On the Statistical Inference of Barcodes *

Leonid Pekelis
Stanford University
Sept 09

Abstract

Statistical methods including (sparse) linear discriminant analysis, and $k$-means clustering are used to extract information from barcodes, the output from persistent homology. Various data is considered including the Mumford natural images data [8] and parametrizations of various shapes in $\mathbb{R}^n$. A particular embedding of the Klein bottle in $\mathbb{R}^8$ is discussed. Area measure is shown to be the correct way to sample from the surface of curved manifolds.

---

*Research supported in part by NSF VIGRE grant.
Adviser: Susan Holmes
Helpful Comments: Gunnar Carlsson, Tigran Ishkanov, Jennifer Kloke

Key words and phrases. Computational Topology, Topology, Klein Bottle, Natural Images, Area Measure, Barcode Statistics, Plex.
1 Algebraic Topology and Persistence

Before delving into numerical analysis, it is instructive to introduce the mathematical techniques employed in persistent homology, and expand on the role each plays in analyzing a data set.

The prime motive of algebraic topology is to study techniques for forming algebraic images of topological spaces. Two topological spaces are considered equivalent if one can continuously deform one space into the other, that is, without ripping or tearing. A classic example of two topologically equivalent spaces is a doughnut and a coffee cup. The reason we are interested in forming algebraic images (groups, rings, modules, etc.) is to quantitatively characterize and compare such groups. If it is possible to construct images such that topologically related spaces have algebraically related images, one can use the algebraic relations to describe the shape of topological space in question.

Still, having a method to classify topological spaces is only one-half of the computational topology puzzle. In general, one can only expect to have a random, and perhaps noisy sample from a given data set. It is necessary to derive methods to extract the most likely topology from a random sample, so that the above techniques may be used. Currently a sufficient method has not yet been identified, although some very good candidates exist. This paper borrows tools from statistical inference in an attempt to improve on one of these methods.

The remainder of this section sets up some terminology, describes one method for forming algebraic images of topological spaces, and then one for inferring topology from a random sample, finishing with a few comments from the author on the strengths as well as most promising areas for improvements for the latter. For an excellent treatment of algebraic topology, see [7]. For thorough and engaging descriptions of persistent homology, the method for inferring topology developed (among others) by H. Edelsbrunner, D. Letscher and A. Zomorodian, and the one described in this paper, see [1] and [2].

1.1 Terminology

We define an n-simplex as the smallest convex set in $\mathbb{R}^m$, $m > n$, containing $n + 1$ points, $v_0, \ldots, v_n$, that do not lie in a hyperplane of dimension less than $n$, or, equivalently with vectors $v_1 - v_0, \ldots, v_n - v_0$ that are all linearly independent. An n-simplex can be thought of as the n-dimensional equivalent of a triangle, and is described as a single point, a line, a triangle, and a tetrahedron in 0, 1, 2, and 3 dimensions. A face of a simplex is the subsimplex with vertices composing any non-empty subset of the original vertices. From n-simplices, one can build a $\Delta$-complex by starting with a disjoint collection of n-simplices, and snapping them together to form sets of faces, or sides of the complex. More formally, if the collection of disjoint simplices of various dimensions is $\Delta_n$, and for each $i$, $\mathcal{F}_i$ is a set of faces of $\Delta_n$’s, the $\Delta$-complex is defined as the quotient space of the disjoint union $\cup \Delta_n$ formed by indentifying the faces in each $\mathcal{F}_i$ into a single simplex by canonical linear homomorphisms between them. In general, each simplex comes with an orientation determined by the order of its vertices, which in turn defines an orientation on a complex, but this technicality will be left out for the sake of brevity. Examples of a torus and a Klein bottle built as a $\Delta$-complexes are given in figure 1. Notice that each is built from two 2-simplices, $U$ and $L$, and the only difference between the objects is in their orientation.

