Hypothesis Testing for Topological Data Analysis

Andrew Robinson & Katharine Turner

October 29, 2013

Abstract

Persistence homology is a vital tool for topological data analysis. Previous work has developed some statistical estimators for characteristics of collections of persistence diagrams. However, tools that provide statistical inference for scenarios in which the observations are persistence diagrams are not developed. We propose the use of randomization-style null hypothesis significance tests (NHST) for these situations. We demonstrate this method to analyze a range of simulated and experimental data.

1 Introduction

Topological Data Analysis (TDA) is about considering shape within data. For example, samples may lie on a submanifold and we may want to learn about this manifold or we may want to understand the higher dimensional correlations between different variables. The main tool in TDA is persistence homology, which summarizes how the topology changes through a filtration of a space. Although it is attracting more and more interest, the use of persistent homology in data analysis remains widely heuristic. There are relatively few papers that establish connections between persistence and statistics and, despite a few promising results, the statistical analysis of homology, persistent homology and more general topological and geometric features of data, is still in its infancy.

The need to introduce statistical techniques to topological data analysis has become increasingly apparent. The overarching idea is that the observations used as input to most analysis techniques are generated stochastically, often using a sampling mechanism from a process or population. Therefore, it is useful to talk not only about a single persistence diagram, but about an entire collection of them, i.e. a sample, drawn from some distribution. Some progress has been made in, for example, calculating means, variances and applying statistical inference techniques (Mileyko et al., 2011; Turner et al., 2012; Balakrishnan et al., 2013; Turner, 2013; Chazal et al., 2013; Bubenik and Kim, 2007). Alternative approaches involve doing analysis on related objects to persistence diagrams. In particular there has been progress doing analysis using an object called a persistence landscapes including randomization tests (Bubenik, 2012).

Null hypothesis significance testing (NHST) is a commonly used and important statistical tool that provides a measure of the strength of evidence against a hypothesis. In the current setting, NHST will quantify the the differences between two different types of underly objects or processes, using persistence diagrams as observations. For example it can provide a necessary condition as to whether particular persistence diagrams could be used for classification.

Unfortunately, the space of persistence diagrams is geometrically very complicated. It is infinite in dimension and arbitrarily curved (Turner et al., 2012; Turner, 2013). As a result it is not plausible to use any parametric models for distributions, so we cannot do NHST using a method that requires an underlying parametric model. Our approach is to instead find a relevant joint loss function and then use a randomization test (also known as a permutation test). This method of NHST is standard in statistical theory and is theoretically rigorous (see, e.g., Casella and Berger, 1990; Welsh, 1996).

The theory behind the randomization test ensures that when the two sets of diagrams are drawn from the same distribution of diagrams, then the p value obtained is a random variable with a uniform distribution over an evenly spaced subset of $[0, 1]$. However we do not know of any theory that the p value will necessarily be low if the distributions are different. Furthermore, since persistence diagrams are summary statistics it is possible that the distribution of the underlying objects under analysis might be different but the corresponding distributions of persistence diagrams are similar. We will therefore show by example that there do exist situations where the null hypothesis might be correctly rejected by our method.

In the following section we provide a brief overview of the background theory for TDA and the rationale behind NHST. We then develop a test procedure in Section 3. In Sections 4 and 5 we apply the resulting algorithm to a range of data including point clouds of shapes and the persistent homology transform of silhouette data, and the concurrence filtration for fMRI data, respectively.

2 Preliminaries

2.1 TDA Background Theory

Persistence diagrams provide a means of capturing and recording a summary, given a filtration, of how the topology changes through that filtrations. The set up is that we are given a filtration $K = \{K_r | r \in \mathbb{R}\}$ of a countable simplicial complex indexed over the real numbers with $K_{-\infty} = \emptyset$. We wish to summarize the change in the topology of the filtration over time.

For $i < j$, the inclusion map

$$\iota^{\{i \to j\}} : K_i \to K_j$$

induces homomorphisms

$$\iota_*^{\{i \to j\}} : H_*(K_i) \to H_*(K_j).$$

We say that a homology class $\alpha \in H_*(K_i)$ is *born* at time i (denoted $\mathrm{b}(\alpha)$) if it is not in the image $\iota_*^{\{i' \to i\}}$ for any $i' < i$. We say that α *dies* at time j (denoted $\mathrm{d}(\alpha)$) if $\iota_*^{\{i \to j\}}(\alpha) = 0$ but $\iota_*^{\{i \to j'\}}(\alpha) \neq 0$ for $i < j' < j$. We say that α is an *essential class* of K if it never dies.

For each pair (i, j) with $i < j$ we can then consider the vector space of homology classes that are born at time i and die at time j. Let $\beta_*^{(i,j)}$ denote the dimension of this space. Similarly, let $\beta_*^{(i,\infty)}$ denote the dimension of the space of essential homology classes that are born at time i.

