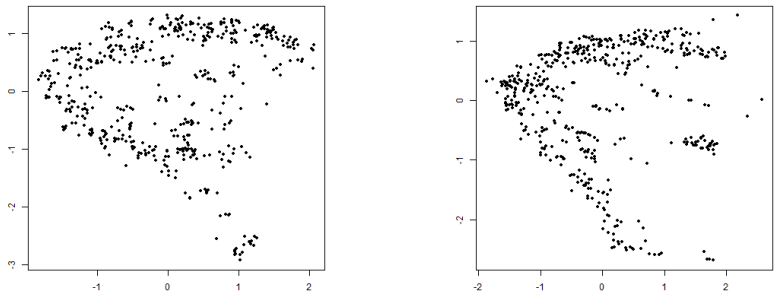

Figure 4: Benign (left) and malignant (right) sampled point clouds

Using the R package `TDA` [13], one can compute both the persistence diagrams and persistence landscapes of these point clouds.

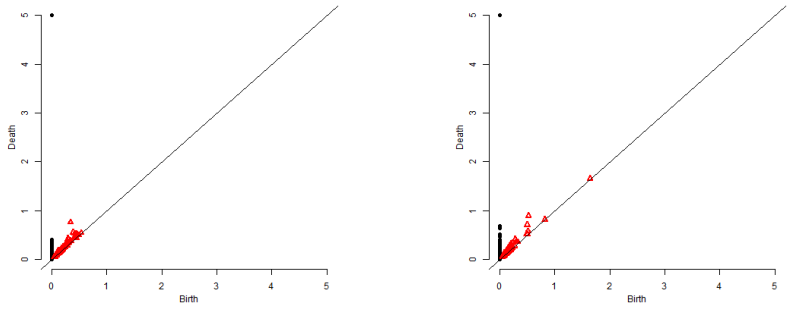

Figure 5: Benign (left) and malignant (right) persistence diagrams

For example, Figure 4 shows the persistence diagrams of the point clouds featured in Figure 4. The persistent homology classes in degree 0 and degree 1 are marked in Figure 5 by black dots and red triangles, respectively. Furthemore, Figure 6 shows the first five persistence landscape functions in degree 1 for the point clouds in Figure 4, and Figure 7 shows the average landscapes and their difference.

To perform the two-sample t-test, the function $f : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ chosen was

$$f(k,t) = \begin{cases} 1, & 1 \leq k \leq 5 \text{ and } t \in [0,2] \\ 0, & \text{otherwise.} \end{cases}$$

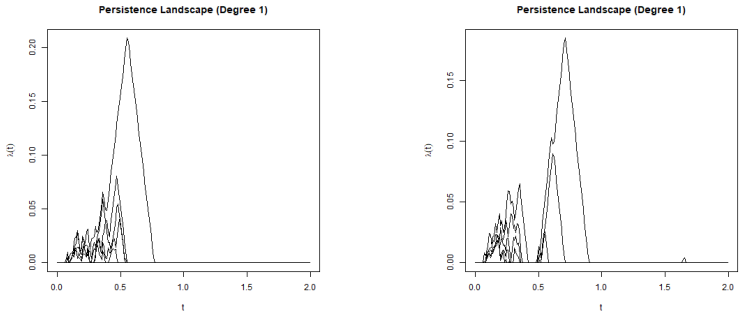

Figure 6: Benign (left) and malignant (right) persistence landscapes

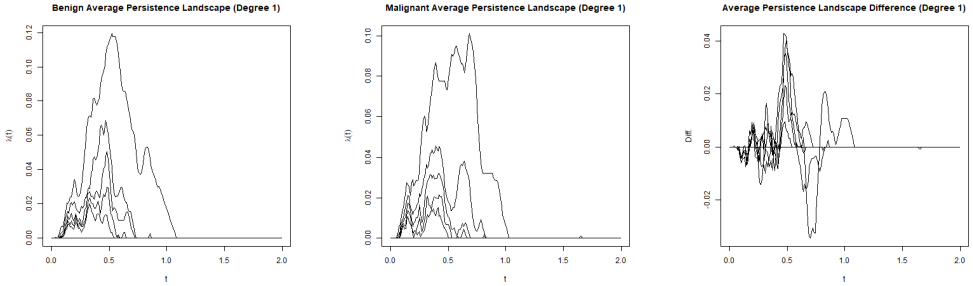

Figure 7: Benign (left) and malignant (right) average landscapes and landscape difference

