Statistical Inference for Persistent Homology applied to fMRI

Hassan Abdallah[1], Adam Regalski[1], Mohammad Behzad Kang[1], Maria Berishaj[1], Nkechi Nnadi[1], Asadur Chowdury[2], Vaibhav A. Diwadkar[2], Andrew Salch[1]

[1]Dept. of Mathematics, Wayne State University, Detroit, MI
[2]Dept. of Psychiatry & Behavioral Neuroscience, Wayne State University, Detroit, MI

October 2020

1 Introduction

The purpose of this paper is to describe a method for determining whether certain time intervals within time series data are topologically distinguished from other time intervals in the same data set, in a statistically significant way. Here is an example, which also is the original motivating example for this paper: suppose we are given the data of an fMRI (functional magnetic resonance imaging) scan for a single patient. This data set consists of, for each time index t and each spatial coordinate (x, y, z) in some representative set of spatial locations within the physical space of the brain, a number $f(x, y, z, t)$, called the *fMRI signal amplitude*, which varies with the ratio of oxygenated hemoglobin to deoxygenated hemoglobin within the brain tissues near spatial location (x, y, z) at time t. The fMRI signal amplitude $f(x, y, z, t)$ is understood to vary, in an indirect and highly nonlinear way, with neuronal activity in the brain near (x, y, z) shortly preceding time t. In a typical fMRI study, the person in the fMRI machine is asked to carry out some temporally-structured cognitive task: for example,

Epoch 1 for fifteen seconds the person is asked to memorize the locations of symbols presented in a grid on a screen within the fMRI test chamber,

Epoch 2 and then for the fifteen seconds the screen is left blank while person is asked to recall the locations of the symbols.

We imagine that the experiment then repeats, returning to epoch 1. A typical data-analytic idea in fMRI is:

- to apply a "mask" to the data, removing all the fMRI data outside of some particular spatial region, to isolate one particular region (e.g. one particular brain lobe or other brain structure, like the hippocampus, the amygdala, etc.),
- and then to ask whether the temporal organization of the data into epochs can be recovered from the masked fMRI data in some statistically significant way.

If so, then the brain region within the mask is implicated in the cognitive task from the experiment. For example, if the mask isolates the hippocampus, then in our two-epoch example described above, a positive answer indicates that the hippocampus exhibits measurably different behavior during the memorization epoch (epoch 1) than during the recall epoch (epoch 2), so one could deduce that the hippocampus plays some role in memorization and recall. Of course there is nothing that restricts this deductive method to only be used for fMRI data: it is a generally useful idea for time series data originating from an empirical study.

With the rise of the use of topological methods in data analysis in the past ten years (see [15] for an introduction and brief survey) and in fMRI in particular (see [14] for an introduction and brief survey), we would like to combine the above idea with topological methods. The idea is to calculate the *persistence diagram* (see [12] for an introduction) of the time-series data at each time index separately, and then to ask whether the temporal organization of the data into epochs can be recovered from the persistence diagrams in some statistically significant way. A Monte Carlo test for statistically significant clustering of persistence diagrams was given in [13] and generalized in [2], but in both of those references, an independence hypothesis on the persistence diagrams makes the resulting test unsuited to time-series data. The purpose of this paper is to describe a multi-level block-sampled version of that Monte Carlo test which is indeed well-suited to time-series data, and to demonstrate its use on simulated fMRI data. We provide the R software package for this test, developed by our group, at `https://github.com/hassan-abdallah/TimeSeriesTDA`. While time-series data is the main area of application for this test, it is also useful on any other set of observations in which the independence hypothesis fails.

Our analytic method takes, as input, a set of observations (each of which is a point cloud) together with a labelling of the observations, a grouping of the observations into exchangeability blocks, and a labelling scheme for those blocks; a careful definition of this kind of structure is given in Definitions 4.1 and 4.2. The output of the analytic method is a p-value which reports on whether the persistent homology of the point clouds of each given label are statistically significantly distinct from the point clouds with other labels.

A brief introduction to TDA is given, though we direct interested readers to a thorough introduction and overview given by [4].

2 Background Knowledge on Topological Data Analysis (TDA)

Topological data analysis involves computations of *homology* and *persistent homology*. In this section, we offer a primer on those ideas, but for reasons of space, it is necessarily only a brief primer. For a more complete introductory treatment of persistent homology, see [12]. For a more comprehensive introduction to topological data analysis in general (rather than specifically persistent homology), we refer the reader to [4]. Even more generally, a more completely introductory treatment of homology can be found in any textbook on algebraic topology, such as the widely used book [8].