Let $\mathbb{R}^{2+} := \{(i, j) \in (-\infty \cup \mathbb{R}) \times (\mathbb{R} \cup \infty) : i < j\}$ We define the k-th persistence diagram corresponding to the filtration K to be the multi-set of points in $\mathbb{R}^{2+}$ alongside countably infinite copies of the diagonal such that the number of points (counting multiplicity) in $[i,\infty)\times[j,\infty]$ is equal to dimension of the image of $\iota_k^{\{i\to j\}}$. That is the dimension of the space of k-dimensional homology classes that are born at or before i and die at or after j. This is achieved by placing at each (i,j) a number of points equal to $\beta_k^{(i,j)}$. The countably many copies of the diagonal play the role of homology classes whose persistence is zero and hence would not otherwise be seen.

Let $\mathcal{D}$ denote the space of persistence diagrams. There are many choices of metrics in $\mathcal{D}$, analogous to the variety of metrics on spaces of functions. We will be considering the distance metric that is analogous to the L2 distance in the space of functions on a discrete space and the 2-Wasserstein distance between probability distributions. A natural family of metrics is discussed in Turner (2013).

Let X and Y be diagrams. We can consider bijections ϕ between the points in X and the points in Y. These are the transport plans that we consider. Bijections always exist because there are countably many points at every location on the diagonal. We only need to consider bijections where off-diagonal points are either paired with off-diagonal points or with the point on the diagonal that is closest to it.

Define

$$d(X,Y)=\left(\inf_{\phi:X\to Y}\sum_{x\in X}\|x-\phi(x)\|_2^2\right)^{1/2}. \tag{1}$$

We will call a bijection between points *optimal* if it achieves the infimum. We can find an optimal bijection, given two diagrams X and Y with only finitely many off diagonal points, using the Hungarian algorithm (also known as Munkres assignment algorithm). Suppose X has n off-diagonal points, labelled $x_1,x_2,\ldots x_n$, and Y has m off-diagonal points, labelled $y_1,y_2,\ldots y_m$. Let $x_{n+1},x_{n+2},\ldots x_{n+m}$ and $y_{m+1},y_{m+2},\ldots y_{n+m}$ be copies of the diagonal. We construct a cost matrix with $n+m$ column and rows where the (i,j) entry is $\|x_i-y_j\|_2^2$. When either x_i or y_j is a copy of a diagonal then this is the perpendicular distance. Each transportation plan corresponds to an assignment of rows to columns — a bijection between the points in X and those in Y.

Given two sets X and Y, and pairwise costs associated to assigning to $x\in X$ the object $y\in Y$, the Hungarian algorithm finds the least-cost bijective assignment. Suppose we have two diagrams X and Y each with only finitely many off-diagonal points. Consider as many copies of the diagonal in X and Y to allow the option of matching every off-diagonal point with the diagonal. The cost of $x\in X$ doing task $y\in Y$ is $\|x-y\|_2^2$. The total cost of an assignment (or in other words bijection) ϕ is $\sum_{x\in X}\|x-\phi(x)\|_2^2$. The Hungarian algorithm gives us a bijection ϕ that minimizes this cost. This means it gives an optimal bijection between X and Y.

There are many ways to create filtrations of interest. One of the most common ways is the forming of Rips complexes from point cloud data. We will use this method for our simulated examples later. Given a point cloud $\{x_1,x_2,\ldots x_N\}$ of points in Euclidean space $\mathbb{R}^n$ we define the *Rips complex* with parameter ϵ (denoted $\mathcal{R}(\epsilon)$) to be the flag complex on the graph whose vertices are $\{x_1,x_2,\ldots x_N\}$ and contains the edge (x_i,x_j) when $\|x_i-x_j\|\leq\epsilon$. We then build a filtration by considering the Rips complexes under an increasing parameter.

2.2 Persistence Diagrams as random elements

If our method of constructing a filtration is in some way random[1] then this randomness can also be seen in the corresponding collection of diagrams. A distribution of filtrations determines a distribution of diagrams. As a result we have a persistence diagram valued random element. This process works for any method of creating a filtration whether it is sublevel sets of a function or the Cech or Rips complexes from a point cloud.

For example, suppose we are sampling points m from a subset K of $\mathbb{R}^d$ with some noise. We are stochastically generating a point cloud that will approximate K. This sample generates a distribution $\rho_{\text{point clouds}}$ of sets of m points in $\mathbb{R}^d$. Each point cloud determines a filtration of simplicial complexes and hence a distribution $\rho_{\text{filtrations}}$ of filtrations of simplicial complexes. In turn each of the filtrations determines a distribution ρ_K of persistence diagrams. Every time we draw m sample points to create a point cloud we are effectively drawing a sample from the distribution $\rho_{\text{point clouds}}$ and hence also drawing a sample persistence diagram from ρ_K. Under certain conditions, for example that the sample is random, we can learn something about K by analyzing ρ_K. Suppose we have another subset L of $\mathbb{R}^d$ which we can similarly sample to form point clouds. We may wish to know if K and L are different. Our null hypothesis would be that they are the same subset. A necessary, but not sufficient, criterion for K to be L is that $\rho_K = \rho_L$. This implies that our null hypothesis for studying persistence diagrams is that the underlying distributions from which ρ_K and ρ_L are drawn are the same. We later consider a simulated examples of this form.