The final piece of terminology from algebraic topology will be a boundary map, $\delta$, which takes a n-simplex to the collection of $(n - 1)$-simplices formed by the sum of faces produced by leaving out one vertex. Such a collection is an example of an $(n - 1)$-chain and denoted by $\sum_i (-1)^i[v_0, \ldots, \hat{v}_i, \ldots, v_n]$, where the carrot notation means leaving out the vertex and $(-1)^i$ is
inserted to preserve orientation. For example, the boundary of a 2-simplex, or triangle, would be the three 1-simplices forming its sides. The boundary of a 1-simplex consists of two points. [insert graphic of boundary map here?]

1.2 Homology

Let $Z$ denote a $\Delta$-complex, $C_n$ be the set of $n$-dimensional complexes formed when all $m$-simplices for $m > n$ of $Z$ are recursively taken to their boundaries, and $\delta_n$ be the boundary map which takes $n$-simplices to $(n-1)$-chains of boundaries. The relation between $C_n$ and $\delta_n$ can be represented as follows,

$$
\cdots \rightarrow C_{n+1} \xrightarrow{\delta_{n+1}} C_n \xrightarrow{\delta_n} C_{n-1} \rightarrow \cdots \rightarrow C_1 \xrightarrow{\delta_1} C_0
$$

where the sequence is extended by 0 on the right hand side, with $\delta_0 = 0$. The $n$-th homology group of $Z$ is defined to be $H_n \doteq \text{Ker} \delta_n/\text{Im} \delta_{n+1}$, and the $n$-th betti number is $\beta_n \doteq \dim(H_n)$. Perhaps the clearest motivation for $H_n$ is that it contains the $n$-dimensional cycles (e.g. loops in 2-dimensions) modulo those belonging to the boundaries of $(n+1)$-simplices. The practical significance is the betti $(n)$ numbers correspond to the number of $n$-dimensional holes in $Z$, with betti 0 representing the number of connected components. Figure 2 gives a few examples of topological spaces and their corresponding betti numbers. Notice that the betti numbers corresponding to dimensions greater than that of the object are 0.

Figure 2: Examples of topological spaces with betti numbers listed. Image borrowed from [10]
1.3 Persistent Homology

Here we describe a technique to build likely topological spaces, and thereby likely homology groups from a finite set of points in Euclidean space sampled from an underlying distribution, called a point cloud. As noted earlier, the major gap between homology and data analysis lies in how to build a likely ∆-complex given a random sample. Once this is accomplished, homology calculations described above take care classifying and categorizing the topological space.

Let \( X \) be a subset of \( \mathbb{R}^n \); the space that our data lives in. And let \( \mathcal{X} \subseteq X \) be the, assumed random, sample of \( X \) that we observe. Throughout this section we will refer to a ∆-complex built from a point cloud as a simplicial complex, but stress that a simplicial complex shares all the properties of a ∆-complex, i.e. it is a collection of ordered vertices belonging to disjoint \( n \)-simplices of various dimensions with rules defining when simplices snap together. Standard examples of simplicial complexes include Cech, Rips and \( \alpha \)-shape complexes. Generally, the vertices are chosen to be \( V = \mathcal{X} \). The problem in practice is that the resulting simplicial complex quickly becomes too unwieldy for computation. For example, it is not uncommon for a Rips complex to have 10 times the number of points in the sample. The witness complex, defined next, instead operates by choosing a small set \( L \subset X \) of landmark points (usually 1/10th to 1/20th of the sample) as the vertex set, and the rest of the unchosen points are “witnesses” in determining the simplexes forming the complex. We define the lazy witness complex, \( W_t(\mathcal{X}, \nu) \), first presented in [2], as follows:

- The vertex set is \( L \).
- For each \( x \in \mathcal{X} \), define \( m(x) \) to be 0 for \( \nu = 0 \), and the distance from \( x \) to the \( v \)th closest landmark, \( l \in L \), for \( v > 0 \).
- The 1-simplex \([ab]\) is then in \( W_t(\mathcal{X}, \nu) \) if there exists a witness \( x \in \mathcal{X} \) such that \( \max\{\|a - x\|, \|b - x\|\} \leq t + m(x) \).
- A higher dimensional simplex is in \( W_t(\mathcal{X}, \nu) \) if and only if all of its 1-faces are.