Before we begin, we note that persistent homology is defined on a choice of "point cloud"—that is, a finite subset of $\mathbb{R}^n$ for some n—together with a choice of coefficient ring. In most practical applications of TDA, the coefficient ring has been chosen to be the field with two elements, $\mathbb{F}_2 = \{0, 1\}$; see for example Table 3.1 in [11] for a 2015 list of commonly-used persistent homology software libraries which use $\mathbb{F}_2$ as either the default coefficient ring or as the only supported coefficient ring, e.g. Perseus, Dionysus, and GUDHI. We adhere to that convention in this paper: throughout, all homology is taken with coefficients in $\mathbb{F}_2$.

Now we sketch the definition of a simplicial complex and its homology. We begin with a set of points $v_0, v_1, ..., v_k$ in $\mathbb{R}^n$ such that the vectors $v_1 - v_0, v_2 - v_0, ..., v_k - v_0$ are linearly independent. Taking the convex hull $[v_0, v_1, ..., v_k]$ of this set, we form its *k-simplex*. A *face* of that k-simplex is then the convex hull of a proper subset of $\{v_0, v_1, ..., v_k\}$. So, for example, a 1-simplex is a line segment, and its faces are the endpoints of that line segment. Similarly, a 2-simplex is a solid triangle, and its faces are the edges of the triangle. A 3-simplex is a solid tetrahedron, and its faces are the triangles comprising the surface of the tetrahedron.

Next, consider a countable set K of simplices in $\mathbb{R}^n$ such that:

- for each simplex in K, each of its faces are also contained in K, and
- the intersection of two simplices in K is either a face of both simplices, or is empty.

Such a set K is known as a *simplicial complex*. The intuition here is that a simplicial complex K is a geometric object which is "built" by taking a union of simplices, allowing any two to intersect only along a common face. If a simplicial complex K has only finitely many simplices, then K is a *finite simplicial complex*.

Let K be a simplicial complex, and for each integer k, consider the vector space $V_k(K)$ of formal $\mathbb{F}_2$-linear combinations of k-simplices in K. That is, $V_k(K)$ is the vector space of *simplicial k-chains*. Then the *boundary map*,

extending to k-chains by linearity, is given by

$$\delta_k(K) : V_k(K) \longrightarrow V_{k-1}(K)$$
$$[v_0, v_1, ..., v_k] \longmapsto \sum_{j=0}^{k} [v_0, v_1, ..., \hat{v_j}, ..., v_k],$$

where $\hat{v_j}$ indicates that v_j is omitted from the simplex. The *simplicial chain complex* of the finite simplicial complex K is the sequence of $\mathbb{F}_2$-vector spaces and $\mathbb{F}_2$-linear functions

$$\cdots \xrightarrow{\delta_{k+1}} V_k(K) \xrightarrow{\delta_k} V_{k-1}(K) \xrightarrow{\delta_{k-1}} \cdots \xrightarrow{\delta_2} V_1(K) \xrightarrow{\delta_1} V_0(K) \xrightarrow{\delta_0} 0.$$

The image (that is, range) of $\delta_{k+1} : V_{k+1}(K) \to V_k(K)$ is called the vector space of *k-boundaries of K*, while the kernel (that is, nullspace) of $\delta_k : V_k(K) \to V_{k-1}(K)$ is called the vector space of *k-cycles of K*.

The boundary maps in the simplicial complex satisfy $\delta_k \circ \delta_{k+1} = 0$ for each integer k, that is, every k-boundary is also a k-cycle. Consequently we have a well-defined quotient vector space $\ker \delta_k / \operatorname{im} \delta_{k+1}$ which is trivial if and only if every k-cycle is a k-boundary. The vector space $\ker \delta_k / \operatorname{im} \delta_{k+1}$ is called the *kth homology of K*, written $H_k(K)$. When it is important to remember that the coefficient ring has been taken to be the field $\mathbb{F}_2$, we write $H_k(K; \mathbb{F}_2)$ instead of $H_k(K)$.

Now, given a simplicial complex K, consider a family $\{K_a : a \in \mathbb{R}\}$ of simplicial sub-complexes of K such that $K_m \subseteq K_n$ whenever $m \leq n$. That is, for each real number a, K_a is a simplicial sub-complex of K, and if a, b are real numbers with $a < b$, then every simplex in K_b is also in K_a. (So, as the subscript a gets smaller, the simplicial complex K_a also gets smaller.) The simplicial complex K together with the family $\{K_a : a \in \mathbb{R}\}$ is known as a *filtered simplicial complex*. For $a \leq b$, denoting the boundary maps on $V_k(K_a)$ and $V_k(K_b)$ by δ_k^a and δ_k^b, respectively, we naturally have inclusion maps $\iota : K_a \longrightarrow K_b$, which, in turn, gives inclusion maps $\iota : Im(\delta_{k+1}^a) \longrightarrow Im(\delta_{k+1}^b)$ and $\iota : Ker(\delta_k^a) \longrightarrow Ker(\delta_k^b)$.