2.3 Null Hypothesis Significance Testing

We now review the algorithm and rationale behind null hypothesis significance tests. Further reading can be found in many introductory and medium-level statistical texts; we mention for example, Casella and Berger (1990), Welsh (1996) and Pawitan (2001). The steps for the test are as follows. First, choose a parameter that represents the data in some way, and about which a pertinent hypothesis can be formed. Popular examples of parameters include the sample mean and median for tests of location, and the variance for tests of spread. Second, choose a statistic to use to estimate the parameter. Third, predict the statistical behavior of the statistic under the null hypothesis, trying to capture the full range of variability that is implied by the model. Commonly, statistical theory is used to nominate a distribution for the test statistic assuming that the null hypothesis is true. For example, the sample mean might be assumed to have a Gaussian distribution, based on the Central Limit Theorem or the assumed distribution of the data. Fourth, compare the observed value of the test statistic with the expected behavior under the null hypothesis. If the test statistic is anomalous compared with the expected behavior under the null hypothesis, then it is considered to be evidence against the null hypothesis.

Formal approaches to testing diverge at the point of comparison, and we discuss two of them here. One approach requires nominating a cutoff, called the *size* of the test, which is by definition the probability of mistakenly rejecting the null hypothesis. That is, the cutoff is set to be the probability of rejecting the null hypothesis when it is true. Then, the probability of observing a result as or more extreme than the observed test statistic is computed based on repeated experiments, assuming that the null hypothesis is true. That is, we imagine a set of identical experiments to be carried out, for which the null

[1] Including, for example, by the random selection of units from a population or process.

hypothesis is true, and ask what is the proportion of that set for which the computed test statistic is more extreme than the observed value, relative to the null hypothesis. Then report the outcome of the comparison of the observed probability against the size of the test. This is a Neyman–Pearson approach to testing, and in the normal case where the variance is unknown and the hypothesis concerns the mean, the reference distribution is Student's t distribution, with degrees of freedom equal to the sample size n minus one. Another approach, following Fisher, simply reports the estimated probability computed above, called the p value. The reader is free to place their own interpretation on the p value. We will follow the latter approach.

In any case, interpretation of the outcome of the usual NHST is conditional on some model, and the hypothesis is stated in terms of parameters of the model. It is due diligence for the analyst to ensure that the model is a defensible approximation to reality. This is usually performed by examining graphical diagnostics of some quantities that arise from the model estimation. For example, the analyst might create histograms of the residuals, which are the differences between the observations and the values that would have been observed had the data followed the assumed model exactly.

2.4 Randomization Tests

Randomization-based tests relieve the analyst of the need to nominate a formal model under the null hypothesis, by providing an empirical estimate of the distribution of the test statistic under the null hypothesis. That is, instead of nominating a theoretical distribution to use as a basis for comparison with the test statistic, an empirical null distribution is created, using simulation. The procedure outlined in the following section is a randomization test. Welsh (1996) provides a readable introduction.

3 A Test Procedure

3.1 Two Sets of Labels (t-test)

Assume that we have a collection of n independent persistence diagrams and a tentative labeling scheme that divides the collection into two possibly dissimilar collections, say $\mathcal{X}_1$ containing n_1 diagrams and $\mathcal{X}_2$ containing n_2 diagrams. For example, we may conjecture that the persistence diagrams that represent fMRI data of two groups of patients – one group with a condition of interest, and one without — are dissimilar. The assumption of independence precludes the possibility that any of the observations may have an influence on any of the other observations and is important for generating the null distribution. Our goal is to assess the strength of evidence that the processes that generated the collections $\mathcal{X}_1$ and $\mathcal{X}_2$ differ.

We realize this goal in the NHST framework as follows. We take as the null hypothesis the claim that the labels are exchangeable; that is, informally, that the current configuration of labels is no less likely than would have happened under a random labeling scheme, relative to the test statistic. An example of this reasoning follows. Given three tosses of a fair coin, each possible configuration has an identical probability — 0.125. However, from the point of view of counting the number of heads in three tosses, as a test statistic, it is much less likely that the count will be three (for a fair coin, 0.125) than two (for a fair coin, 0.375). The same reasoning holds in the proposed test: even though each possible configuration of the label is equally possible under the null hypothesis, we conjecture that very many of the random configurations lead to a value of the test statistic that is quite different to that in the observed sample.