There does not seem to be much theory motivating this definition, but de Silva and Carlsson (2004) note that for \( \nu = 1 \), \( W_t(\mathcal{X}, \nu) \) can be linked to Vitori-like regions surrounding each point. For the remainder of the paper we take \( \nu = 1 \), and drop both the parameters \( \nu \) and \( \mathcal{X} \), except for where the sample taken is not obvious.

The utility of the parameter \( t \), which is characteristic to every simplicial complex, is that it allows for a multi-resolution approach to determining topology. For example, consider the case of a random sample consisting of 2 clusters of points separated form each other by a small distance, \( \epsilon \). An examination of the definition of \( W_t \) shows that we could either get a complex with \( \beta_0 = 1 \) (one connected component) or 2, depending on our choice of \( t \). Therefore, how small \( \epsilon \) should be before we classify our sample as one connected component also depends on \( t \). To avoid making a blanket, and possibly false, statement for all samples and distributions, \( t \) is allowed to vary.

At the heart of persistence is the property that whenever \( t \leq t' \), there in an inclusion of simplexes in the complex, \( W_t \subseteq W_{t'} \). Therefore, by the property of functoriality, there also exists a linear transformation of homology groups, \( H_n(W_t) \rightarrow H_n(W_{t'}) \), which can be classified according to a barcode of their betti numbers.

A barcode is a finite set of intervals in \( \mathbb{R}^+ \), with each interval bounded below. It is a theorem that barcodes are in bijective correspondence with isomorphism classes of directed systems of vector
spaces [13], which is what we have by the inclusion property described above. There exists a separate barcode for each homology group of dimension $n \geq 0$ on a random sample $X$. Each interval $[t_0, t_1)$ in a barcode corresponds to an $n$-dimensional hole opening in the complex at time $t_0$ and being filled in at time $t_1$. An interval with $t_1 = \infty$ corresponds to a perpetually open hole. Imagine computing the betti numbers for a finite set of indices $0 = t_0 < t_1 < \cdots < t_n$, and describing the $n$-cycle that is present at both $t_i$ and $t_j$, $i < j$, but not at any $t < t_i$ or $t_j$, by the interval $[t_i, t_{j+1})$. Hence, at any $t$ one may count the numbers of open intervals in a barcode for dimension $n$, obtaining $\beta^t_n = \dim(H_n(W_t))$. An example barcode is given in figure 3.

A natural interpretation of barcodes is that longer intervals correspond to intrinsic properties of the underlying space, while short intervals reflect “topological noise” in the form of artifacts from the construction of simplices, or perhaps sampling noise.


2 Sampling from a Curved Manifold

Imagine forming a torus according to the schematic in figure x. The steps are first roll a square sheet of paper into a tube by identifying 2 sides, and then attach both ends of the tube to each other. In order for the outer ends of the tube to meet, the paper will have to stretch along the outer boundary, or, conversely, compress to form the smaller, inner hole. It is plausible that this stretching and compressing of space would influence a distribution on the surface of a torus. This influence can be described by the area measure, which, roughly, is formed by taking a region of the surface of a torus. This outer boundary, or, conversely, compress to form the smaller, inner hole. It is plausible that this stretching and compressing of space would influence a distribution on the surface of a torus. In order for the outer ends of the tube to meet, the paper will have to stretch along the outer boundary, or, conversely, compress to form the smaller, inner hole. It is plausible that this stretching and compressing of space would influence a distribution on the surface of a torus.

Theorem 2.1. For $m \leq n$ and $g : \mathbb{R}^m \to \mathbb{R}$ a Borel function,

$$
\int_A g(f(x)) J_m f(x) \lambda^m(dx) = \int_{\mathbb{R}^n} g(y) N(f|A, y) \mathcal{H}^m(dy)
$$

(1)

Proof. See [5], (3.2.5)

A similar result holds for $m > n$ using the coarea of $f|A$ and is given in (3.2.12) of [5].

Equation (1) is a useful formula because it allows one to draw a sample with density $g$ on a curved manifold by sampling from the density $g \circ J_m f(x)$, after normalization, in the parameter space $\mathbb{R}^m$. In the cases considered in this paper, we take $g = \frac{1}{B}$, the uniform density with $B \subset \mathbb{R}^m$, the domain of $f$.