If the simplicial complex K is finite, then for most pairs of real numbers $a < b$ with a sufficiently close to b, the subcomplex K_a of K_b is simply the entirety of K_b. There is only a *finite* list of real numbers b such that K_a differs from K_b for all $a < b$, no matter how close a is to b. Writing $b_1, b_2, \ldots, b_m$ for that finite sequence of real numbers, we have a sequence of $\mathbb{F}_2$-linear functions

$$0 \to H_k(K_{b_1}) \to H_k(K_{b_2}) \to \cdots \to H_k(K_{b_m})$$

called the *persistent homology groups of K*.

An element z of $H_k(K_{b_i})$ has a *birth radius*, that is, the least real number b_h such that z is in the image of the function $H_k(K_{b_h}) \to H_k(K_{b_i})$. Similarly, z has a *death radius,* that is, the least real number b_j such that z maps to zero under

the function $H_k(K_{b_i}) \to H_k(K_{b_j})$. If the image of z is nonzero in $H_k(K_{b_j})$ for all b_j, then the death radius of z is defined to be ∞. (The birth radius of z, however, is always finite.)

We are now prepared to define the *persistence diagram*. The k^{th} *persistence diagram* of the k^{th} persistence module is a multiset[1] of points in $\mathbb{R} \times (\mathbb{R} \cup \{\infty\})$. Each point in the diagram represents a homology class; the x-coordinate of the point representing a homology class z is the birth radius of z, while the y-coordinate of that point is the death radius of z. By convention, we include (with infinite multiplicity) all points such that $x = y$ (that is, the points lying along the diagonal). The further a point is from the diagonal of a persistence diagram, the longer the homology class *persists* (i.e., is nonzero) as the filtration parameter ranges over the real numbers. The intuition here, then is that the closer a point in the persistence diagram is to the diagonal, the more we think of the topological feature represented by that homology class as a kind of "topological noise," rather than a meaningful topological pattern involving and organizing a large part of the data set.

The typical intended use of persistent homology for the sake of data analysis is that one begins with a point cloud, one builds a finite filtered simplicial complex whose structure reflects the geometry of the point cloud in some desired way, and then one calculates the persistent homology groups of that filtered simplicial complex. We have explained the last step, but we have not yet explained how to build a finite filtered simplicial complex from a point cloud. There are several ways to do this: a point cloud has an associated Čech complex, Vietoris-Rips complex, Delaunay complex, witness complexes, and others, each of which is a finite filtered simplicial complex whose structure "encodes" the geometry of the point cloud in some particular way. See [5] for discussion and comparison of the Čech and Vietoris-Rips complexes, for example. For brevity, here we do not attempt a survey of these various filtered simplicial complexes, but we at least give a definition of the Čech complex, since it is the most geometrically straightforward: given a point cloud $X \subseteq \mathbb{R}^n$ and a subset U of X, the *diameter of* X is the least real number ϵ such that every element of U is contained in a closed ball of radius ϵ in $\mathbb{R}^n$. The *Čech complex of* X is the filtered simplicial complex $\{K_a : a \in \mathbb{R}\}$ such that K_a is the union of the convex hulls of each of the subsets of X of diameter $< a$.

This has necessarily been only a brief sketch. For more details, the reader can consult the references cited at the start of this section.

[1] Recall that a "multiset" is a set with (unordered) multiplicities, that is, an element of a multiset can be contained in that multiset "multiple times." A typical way to make this intuitive idea rigorous is to simply think of a multiset as an ordinary set S equipped with an equivalence relation. Given an element s of S, the multiplicity of s in S is understood to be the number of elements in the equivalence class of s.

3 Summary of Existing Hypothesis Testing Methods for Topological Data Analysis

The purpose of this section is to discuss (and extend via [2]) the methods used in [13] which we will apply to our fMRI data. To clarify, everything we discuss in this section is due to [13] and [2].