Randomization tests that are used to compare two numerical samples usually focus on some function of the distance of the means of the samples. In the current study, computing the means is expensive, therefore computing the distance from each observation to the means for each simulated set of labels will also be expensive. We therefore instead nominate a function of the within-group pairwise distances as a test statistic. This statistic needs to be computed only once for each possible pair, and be stored in a table. Then simulation can proceed by summing the distances of pairs of observations that are randomly allocated to the same group.

When the observations are on the real line, and the measure of location is obtained by minimising the L2 norm, the location estimate is the mean, and the L2 norm is a monotonic function of the variance. In the proposed setup we have two putative means and two putative variances to consider. The joint loss of any labelling scheme, conditional on the sample sizes, can be expressed as the sum of the group-wise variances. Hence we propose that taking the mean or the sum of the variances of the two groups would be a sensible test statistic. The usual expression for the sample variance (for sets of real numbers), which is in the form closest to the L2 norm evaluated at its minimum, is

$$\sigma^2_{\mathcal{X}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

however, an equivalent variation can be computed without first calculating the mean, namely

$$\sigma^2_{\mathcal{X}} = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 \tag{3}$$

We suggest the sum of the group variances computed in this way as the test statistic.

For persistence diagrams with labeling L into the sets $\mathcal{X}_1 = \{X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}\}$ and $\mathcal{X}_2 = \{X_{2,1}, X_{2,2}, \ldots X_{2,n_2}\}$ we thus get the joint loss function

$$\sigma^2_{\mathcal{X}_{12}}(L) = \sum_{m=1}^{2}\frac{1}{2n_m(n_m-1)}\sum_{i=1}^{n_m}\sum_{j=1}^{n_m} d(X_{m,i}, X_{m,j})^2 \tag{4}$$

where $d(\cdot,\cdot)$ is the distance function in (1). It is possible to use alternative distance functions and indeed a different distance function may be more appropriate for persistence diagrams generated from some data sets.

If the grouping is sensible then this statistic will be small.

The total number of permutations of the labels is $\binom{n_1+n_2}{n_1}$, which is usually far too large to check exhaustively. We instead sample from the set of permutations with uniform probability. The randomization NHST algorithm is presented as Algorithm 1.

The advantage of using (3) instead of (2) is that the matrix of pairwise distances only has to be computed once, and the means of the randomly generated samples are not calculated. Randomly shuffling the group labels amounts to reading cells from the precalculated distance matrix.

The output, Z, of the Algorithm 1 is closely related to various probabilities. Since each of the relabelings is chosen independently we know that

$$\mathbb{E}(Z) = \mathbb{P}(\sigma_{\mathcal{X}_{12}}(L) \leq \sigma_{\mathcal{X}_{12}}(L_{\text{observed}}))$$

Algorithm 1: NHST algorithm for persistence diagram.

```
Data: $n_1 + n_2$ persistence diagrams with labels $L_{\text{observed}}$ in disjoint sets of size $n_1$
        and $n_2$, number of repetitions $N$
Result: p value
initialization - Z=0;
Compute $\sigma^2_{X_{12}}(L_{\text{observed}})$ for the observed labels;
for $i = 1$ to $N$ do
    Randomly shuffle the group labels into disjoint sets of size $n_1$ and $n_2$ to give
    labeling $L$;
    Compute $\sigma^2_{X_{12}}(L)$ for the new samples;
    if $\sigma^2_{X_{12}}(L) \leq \sigma^2_{X_{12}}(L_{observed})$ then
        $Z$ += 1
    end
end
$Z$ /= $N$;
Output $Z$
```

and the law of large numbers further ensures that $Z \to \mathbb{E}(Z)$ as $N \to \infty$ with probability one. The convergence rate is exponential.

We can justify calling our output a p-value by the following lemma.

Lemma 1. *Let $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ and $X_{2,1}, X_{2,2}, \ldots X_{2,n_2}$ be persistence diagrams drawn i.i.d. (the null hypothesis) and let α be the p-value computed by the above algorithm. Then for all $p \in [0,1]$ we have $\mathbb{P}(\alpha \leq p) \leq p$.*

Proof. Within the algorithm we randomly choose N different sets of labels of the set $\{X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}, X_{2,1}, X_{2,2}, \ldots X_{2,n_2}\}$, alongside the observed labels. Order these labelings by the cost, lowest first, randomly arranging amongst ties. Let the random variable W be the number of different labels appearing before the observed labels. Since under the model, the persistence diagrams are i.i.d., there is a uniform probability of the location of the original labeling over all the rankings. That is, W has a uniform probability over the natural numbers from 0 to N. This implies that $\mathbb{P}(W \leq k) = k$ for all $k \in \{0, 1, \ldots N\}$. Furthermore $\mathbb{P}(W/N \leq p) \leq p$ for all $p \in [0,1]$.