The normalized area measure for 3 simple shapes is given in table []. A fourth example, the Klein bottle embedded in $\mathbb{R}^3$ (see section 3) is also presented, but without a normalizing constant. As one may see, the difficulty of area measure calculations grows quickly with increasing complexity of the parametrization, forcing the use of numerical methods for all but the simplest manifolds. Plots of the area measure for the shapes considered in section [] are given in figure [], with visual
representations of the shapes colored according to area measure in figure[]. We exclude the circle from the images as its area measure is trivial. Figure[] shows that the earlier statement that area measure is proportional to local curvature is supported.
3 Parametrizing a Klein bottle in $\mathbb{R}^8$

Need linking sentence here.

The key feature of this parametrization is that is developed with the topology of the Klein bottle in mind. Other commonly found parametrizations attempting to trace out an object in $\mathbb{R}^3$ often employ shortcuts that although give a faithful visual representation, should not be expected to hold up under tests of persistent homology. It is hoped that this distinction will not only give a more pronounced barcode signature, but also more accurately represent a Klein bottle structure one may see in actual data. This section will describe the parametrization and discuss the associated area measure. Much of the construction is taken from [4], where a similarly formed manifold was actually shown in a database of natural images.

Let $K$ be the space of all degree 2 polynomials of the form $c(ax + by)^2 + d(ax + by)$ where $(a,b) \in S^1$ and $(c,d) \in S^1$. Each pair of coordinates is chosen on the unit circle. Defining that map $g : S^1 \times S^1 \to K$ by $g(a,b,c,d) = c(ax + by)^2 + d(ax + by)$. As shown next, $g$ maps the unit torus to a space homeomorphic to the Klein bottle.

First note that $g$ is onto, but not $1 - 1$, as the points $(a, b, c, d)$ and $(-a, -b, c, -d)$ give the same polynomial. Parametrizing $(a, b)$ and $(c, d)$ by $\theta$ and $\psi$, the non-uniqueness property can be represented identically by two equivalence relations

$$(a, b, c, d) \sim (-a, -b, c, -d) \text{ for } (a, b) \in S^1 \text{ and } (c, d) \in S^1 \text{ and } (\theta, \psi) \sim (\theta + \pi, 2\pi - \psi) \text{ for } \theta, \psi \in [0, 2\pi].$$

This periodicity gives the required reverse orientation of one of the edges for a Klein bottle when constructing it from a square. Figure 4 shows this by giving the image of $g$ on the square representation of a torus. Equivalent regions under $g$ are shaded in like colors, and arrows indicate the direction of equivalence. Both the left and right halves of $im(g)$ show the standard representation of a Klein bottle. Hence we can consider the space $K$ to be homeomorphic to a Klein bottle.

We now embed $K$ in $\mathbb{R}^8$ by a map $q : K \to \mathbb{R}^8$. Note that in general we cannot expect $im(q)$ to also be homeomorphic to a Klein bottle, but a result from general topology shows that a continuous $1 - 1$ map on a compact space is a homeomorphism onto its image. Define $q$ by evaluating the polynomial in $K$ at nine points on a grid obtained from all pairwise combinations of $x, y \in \{-1, 0, 1\}$. Although $q$ gives a 9-vector for each polynomial in $K$, $q(p)$ actually sits in $\mathbb{R}^8$ since for all $p$ in $K$, $p(0, 0) = 0$. Hence $q$ can be equally formulated as evaluating $p$ on 8 grid points excluding the origin, keeping in mind that a ninth constraint on $p$ is decided by the space $K$. Since a second degree polynomial of one variable can be uniquely determined from 3 points, we expect 9 to be sufficient for two variables, and hence for $q$ to be $1 - 1$. The following proposition proves this explicitly. Note that the proposition also shows that $im(q)$ is homeomorphic to a Klein bottle since the space $K$ is parametrized by a closed subset of $\mathbb{R}^2$, and therefore compact.

**Proposition 3.1.** $q$ restricted to $K$ is one-to-one.

**Proof.** Suppose there exist $p, p_2$ in $K$ such that $q(p) = q(p_2)$. Evaluating the two polynomials at $(-1,0), (1,0), (0,-1)$ and $(0,1)$ gives four equations relating the coefficients $(a,b,c,d)$ and $(a_2, b_2, c_2, d_2)$, adding the first and second two gives the following two equations $da = d_2a_2$ and $db = d_2b_2$.