We begin by involving some ideas from statistics alongside the basic notions in persistent homology. The idea here is as follows: imagine that we calculate the persistent homology of each data set in a collection of data sets gathered from one real-world process (for example, fMRI scans of the brains of people who are reading words from a screen), and we also calculate the persistent homology of each data set in a collection of data sets gathered from a different real-world process (for example, fMRI scans of the brains of people who are looking at a blank screen). We would like a statistical test (namely, a hypothesis test) to determine whether the persistence diagrams from the first group are distinguishable from the persistence diagrams from the second group. In order to assess the strength of evidence against the claim that the two processes are the same, we can study the distributions of persistence diagrams from each process. The goal of Robinson and Turner's work in [13] is to use hypothesis testing to compare two groups of persistence diagrams. The methods discussed in [13] are extended in [2] in order to use hypothesis testing to compare multiple groups of persistance diagrams. The need for us to extend the comparisons between persistence diagrams to 3 or more groups of persistence diagrams comes from the multi-level block sampling framework that we apply to our time-series data in the next section of this paper, where we freely permute multiple blocks (and, hence, multiple groups of persistence diagrams) to carry out our hypothesis test.

Our hypothesis test begins with a set of n persistence diagrams divided into s groups $\beta_1 = \{X_{1,1}, X_{1,2}, ..., X_{1,n_1}\}$, $\beta_2 = \{X_{2,1}, X_{2,2}, ..., X_{2,n_2}\}$, ... , $\beta_s = \{X_{s,1}, X_{s,2}, ..., X_{s,n_s}\}$ containing $n_1, n_2,$ ... , n_s diagrams, respectively, with this division into multiple groups done according to some initially-chosen labeling scheme. The hypothesis test corresponding to the case $s = 2$ is the subject of [13], while the generalization to arbitrary finite s was the focus of [2]. The null hypothesis is that, if one "shuffles" the set of persistence diagrams between the two groups by randomly permuting the labels, then the difference observed between the sets of diagrams based on this "shuffled" labeling would be just as likely as the differences observed between sets of diagrams with the original, correct labels. More generally, the null hypothesis is that the two processes that determine the original labeling scheme one has chosen (and, hence, the initial split into two groups of persistence diagrams) are not significantly different. An observed test statistic is computed using the initial labeling scheme, and computed further for each permutation of labels in the permutation test. The key to computing the final p-value, which assesses the strength of evidence against the null hypothesis, then, is to compute the ratio of permutations that

yield a test statistic more extreme than the observed statistic to the total number of permutations. We note that a necessary assumption for the test is that observations (respectively, persistence diagrams) are independent.[2] Also, the permutation test we carry out is a randomization test. As mentioned in Section 2.6 of [13], using a randomization test avoids any need to hypothesize a distribution model from which persistence diagrams are drawn under the null hypothesis .

3.1 Metric on Persistence Diagrams

In order to carry out our hypothesis test, we first need to introduce a metric, i.e., a distance function, on the space of persistence diagrams. This metric allows us to compare two persistence diagrams and is a key piece of the test statistic that we'll utilize in this hypothesis test.

The appropriate distance metric between persistence diagrams X and Y that we consider is the bottleneck distance

$$d_\infty(X,Y) = \inf_{\text{bij. } \phi: X \longrightarrow Y} \sup_{x \in X} ||x - \phi(x)||_\infty$$

which occurs as the limit of the metric

$$d_p(X,Y) = \left(\inf_{\text{bij. } \phi: X \longrightarrow Y} \sum_{x \in X} ||x - \phi(x)||_p^p \right)^{1/p}$$

as p goes to infinity, where $||x - \phi(x)||_p^p$ is the L^p-norm between x and $\phi(x)$ raised to the p-th power, and the infimum is taken over all bijections ϕ between the points of X and the points of Y. Note that, as a metric on the space of persistence diagrams, the bottleneck distance $d_\infty(X,Y)$ between X and Y is indeed symmetric. This follows since a bijection $\phi : X \longrightarrow Y$ also defines a bijection $\phi^{-1} : Y \longrightarrow X$. We now unpack the construction of these metrics.

The metrics take into account an optimal bijection $\phi : X \longrightarrow Y$ between the points of X and the points of Y. A bijection $\phi : X \longrightarrow Y$ is said to be "optimal" if it minimizes the total cost $\sum_{x \in X} ||x - \phi(x)||^2$. Optimal bijections are found by using the Hungarian algorithm. Given two sets of elements $S = \{s_1, ..., s_n\}$ and $T = \{t_1, ..., t_n\}$, and a square matrix A, where the ith row of A is represented by the element s_i and the jth column is represented by the element t_j, one can apply the *Hungarian algorithm* to A to find the optimal bijection between elements of S and elements of T. (The original reference for the Hungarian algorithm is the 1955 paper [9], but today the Hungarian algorithm is a standard topic

[2] Since our goal is to have a statistical test that can be applied to the persistence diagrams of *non-independent* time series data, in the next section, we apply a multi-level block sampling framework to satisfy the exchangeability criteria for our permutation test, thereby removing the requirement of our observations being independent.

covered in many discrete mathematics textbooks, so many modern expositions are available.)