Note that there is a coupling between W/N and Z with $W/N \leq Z$. They agree except potentially in the case where there are multiple labelings with the same cost as the originally observed. Using this coupling we conclude that for all $p \in [0,1]$

$$\mathbb{P}(Z \leq p) \leq \mathbb{P}(W/N \leq p) \leq p.$$

□

4 Simulated examples

4.1 Point clouds of different shapes sampled with varying noise

Let K be the unit circle and let L be the wedge of a circle with radius 3/5 with a circle of radius 4/5. K and L have the same length. These are illustrated in Figures 1 and 2. The following process creates a point clouds from K and L.

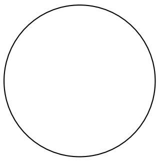

Figure 1: K

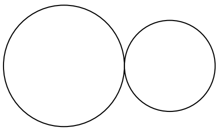

Figure 2: L

Fix the noise parameter $\sigma \geq 0$. Let μ_K and μ_L be uniform measures over K and L respectively. Draw 50 points i.i.d. using μ_K. Then to each of these points add an error drawn from $\mathcal{N}(0, \sigma)$. This is the same as drawing the original 50 points from the convoluted measure $\mu_K * \mathcal{N}(0, \sigma)$. We do the same procedure of drawing points from μ_L and adding noise.

For each run of the simulation we created 20 point clouds for K and L and computed their first homology persistence diagrams. We will just consider the first homology as this should be enough to distinguish them and they involve less off diagonal points and hence will be much faster. We then computed the corresponding p value within the algorithm above.

We ran this simulation 5 times each for each 0.01 increment of σ from 0 to 0.5. The results are tabulated in Figure 3.

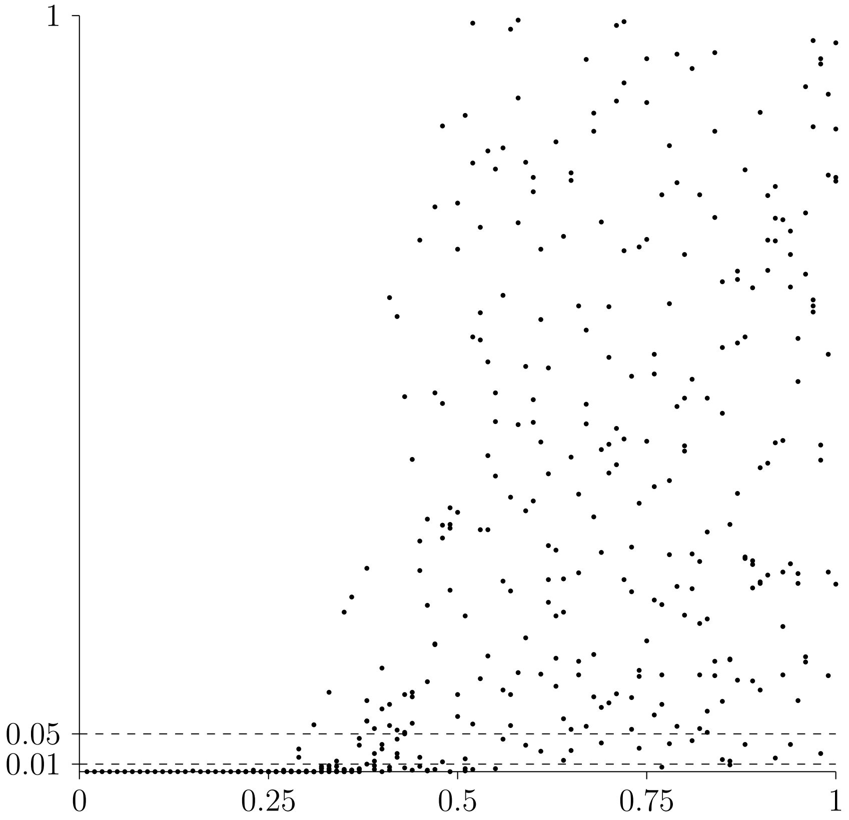

Figure 3: Simulated p values given the noise parameter for 20 point clouds each of K and L

This simulation demonstrates that, for this K and L, when the noise is sufficiently low then the p-values are low and hence we can tell that the persistence diagrams come from different distributions and hence the underlying shapes K and L must be different. When the noise increases we cannot reject the null hypothesis. This makes sense as the distributions $\mu_K * \mathcal{N}(0, \sigma)$ and $\mu_L * \mathcal{N}(0, \sigma)$ are closer when σ increases.

We also ran some analysis on the distribution of p values for a very similar simulation. For each run of the simulation we now only drew 10 point clouds (instead of 20) of K and L respectively and computed their first homology persistence diagrams. We then computed the corresponding p values by the algorithm above. We ran this simulation 200 times each for each 0.05 increment of σ (this is the parameter of normal noise convoluted with the uniform measure on K or L) from 0 to 0.5 and constructed the box plots as Figure 4.