There are two possible cases. Case (1) $\frac{a}{a_2} = \frac{b}{b_2}$. Case (2) $d = d_2 = 0$. 

8
First addressing case (1), if \((a, b) = (a_2, b_2)\) or \((a, b) = -(a_2, b_2)\), evaluating the polynomials at \((1, 1)\) and dividing gives \(c = c_2\). In the first instance, all 4 sets of coefficients are equal. In the second, \((a_2, b_2, c_2, d_2) = (-a, -b, c, -d)\), which is exactly the equivalence relation (2). Hence \(p = p_2\).

In case (2), there are two distinct possibilities, (a) \(c = c_2 = -1\), (b) \(c = c_2 = 1\). For (a), evaluating \(p\) and \(p_2\) at \((1, 1)\) gives \((a + b)^2 = -(a_2 + b_2)^2\), a contradiction. For (b), evaluating the polynomials at the \((1, 1)\) and \((-1, 1)\) gives the equations \(a^2 + 2ab + b^2 = a_2^2 + 2a_2b_2 + b_2^2\) and \(a^2 - 2ab + b^2 = a_2^2 - 2a_2b_2 + b_2^2\). Subtracting the equations gives \(ab = a_2b_2\), which, since \(c = c_2\), implies \((a, b) = \pm(a, b)\), and that \(p = p_2\) by case (1).

Letting \(S^1_r\) denote a circle with radius \(r\), and defining \(h(\theta, \psi; r, R) = (r \cos(\theta), r \sin(\theta), R \cos(\psi), R \sin(\psi))\), the above parametrization can be generalized and fully specified by \(q \circ g \circ h(\cdot, \cdot; r, R) : [0, 2\pi] \times [0, 2\pi] \to IR^8\) for any choice of parameters \(r, R > 0\). Note that proposition 3.1 and all the above calculations carry over exactly if we instead assume the coefficients come from circles of non-unit radius. The parametric equations and corresponding (non-normalized) area measure are given in table [], while a plot of the area measure is shown in figure []. Since the function \(g\) produces 2 identical copies of the Klein bottle from a torus, only one half of the parameter space is necessary. Hence the parameter space for the Klein embedding in \(R^8\) will be \([0, \pi] \times [0, 2\pi]\).

Figure 4: The image of the map \(g\).
4 Statistical Inference & Numerics

To compute persistent homology of point cloud data, we use the freely distributed software package PLEX, developed by G. Carlsson, A. Zamordian, V. De Silva and P. Petty [9]. In addition, statistical analysis is done in R, utilizing the additional packages sparseLDA, mda, mcmc, lattice, and scatterplot3d all available from CRAN, http://cran.r-project.org/. All scripts, both to produce the shapes data, and to perform analysis are publicly available at http://www.stanford.edu/~lpekelis/comptop. For the actual data sets generated in the analysis, please contact the author.

4.1 LDA on barcodes - natural images database

The first set of numerical analysis is carried out on a database of '5 by 5' patches of high contrast, denoted $\mathcal{M}$, taken from a collection of photographs gathered by H. van Hateren and A. van der Schaff. The collection consists of 4,167 digital images of outdoor scenery and is described in [12]. The steps to form a set of high contrast patches from an image involve mean and variance normalization, before selecting the set of top 20% highest contrast patches. The procedure is described in detail in [8]. The dataset $\mathcal{M}$ contains 15,000 such high contrast patches. These patches were obtained by first selecting a random subset of the full space of patches, and then applying a $k$-nearest neighbors density threshold. Each patch is saved as a 25-vector sitting in a 23-dimensional space due to normalizations.

For a similar space of '3 by 3' patches, it was shown that under a suitable density threshold, the topology of the dataset can be represented by a 3 ring diagram [4]. In a later work Carlsson and Ishkarov find that the space of '5 by 5' patches also supports this claim [3]. Each ring in the diagram represents an ordering of high and low contrast bars, or gradients. The gradients are rotated around the grid center as one moves around the ring. Figure [] is borrowed from Carlsson et. al. (2008) to show the ring representation. This diagram can be parametrized by the Klein bottle (compare figure [] to the Klein bottle schematic in figure []). Persistent homology calculations also support this claim.