Here is a bit more detail about what the bottleneck distance between two persistence diagrams is. If X has points $x_1, ..., x_n$ and Y has points $y_1, ..., y_m$, one takes copies $x_{n+1}, ..., x_{n+m}$ and $y_{m+1}, ..., y_{m+n}$ of the diagonal in a persistence diagram, where this diagonal is the line of slope 1 in the birth-death plane, and constructs the $(n+m) \times (n+m)$ matrix in which the (i, j) entry is the cost $||x_i - y_j||_2^2$. When one of x_i or y_j is a copy of the diagonal, this is simply the perpendicular distance between x_i and y_j. When both x_i and y_j are copies of the diagonal, the cost is simply 0.

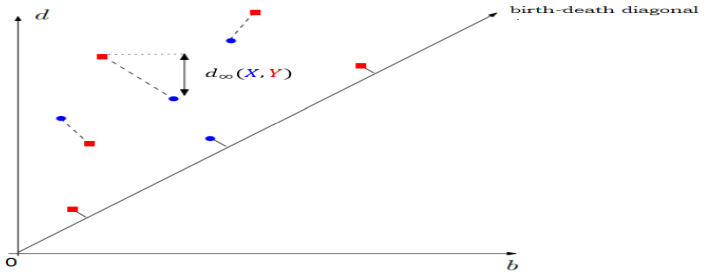


Figure 1: A visual representation of bottleneck distance $d_\infty(X, Y)$ between two persistence diagrams X and Y. The blue points represents points in persistence diagram X, while red points represents points in persistence diagram Y. A line drawn from a point $x \in X$ (respectively, $\phi(x) \in Y$) to a point $\phi(x) \in Y$ (respectively, $x \in X$) represents the assignment of x to $\phi(x)$ using the definition of the bottleneck distance between X and Y. Note that, one can see that the totality of assignments of points in X to points in Y is one that minimizes the sum of distances between points in X and points in Y under the described bijection ϕ. This figure is essentially taken from [3].

3.2 Test Statistic and p-value for Comparing Groupings of Persistence Diagrams

Now that we have established an appropriate metric on two persistence diagrams, we can formulate a test statistic for our hypothesis test. The test statistic is the joint loss function given by

$$F'_{p,q}(\{X_{1,i}\}, \{X_{2,i}\}, ..., \{X_{s,i}\}) := \sum_{m=1}^{s} \frac{1}{2n_m(n_m - 1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} d_p(X_{m,i}, X_{m,j})^q,$$

where $p \in [1, \infty)$, $q \in [1, \infty)$. This joint loss function, as a test statistic, was introduced in [2] as a generalization of the $s = 2$ case considered in [13]. (In our application of these ideas, we take p to be infinity and q to be 1.) Since the groups β_1, β_2, ... , β_s are determined by a choice of labeling L, we will use

the notation $F'(L)$ to mean the joint loss function computed on the s groups of persistence diagrams determined by L, and $F'(L_{observed})$ to mean the joint loss function computed on the s groups of diagrams determined by the initial choice of labeling. When implemented in software, the pairwise distances between persistence diagrams are only computed once and stored in a table. Note that the test statistic given by the joint loss function takes into account distances between observations (respectively, persistence diagrams), rather than distances between observations and the mean. This is because the latter consideration is very computationally expensive[3].

Taking α to be the proportion of all labelings L such that $F'(L) \leq F'(L_{observed})$, where all of the possible labelings L are determined by the permutation test carried out that permutes the labels on persistence diagrams, we can now generalize (via [2]) the algorithm developed in [13] to compute the proportion α to be taken as the p-value (after a standard modification to α in order to avoid a p-value of 0). The difference is that, rather than having $n_1 + n_2$ persistence diagrams with labels $L_{observed}$ in disjoint sets of size n_1 and n_2 and randomly shuffling the group labels into disjoint sets of size n_1 and n_2 to give the labeling L, we now have $n_1 + n_2 + ... + n_s$ persistence diagrams with labels $L_{observed}$ in disjoint sets of sizes $n_1, n_2, ..., n_s$ and we randomly shuffle the group labels into disjoint sets of sizes $n_1, n_2, ..., n_s$ to give the labeling L. It is shown in [13] that the modified α is a true p-value, and by Lemma 1 of [13], α is an unbiased estimator of the permutation p-value under the assumption that the persistence diagrams are i.i.d. As mentioned before, our goal in this paper is to adapt the Robinson-Turner test to the common real-world situation of time series data which is not independent, and consequently our persistence diagrams, regarded as observations, are not independent observations. However, again, we're able to correct for this using our methods in the next section.