If one is interested in testing

$$H_0 : \mu_Y = \mu_{Y'} \qquad \text{and} \qquad H_1 : \mu_Y \neq \mu_{Y'},$$

then the test statistic provided by Equation (4.1) is $t = 0.6534$, which yields a p-value of 0.5135.

As such, it would seem from this particular procedure that no difference (at least according to the chosen measurement) has been detected between the benign and malignant mass groups.

For each image's landscape Λ, the norms $\|f\Lambda\|$ are summarized in Table 2.

Table 2: CBIS-DDSM Sample

	Norms					Mean	Std. Dev
Benign (Y)	0.0882	0.0803	0.1331	0.0813	0.0957	0.0959	0.0217
Malignant (Y')	0.0686	0.0730	0.0688	0.1404	0.0729	0.0847	0.0312

There are several limitations to consider with this particular problem, specifically concerning the data's quantity and quality. First, the quantity of data is lacking in perhaps two regards: sample size and sampled points. Five images per group can

certainly be considered a small sample size, which may be affecting the Normal approximation provided by the asymptotic results in Theorem 4.1 above. Furthermore, due to computation limitations, the entire image point clouds could not be used and had to be randomly sampled. As such, extra information about the images could have been excluded due to not being sampled.

As far as the data's quality is concerned, one must remember that 2D images are being used to study 3D phenomena, which means that 2-dimensional homological features cannot be included. Moreover, this example considered only one particular perspective (MLO) in mammography imaging, whereas including other perspectives might have yielded other features.

6 Conclusions

TODA is a collection of tools that allows more options for summarizing and analyzing complex object data, such as the mammogram images feautured above. In summarization, TODA allows one to capture topological features, such as connected components, holes, and voids, which provide further insight into a particular data set and can even aid in distinguishing between two data sets, particularly when traditional summary statistics, such as sample means and sample covariances, appear similar. In analysis, persistence diagrams can be used to discriminate between "significant" and "non-significant" features, and persistence landscapes can be used to perform hypothesis testing through the Central Limit Theorem for Persistence Landscapes and the nonparametric bootstrap.

Other developments in TDA could also be directed toward studying object data. For example, one drawback of using the distance function to construct the Vietoris-Rips complex is that it is highly sensitive to outliers. A robust alternative is the *distiance-to-measure function* [10, 7, 8] whose sublevel sets can still capture features of the underlying support.

Another area of research is finding representations of persistence diagrams that are more amenable to statistics. In this work, one such representation has been found in the persistence landscape, but other vector representations can be found in the *persistence image* [1] and the *persistence codebook* [24]. Finally, instead of topological features, TODA methods can be used to study geometric features [18, 3].

The intersection of Statistics, TDA, and Object Data Analysis found in TODA represents an exciting area of research. For further introduction, some excellent resources are available [4, 19, 22, 9].

References

[1] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, *Persistence Images: A Stable Vector Representation of Persistent Homology*, Journal of Machine Learning Research, 18, 8 (2017), 1-35.

[2] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, arXiv:1207.6437 [cs, math, stat], Jan. 2015, arXiv: 1207.6437.