Having strong suggestion that $\mathcal{M}$ has a non-trivial and fairly complex topological structure affords an opportunity to use it to search for methods that give quantifiable differences between barcodes emerging from spaces with different topological structures. To this end, we create two new datasets to append to $\mathcal{M}$.

The first, denoted $\mathcal{R}$, can be considered a collection of random high contrast patches obtained by choosing a large set of uniformly random '3 by 3' intensity patches, performing the same normalizations and keeping only the top 20% highest-contrast patches as in [8]. As in [4], we then take the 30% densest points using a 300th nearest neighbor density estimate.

The second, $\mathcal{M}_{10}$, is a multinomial sample of 15,000 from $\mathcal{M}$, with probability of observation $m$ being chosen $1/15,000$ for each $m$ in $\mathcal{M}$ (sampling with replacement). To this sample we add noise distributed $\mathcal{N}(0, \hat{\sigma}^2)$ where $\hat{\sigma}^2$ is the empirical variance, column-wise, of $\mathcal{M}$.

From each data set, we produced multiple random samples - 60 samples from $\mathcal{R}$, and 400 from each $\mathcal{M}_{10}$ and $\mathcal{M}$. The sample size was chosen according to computational difficulty of data generation. Next, persistent homology calculations through PLEX were performed, giving a final count of 860 barcode sets, each set containing 3 barcodes corresponding to betti 0,1, and 2 intervals.

With this may barcodes it is difficult to compare each barcode individually, and so a representation which we call a 'barseries' was developed. The barseries gives a piece-wise linear curve with
Figure 5: Barseries for $\mathcal{R}$

Figure 6: Barseries for $\mathcal{M}$

Figure 7: Barseries for $\mathcal{M}_{10}$
height equal to the number of intervals a vertical bar would intersect in a barcode if the bar moved from $t_0 = 0$ to some large $t_n$, pausing to take measurements at a uniformly spaced set of times $t_0 < t_1 < \cdots < t_{n-1} < t_n$. This gives the number of betti $n$ roles present at times $t_i$, modulo the interval each hole represents. An example comparing a barcode and a barseries is shown in figure \[\]. A local polynomial regression is shown in red to approximate the general shape of the barseries. The advantage of removing the identity of individual intervals is that it enables the comparison of barcodes from many samples at once, forming an empirical distribution of the number of intervals in a barcode as $t$ increases. Figures 5, 6, and 7 show betti 1, 2 and 3 barseries for all 3 data sets. Each figure has has series for all $n_j$ samples, $j \in \{R, M, M_{10}\}$, overplotted, with 90% confidence regions marked.

It is clear that the barseries for randomly generated high-contrast pixels are different from both other datasets. The final sampling time for the series ($t_n$) is determined according to the endpoint of the longest, finite interval among all barcodes in a sample. Topologically, given a sample $\mathcal{X}$, for all $t \geq t_n(\mathcal{X})$, $W_t(\mathcal{X})$ is homeomorphic to a single point. The longest finite interval for $\mathcal{R}$ among all samples is only about 1/20th the same measurement for $\mathcal{M}$ or $\mathcal{M}_{10}$. In contrast $\mathcal{M}_1$ and $\mathcal{M}_{10}$ have approximately similar values for $t_n$. This result can be expected, as we do not expect to see evidence of topological structure as output from persistence, when none was present in the input.

![Figure 8: Plots of pixel square data projected onto first two discriminant directions.](a) lda1

(b) lda2

Figure 8: Plots of pixel square data projected onto first two discriminant directions. lda1 - does not include data value $r$, lda2 - includes $R$. Orange - $R$, Green - $M$, Blue - $M_{10}$

It is not as straightforward to tell whether introducing noise to $\mathcal{M}$ produces different barcodes. To test this hypothesis, we use linear discriminant analysis. Linear discriminant analysis picks an optimal classification of data into groups, under the assumptions that the number of groups is known, each group is distributed normally about a centroid, and all groups have equal variance. A benefit of lda is that a projection of the data onto its discriminant directions gives a view of the
data that maximizes between group variance, while keeping within group variance small. For the barcode data we select a training sample of size approximately 75% (2133 observations) and leave the rest for a test set. Figure 8 shows a plot of the test data projected onto the 1st two discriminant directions, color coded according to the true group labels, for two separate lda tests. The centroids are labeled by dataset name. Table 1 gives the corresponding confusion matrices.