4 Hypothesis Testing for Topological Data Analysis extended to Non-Independent Data

In this section, we describe a single and multi-level block variation of the original Monte Carlo test that allows for the analysis of non-independent data sets. The primary idea involves accommodating the unique *exchangeability* structure of a particular set of data.

4.1 Exchangeability

A sequence of random variables $X_1, X_2, ..., X_n$ is *exchangeable* under a set of permutations Π of $\{1, 2, .., n\}$ if it has the same joint distribution as the sequence $X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(n)}$ for every $\pi \in \Pi$. Determining the set Π for which

[3] Also, there is not a clear notion of the mean of a set of persistence diagrams.

exchangeability holds is critical to perform sound statistical inference via a permutation test. If the joint distribution of a set of data changes under particular permutations of labels, then the distribution of a test statistic under those permutations is not suitable to be compared to the observed test statistic and could elicit spurious results. In the case of independent and identically-distributed random variables, the set Π contains all permutations of $\{1, 2, .., n\}$, meaning labels may be freely exchanged during a permutation test. As a result, the hypothesis testing framework described in Section 2 did not require any considerations of exchangeability. In many cases, however, the set of permutations that satisfy the above exchangeability criterion is far more restrictive. Fortunately, by restricting permutations to the set Π while generating the distribution of a test statistic, a permutation test may proceed without the iid requirement.

In practice, implementing a restrictive set of permutations is done via a multi-level block shuffling scheme, as in [17]. Instead of exchanging the label of one observation with another, shuffling takes places across blocks of data called *exchangeability blocks*. Exchangeability blocks can either be shuffled as a whole(defined as whole-block exchangeability) or labels may be shuffled within a block(defined as within-block exchangeability). The block sizes and attributes are chosen in accordance with the permitted set of permutations.

For example, consider an fMRI experimental design that consists of 120 total scans. Suppose a stimulus or task is administered every 10 scans and lasts for 10 scans. Our whole-block exchangeable level in this scenario would be defined as each contiguous set of 10 scans starting at 1, accounting for 12 blocks in total. These first-level exchangeability blocks ensure that any label shuffling would result in a set of labels exhibiting a similar contiguity to the initial set of labels (those assigned stimulus/no stimulus during the experiment) and retain structure related to the experiment design. They do not, however, account for a temporal dependence structure across the whole of the experiment. Without additional restrictions, labels may be shuffled into two groups where one group is the data associated with scans 1-60 and the other is the data associated with 61-120. A distinguishing pattern across early versus late stages of fMRI experiments have been noted so the test statistic computed for this set of labels could display an extremeness resulting from this temporal phenomena [10]. As such, a within-block exchangeability level is necessary.

A within-block exchangeability level would consist of two blocks, one covering scans 1-60 and another covering scans 61-120. In this design, the whole-block exchangeable blocks present in labels 1-60 could only be exchanged amongst themselves and not with their corresponding blocks in labels 61-120. This ensures that the first half of the experiment and the second half of the experiment would have equal representation in any set of labels shuffled under this scheme, accounting for early versus late confounding.

In the context of TDA hypothesis testing, we define a *two-level point cloud grouping* and associated *labelling scheme* to encode the multi-level block shuf-

fling technique described above:

Definition 4.1. A *(two-level) point cloud grouping* is the following data:

1. A set T.
2. A function pc from T to the set of all point clouds.
3. A partition T into subsets $T_1, ..., T_n$ and,
4. A partition T into subsets $T_1^{'}, ..., T_m^{'}$ which is finer than the partition $T_1, ..., T_n$ of T.

Definition 4.2. Given a two-level point-cloud grouping X, a 2-group *labelling scheme* on X is a partition of T into subsets S_1 and S_2 by the following:

1. For each $i \in \{1, .., n\}$, $\exists\ j_1, .., j_{k_i} \in \{1, .., m\}$ such that $\bigcup_{a=1}^{k_i} T_{j_a}^{'} = T_i$. For each i, choose $k_i/2$ elements without replacement, i.e. without repetition, from the set $\{j_1, ..., j_{k_i}\}$, denoted $\{s_1^i, ..., s_{k_i/2}^i\}$. Then define $S_1 = \bigcup_{i=1}^{n} \bigcup_{k=1}^{\frac{k_i}{2}} T_{s_k^i}^{'}$ and $S_2 = T - S_1$

The partition $T_1, .., T_n$ represents the whole-block exchangeable level and the partition $T_1^{'}, .., T_m^{'}$ represents the within-block exchangeable level. Permutations of labels are then obtained by generating distinct labelling schemes as defined above. In the fMRI example described above, T would be the set of a time indices from 1 to 120 and the partitions would be defined as follows:

$$
\begin{array}{llll}
T_1 = \{1:10\} & T_2 = \{11:20\} & T_3 = \{21:30\} & T_4 = \{31:40\} \\
T_5 = \{41:50\} & T_6 = \{51:60\} & T_7 = \{61:70\} & T_8 = \{71:80\} \\
T_9 = \{81:90\} & T_{10} = \{91:100\} & T_{11} = \{101:110\} & T_{12} = \{111:120\} \\
T_1^{'} = \{1:60\} & T_2^{'} = \{61:120\} & &
\end{array}
$$

The function pc would map an element of T to its corresponding point cloud.