[3] P. Bubenik, M. Hull, D. Patel, and B. Whittle, *Persistent homology detects curvature*, Inverse Problems, 36(2):025008, Feb. 2020, arXiv: 1905.13196.
[4] G. Carlsson, *Topology and data*, Bulletin of the American Mathematical Society, 46(2):255–308, Jan. 2009.
[5] F. Chazal, *High-Dimensional Topological Data Analysis*, 2016, 22.
[6] F. Chazal, V. de Silva, M. Glisse, and S. Oudot, *The structure and stability of persistence modules*, arXiv:1207.3674 [cs, math], Mar. 2013, arXiv: 1207.3674.
[7] F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman, *Robust topological inference: distance to a measure and kernel distance*, arXiv:1412.7197 [cs, math, stat], Dec. 2014, arXiv: 1412.7197.
[8] F. Chazal, P. Massart, and B. Michel, *Rates of convergence for robust geometric inference*, arXiv:1505.07602 [cs, math, stat], Mar. 2016, arXiv: 1505.07602.
[9] F. Chazal and B. Michel, *An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists*, arXiv:1710.04019 [cs, math, stat], Oct. 2017, arXiv: 1710.04019.
[10] F. d. Chazal, D. Cohen-Steiner, and Q. Marigot, *Geometric inference for probability measures*, Foundations of Computational Mathematics, 6 (2011), 733-751.
[11] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, *The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository*, Journal of Digital Imaging, 26, 6 (2013), 1045-1057.
[12] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of Persistence Diagrams*, Discrete & Computational Geometry, 37, 1 (2007), 103-120.
[13] B. T. Fasy, J. Kim, F. Lecci, C. Maria, D. L. Millman, and V. Rouvreau, *Introduction to the R package TDA*, The Comprehensive R Archive Network (CRAN).
[14] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh, *Confidence sets for persistence diagrams*, Ann. Statist., Publisher: The Institute of Mathematical Statistics, 42, 6 (2014), 2301-2339,
[15] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2002.
[16] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, *A curated mammography data set for use in computer-aided detection and diagnosis research*, Scientific Data, 4, Article number: 170177 (2017); https://doi.org/10.1038/sdata.2017.177
[17] R. S. Lee, F. Gimenez, A. Hoogi and D. Rubin, *Curated Breast Imaging Subset of DDSM [Dataset]*, The Cancer Imaging Archive, 2016.
[18] V. Patrangenaru, P. Bubenik, R. L. Paige, and D. Osborne, *Challenges in Topological Object Data Analysis*, Sankhya A, 81, 1 (2019), 244-271.
[19] V. Patrangenaru and L. Ellingson, *Nonparametric Statistics on Manifolds and Their Applications*, Crc Press, Boca Raton, 2015.
[20] C. Shen and V. Patrangenaru, *Topological Object Data Analysis methods with an application to medical imaging*, International Workshop on Functional and Operatorial Statistics, 237-244.
[21] R. van de Weygaert, G. Vegter, H. Edelsbrunner, B. J. T. Jones, P. Pranav, C. Park, W. A. Hellwing, B. Eldering, N. Kruithof, E. G. P. Bos, J. Hidding, J. Feldbrugge, E. t. Have, M. van Engelen, M. Caroli, and M. Teillaud, *Alpha, Betti and the Megaparsec Universe: on the Topology of the Cosmic Web*, arXiv:1306.3640 [astro-ph], June 2013, arXiv: 1306.3640.

[22] L. Wasserman, *Topological Data Analysis*, arXiv:1609.08227 [stat], Sept. 2016, arXiv: 1609.08227.

[23] X. Xu, J. Cisewski-Kehe, S. B. Green, and D. Nagai, *Finding cosmic voids and filament loops using topological data analysis*, Astronomy and Computing, 27 (2019), 34-52; arXiv: 1811.08450.

[24] B. Zielinski, M. Lipinski, M. Juda, M. Zeppelzauer, and P. Dlotko, *Persistence Codebooks for Topological Data Analysis*, arXiv:1802.04852 [cs, math, stat], June 2019, arXiv: 1802.04852.

[25] ___, Image Coordinate Systems, MathWorks®, https://www.mathworks.com/help/images/image-coordinate-systems.html

Authors' address:

Adam Dixon, Victor Patrangenaru and Chen Shen
Florida State University, U.S.A.
E-mail: ad18g@stat.fsu.edu
vic@stat.fsu.edu
cs15j@my.fsu.edu