Both lda tests had training error close to 1% and 10% test error. The highest number of misclassifications occured between $M$ and $M_{10}$ each time. With fairly low test error, it seems lda is fairly effective at classifying barseries, even for datasets of similar topological structure but varying levels of noise.

The distinction between the two lda tests was that the second included an additional piece of data, defined as $R = \max_{z \in Z} \{d(z, L)\}$, where $Z = X \setminus L$, the points left over after choosing landmarks. Intuitively $R$ represents how finely a set of landmark points covers a data set. It seems as though this value is influences by the presence of topological structure in the setting of pixel patches of natural images. Including the $R$ value allowed lda to separate out the random dataset very cleanly, reducing test error pertaining to the group $R$ to 0.

### 4.2 Classifying random samples from curved manifolds

The dataset for the second set of analysis consists of 300 random samples drawn from the surface of 5 shapes (1500 samples in total), along with 300 more samples drawn within the unit cube in $\mathbb{R}^3$. Hence we have 6 different sources of data:

1. Circle
2. 3-Sphere
3. Torus
4. Klein Bagel (simplified representation of Klein bottle)
5. Klein Bottle immersed in $\mathbb{R}^3$

Table 1: Confusion Matrices for 2 Linear Discriminant Models. lda1 - does not include data value $r$, lda2 - includes $R$. The percent error for each model is: lda1 - 0.02%, 0.07%, lda2 - 0.01%, 0.05%.
6. Within the unit cube in $\mathbb{R}^3$ (topology of single point)

Each of the 6 sources above were randomly rotated and embedded in $\mathbb{R}^{40}$. The data itself is obtained by sampling 5000 points uniformly from the surface of shapes 1-5 and from the interior of the cube. The cube can be regarded as a control since it exhibits the degenerate topology of a single connected component with no $n$-dimensional holes for $n > 1$. The corresponding betti sequence is $(1, 0, \ldots)$. For reasons expanded upon later, this should help to distinguish “noise” common to all inputs from information describing unique topological structure.

Since our shapes exhibit curvature, we use area measure calculations in order to correctly sample uniformly. Heuristically, this amounts to increasing the probability density in regions that are highly curved, as points arising uniformly on the surface would tend to “bunch up” there. A discussion of area measure can be found in section 2.

After sampling 5000 points from each surface 300 times, the persistent homology software PLEX is run to obtain barcodes. Figure 9 gives the intervals for each shape and all samples, plotted according to the start and end time of the intervals. The plots are colored by betti number.

Even without clustering analysis, we can infer some general properties for intervals produced by PLEX. First, shapes with natural embeddings in three or more dimensions seem to all have a similar distribution of intervals, which is different from our one example of a two dimensional shape. This should be expected as shapes with embeddings in two dimensions can be completely characterized by 2-simplexes. Second, the intervals showing the greatest amount of variation in cluster shape from shape to shape are the ones that start relatively early and end relatively late. This would support general assumption by Carlsson (see, for instance, [1]) that longer intervals translate to a feature correctly characterizing the topology intrinsic to a dataset. Third, the start time of an interval appears related to the betti number, with higher betti number intervals starting later. Again, this is intuitive since more higher-dimensional simplicies are needed to form the higher dimensional holes that characterize larger betti numbers. Fourth, the “noise” given by the algorithm seems to consist of short intervals, which can occur at any time. The author’s hypothesis is that the bottom region in all plots amounts to “noise,” or intervals that one would expect even without any topological structure. The cube plot (a) supports this by having no other intervals. What is impressive, is the almost a linear relationship for the noise intervals, suggesting that their length stays relatively constant and independent of start time.

More analysis to follow in this section as the author has time.
References