4.2 Overview of Analysis Pipeline

In this section, a broad overview of the steps to go from a data set to a p-value for a hypothesis are given. Several of the computational tasks involved can be accomplished using our R package "TimeSeriesTDA". Beginning with a data set of interest, carry out the following:

1. Compute persistent homology on all observations of your data set to produce a collection of persistence diagrams. Each observation should be a point cloud.

2. Generate a hypothesis that conjectures a significant difference between two sub-collections of your collection of persistent diagrams, called *groupings*. The null hypothesis is that the two groupings are not significantly different from each other. Choose an α level for rejecting the null hypothesis. For example, for $\alpha = 0.05$, the null hypothesis will be rejected if the resulting p-value of this hypothesis test is less than 0.05.

3. Compute the value of the appropriate joint loss function given labels for the above groupings, called $F'(L_{observed})$ as defined in Section 3.

4. Determine the exchangeability structure of your observations and encode it in a two-level point cloud grouping. In particular, define the set T and partitions of T corresponding to whole-block and within-block exchangeability levels, as in Definitions 4.1 and 4.2.

5. Generate distinct labelling schemes and recompute the joint loss function value for each new set of labels. Compute a p-value by taking the proportion of permuted labels L such that $F'(L) \leq F'(L_{observed})$. Compare the p-value to the pre-determined α-threshold to evaluate whether the null hypothesis will be rejected or not.

5 Application to fMRI data.

fMRI imaging is an excellent and rich source of non-independent time-series data. The data obtained from an fMRI scan is time-series data consisting of, at each time index t, a real number $f(x, y, z, t)$ at each point (x, y, z) in a certain set of lattice points in $\mathbb{R}^3$. The number $f(x, y, z, t)$ is the *fMRI signal amplitude*, which is understood to vary (non-linearly) with the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood in the tissues near physical location (x, y, z) at time t within the fMRI machine. That is, fMRI data is time series data, such that at each time index, we have a point cloud in $\mathbb{R}^4$: three spatial dimensions, and one signal amplitude dimension. The fMRI signal amplitude has a relationship to unfolding biological processes in the brain. These processes are, at each moment in time, potentially dependent on their states at prior moments in time.

Before applying persistent homology, a suitable normalization technique for the fMRI signal needs to be identified such that the 4-dimensional point clouds obtained from fMRI data are organized in such a way that persistent homology is adequately sensitive to evolving topological structure. Additionally, parameters related to persistence, such as maximum birth and death radiuses to compute persistent homology to, and cutoffs for persistence diagram feature selection, need to be explored in the context of fMRI data. In this section, we discuss each of these decisions (normalization method, and parameter choices) in turn.

5.1 Normalization

At each individual time index, the structure of fMRI data consists of three spatial coordinates and a signal amplitude coordinate. When discovering topological features in the 4-dimensional point cloud, one would hope that the same features would be obtained regardless of the choice of units. In the case of fMRI data, this poses an issue as the three spatial coordinates are measured in millimeters, whereas the signal amplitude is unitless. A change in units of distance would rescale the three spatial dimensions but not the fourth (signal amplitude), changing the topological features and persistence diagrams acquired. Consequently, in order to yield results that are invariant under changing units, fMRI data must be normalized before calculating persistence diagrams. Different choices of how to normalize fMRI data may yield different persistence diagrams, so topological structure in fMRI data is impacted by *how* we normalize the data. Below, two methods of normalizing fMRI data are discussed. In section 5.4, we report on which of these two normalization methods, when applied to our simulated fMRI data, allow our statistical test to achieve greater statistical power.