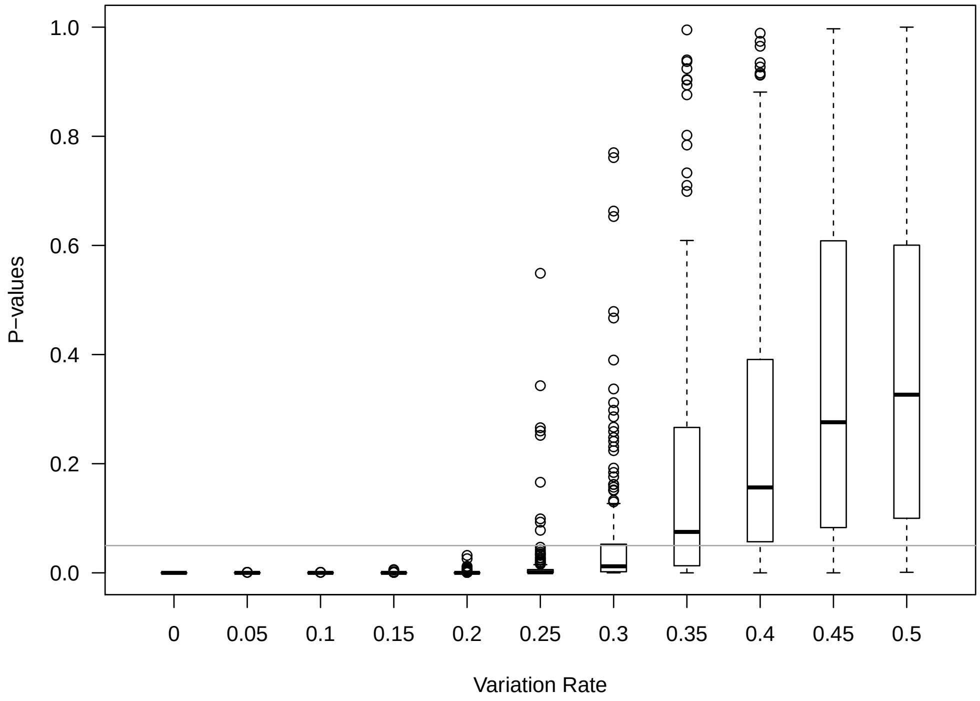

Figure 4: The box plots for the distributions of p values found by Algorithm 1 using the first homology persistence diagrams for noisy point clouds of K and L.

4.2 Point clouds of nearby shapes without noise

Now let M_0 be the circle of radius 1 as shown in Figure 5. Let M_β be two concentric circles with radius $1 - \beta$ and $1 + \beta$ as shown in Figure 6 for the case $\beta = 0.2$. We have the Hausdorff distance between K and M_β as β.

We will calculate the p-values of ρ_K and ρ_{M_β} for varying β and varying numbers of points in the point cloud. We will always have no noise.

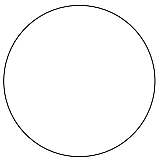

Figure 5: $K = M_0$

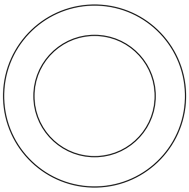

Figure 6: $M_{0.2}$

Fix the parameter $\beta \geq 0$ and a number of points m. Let μ_K and μ_{M_β} be the uniform measures over K and M_β respectively. Draw m points i.i.d. using μ_K and μ_{M_β} to create point clouds of K and M_β respectively.

For each run of the simulation we created 20 point clouds for K and L and computed their zeroth homology persistence diagrams (summaries of the changes in the set of connected components). We then computed the corresponding q_{12} within the algorithm above.

We ran this simulation 5 times each for each 0.01 increment of β from 0 to 0.5 and $m = 5, 10, 20$. The results are tabulated in Figures 7, 8 and 9.

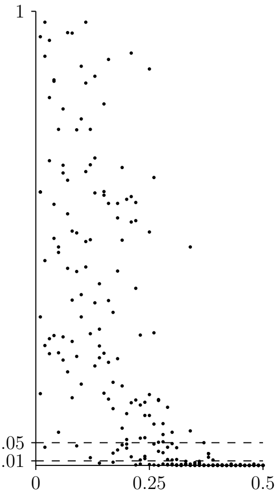

Figure 7: $m = 5$

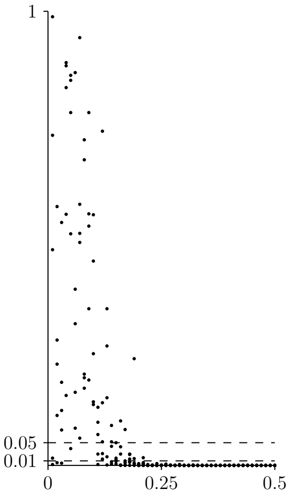

Figure 8: $m = 10$

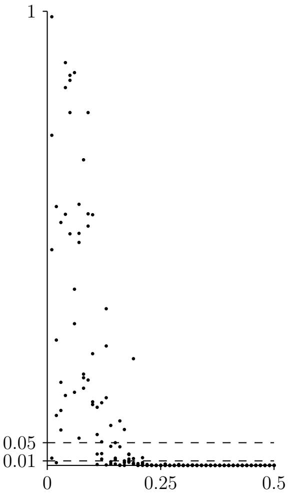

Figure 9: $m = 20$

As would be expected the ability to distinguish the two sets of diagrams increases as the sets are further apart and as the number of points in the point cloud increases.

4.3 Distinguishing sets of shapes from silhouette databank

In this example we will be using a variation on the the theme of a persistence diagram as a random element. Given a simplicial complex M in Euclidean space and a unit vector v we can create a filtration of M by the height function h_v in the direction of v and hence we can construct a persistence diagram $X(K, v)$ from the filtration of M by sublevel sets of h_v. The persistent homology transform of M is the function from the sphere of directions to the space of persistence diagrams where v is sent to $X(M, v)$. This process is explored in detail in Turner et al. (2013). There it is shown that the persistent homology transform of a shape is a sufficient statistic and is stable under perturbations of the shape. As such it is reasonable, given sets of shapes, to analyze the sets of their persistent homology transforms.

The distance squared between the persistent homology transforms of two shapes is effectively the integral over the unit sphere of the distances squared between the corresponding diagrams. This process can be made scale and translation invariant by appropriately modifying the diagrams $X(M, v)$. Furthermore it can be made rotation invariant by taking the infimum of all possible rotations. For details the reader is referred to Turner et al. (2013).[2]

A shape database that has been commonly used in image retrieval is the MPEG-7 shape silhouette database Sikora (2001). We used a subset of this database Latecki et al. (2000) which includes seven class of objects: Bone, Heart, Glass, Fountain, Key, Fork, and Axe. There were twenty examples for each class for a total of 1400 shapes. The shapes are displayed in Figure 10.

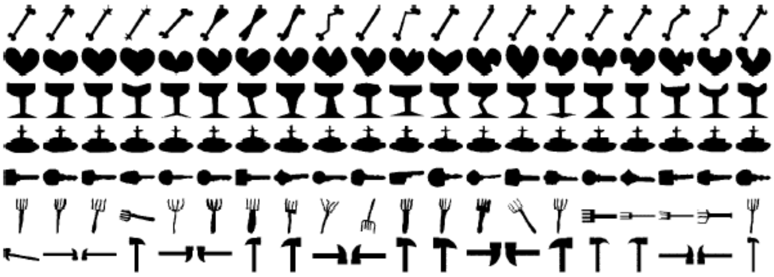

Figure 10: The subset of the silhouette database. Each row corresponds to one of the objects: Bone, Heart, Glass, Fountain, Key, Fork, and Axe. Note that although the objects are distinct, there is a great deal of variation within each object.

We used the perimeters of the silhouettes which are available at Gao (2004). We applied the alignment algorithm we stated in Section 3.3 of Turner et al. (2013) to shift, scale, and rotate the silhouettes. These perimeters are all homotopic to a circle so we used the 0-th dimensional persistent homology transform with 64 evenly spaced directions. We then computed the pairwise distances between each of the images. For each pair of classes we then computed the corresponding p values using 10000 repetitions. The algorithm always resulted in 0. This implies that we expect that $\mathbb{P}(\sigma_{\mathcal{X}_{12}}(L) \leq \sigma_{\mathcal{X}_{12}}(L_{\text{observed}})) < 0.0001$ and that the distributions of persistent homology transforms are very significantly different. This result implies that NHST-based classification via the persistent homology transform should be possible.

5 Example: Concurrence Topology in fMRI data

Given a set of variables and and samples of dichotomized data across those variables, concurrence topology is a method of creating filtration of a simplicial complex (and hence also persistence diagrams) to reflect the frequency of when subsets of the variables are simultaneously active. This method is studied in Ellis and Klein (2012) and applied to fMRI data for both subjects diagnosed with ADHD and healthy controls. We briefly describe the procedure and refer the reader to Ellis and Klein (2012) for details.

[2] The reader should note that in Turner et al. (2013) we focus on the L1 distance whereas here we are using the L2 distance. The definitions and results are analogous.

Some set locations in the brain were measured. For each time interval we get a number associated to how active that location in the brain is. These data are dichotomized by choosing a cutoff value. For each location in the brain we associate a vertex v_i. We assign to the vertex v_i the value of number of times that location was active. We assign to the edge $[v_i, v_j]$ the number of times both v_i and v_j were active simultaneously. Similarly assign to the face $[v_i, v_j, v_k]$ the number of times all three of v_i, v_j and v_k were active simultaneously. The same process assigns values to all simplices in the complete simplicial complex. The filtration is by superlevel sets. This is a simplification of the procedure. As part of the cleaning process some of the locations of in the brain are ignored and this set is different depending on the subject.

We calculated the p values associated with the sets of persistence diagrams in the "default mode network" that Ellis and Klein computed and kindly provided. In red are the p values which are ≤ 0.01. If we take a significance cutoff at $p = 0.01$ then our expected false discovery rate is much less than 1.

Table 1: Output of the algorithm with 10000 repetitions

Dimension	0	1	2	3	4	5
ADHD vs Control	0.75272	0.20537	0.50679	0.41815	0.10146	0.01162
ADHD vs Control in Females	0.68016	0.59175	0.77673	0.90267	0.00588	0.30057
ADHD vs Control in Males	0.46101	0.22070	0.59409	0.48437	0.41364	0.00975
Females vs Males in Control	0.00930	0.59964	0.33578	0.09851	0.19303	0.26304
Females vs Males in ADHD	0.48694	0.45473	0.60937	0.59045	0.02443	0.83618

A few comments should be made about the data set. The fMRI data set was generated at New York University and distributed as part of the 1000 Functional Connectomes project (`http://fcon1000.projects.nitrc.org/`). It includes 41 healthy controls (NewYork a part1) and 25 adults diagnosed with ADHD (NewYork a ADHD). Unfortunately the samples were highly imbalanced with respect to age and gender. Only 20% of the ADHD group was female, while about half of the controls were. About 25% of the controls were children (younger than 20; median age = 12), while there were no children in the ADHD group. Among adults, ages ranged from about 21 to about 50 in each group. The median age in the ADHD group was 37, while in the control group the median adult age was 27. We did not compute the p values while controlling for age.

6 Discussion / Conclusion

Here we have considered a cost function based of the sums of squared distance. This corresponds to the cost function that gives the standard deviation when applied to real numbers. Instead we could have calculated the actual (Frechet) variance. Alternatively, it is possible to consider quite different cost functions. In particular one can consider the total cost with sums of distance (rather than distance squared) or the maximal cost. A future direction is to study other cost functions in this framework of randomized tests to form related null hypothesis testing methods.

Another related problem is whether it is possible to do alternate hypothesis testing when the observations are persistence diagrams.

6.1 More than Two Label Sets

When the goal is to test whether K, $k > 2$ groups of observations differ, we can use an extension of the test statistic in equation 4 that is analogous to the F statistic in analysis of variance.

$$\sigma_{x_k}^2 = \sum_{m=1}^{k} \frac{1}{2n_m(n_m - 1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} (x_{mi} - x_{mj})^2 \tag{5}$$

This is not the same as the F statistic because we do not propose to compute the between-groups sums of squares, as that is expensive in this setting.

7 Acknowledgements

We thank Steve Ellis and Arno Klein for providing us with the persistence diagrams produced in their work. The authors would like to acknowledge the assistance of the Defence Science Institute in facilitating this work.

References

Balakrishnan, S., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. (2013). Statistical inference for persistent homology. *arXiv preprint arXiv:1303.7117.*

Bubenik, P. (2012). Statistical topology using persistence landscapes. *arXiv preprint arXiv:1207.6437.*

Bubenik, P. and Kim, P. T. (2007). A statistical approach to persistent homology. *Homology, Homotopy and Applications*, 9(2):337–362.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Duxbury Press, Belmont, CA.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2013). Optimal rates of convergence for persistence diagrams in topological data analysis. *arXiv preprint arXiv:1305.6239.*

Ellis, S. P. and Klein, A. (2012). Describing high-order statistical dependence using" concurrence topology", with application to functional mri brain data. *arXiv preprint arXiv:1212.1642.*

Gao, J. X. (2004). Visionlab. WWW. `http://visionlab.uta.edu/shape_data.htm`.

Latecki, L. J., Lakamper, R., and Eckhardt, T. (2000). Shape descriptors for non-rigid shapes with a single closed contour. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 424–429. IEEE.

Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Clarendon Press, Oxford.

Sikora, T. (2001). The mpeg-7 visual standard for content description—an overview. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):696–702.

Turner, K. (2013). Means and medians of sets of persistence diagrams. *arXiv preprint arXiv:1307.8300*.

Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2012). Fr\'echet means for distributions of persistence diagrams. *arXiv preprint arXiv:1206.2790*.

Turner, K., Mukherjee, S., and Boyer, D. M. (2013). Sufficient statistics for shapes and surfaces. *arXiv preprint arXiv:1310.1030*.

Welsh, A. H. (1996). *Aspects of Statistical Inference*. John Wiley & Sons, Inc., New York, NY.