Definition 5.1 (Normalization Scheme 1)**.** Define the following notation:

$$S_{\min} = \min\Big\{ \min\{x-\text{coordinates}\}, \min\{y-\text{coordinates}\}, \min\{z-\text{coordinates}\}\Big\}$$
$$S_{\max} = \max\Big\{ \max\{x-\text{coordinates}\}, \max\{y-\text{coordinates}\}, \max\{z-\text{coordinates}\}\Big\}$$
$$A_{\min} = \min\Big\{\text{signal amplitude}\Big\}$$
$$A_{\max} = \max\Big\{\text{signal amplitude}\Big\}$$

where the minimums and maximums are taken for each time slice and each subject individually. For any given coordinate (x, y, z, ϵ), replace ϵ (the signal amplitude) with

$$\left[\frac{\epsilon - A_{min}}{A_{\max} - A_{\min}} \cdot (S_{\max} - S_{\min})\right] + S_{\min}.$$

Definition 5.2 (Normalization Scheme 2)**.** Define the following notation:

$$S_{\min} = \frac{\min\{x-\text{coordinates}\} + \min\{y-\text{coordinates}\} + \min\{z-\text{coordinates}\}}{3}$$
$$S_{\max} = \frac{\max\{x-\text{coordinates}\} + \max\{y-\text{coordinates}\} + \max\{z-\text{coordinates}\}}{3},$$

where the minimums and maximums are taken for each time slice and each subject individually. We let $A_{\min}$ and $A_{\max}$ be as in Definition 5.1. For any given coordinate (x, y, z, ϵ), replace ϵ (here, the fMRI signal amplitude) with

$$\left[\frac{\epsilon - A_{min}}{A_{max} - A_{min}} \cdot (S_{max} - S_{min})\right] + S_{min}.$$

Normalization Scheme 2 is preferred due to the fMRI signal amplitude similarity to the spatial coordinates range, maximums, minimums, and magnitudes. This occurs because Scheme 2 utilizes the averages of the spatial coordinates. We demonstrate this preference in the following example:

Example 5.1. Suppose we consider fMRI data whose spatial coordinates have x, y, and z coordinates in the ranges $x \in (35, 57), y \in (65, 90), z \in (32, 63)$. Using Normalization Scheme 1 linearly rescales the fMRI signal amplitudes, at each time slice, so that the normalized signal amplitudes lie in the range $(32, 90)$. On the other hand, Normalization Scheme 2 linearly rescales the fMRI signal amplitudes so that the normalized signal amplitudes lie in the range $(44, 70)$. The reason an average is preferred is due to the sizes of the ranges: the normalized amplitudes under the first method lie in an interval of length 58, while the normalized amplitudes under the second method lie an interval of length 26, which is closer to the ranges of the spatial coordinates, since the lengths of the ranges for the x, y, and z coordinates are 22, 25, and 31, respectively. In this example, we see that the second normalization scheme yields a normalized fMRI amplitude whose properties more closely mirror the properties of the spatial coordinates. See section 5.4 for empirical calculations of the power of our statistical test when applied to simulated fMRI data with each of the two normalization schemes.

5.2 Parameter considerations

The parameters necessary for computing persistent homology either determine how comprehensive of a view of our data is obtained, or rely on canonical choices informed by characteristics of our data. The first consideration is the maximum filtration parameter to which to compute the persistent homology. For example, for the Čech filtration (defined in section 2), the ideal choice for this parameter is half the distance of the two farthest points in a data set, since there are no non-trivial changes in the topology of the space beyond that radius (the topology is that of a single convex body). This is an example of a "canonical choice" of the filtration parameter.

However, in most fMRI data sets, computing persistent homology up to that distance is not computationally feasible. Instead, a threshold value is chosen such that it has the potential to capture nontrivial topology, and the process completes in a reasonable amount of time. For example, using maximum filtration parameter 1 or 2 with the two normalization techniques previously discussed yields virtually no one-dimensional homological features in fMRI data. This is not because those features are not present, but rather because the birth radius or death radius of those features is greater than 1 or 2. Using maximum radius 3 or 4, on the other hand, is large enough to capture interesting topological information. It is important to determine whether tweaking this choice of parameter alters results (and we present some conclusions to this effect in section 5.4), since as this parameter changes, so does the hypothesis and statistical conclusion of our test. For example, rejecting the null hypothesis would

show that there is enough evidence to support the claim that the two groups of persistence diagrams, *up to persistence* (= maximum radius), are statistically significantly different from each other. This not only indicates differing topological structure, it also indicates the maximum size and scale of the topological structures that influence the result.

Another parameter is related to the number of features of persistence diagrams retained for our analysis. It has been found that reasonably sized sets of fMRI data (>1000 4-dimensional points) contain potentially thousands of 1-dimensional homological features. It is not tractable to compute a distance matrix between more than a few dozen persistence diagrams when each has that many features. Fortunately, there are methodological considerations for filtering out a large subset of features. Masked fMRI data is composed on a lattice with distance 1 between adjacent points. Setting signal to zero for all points, computing persistent homology on such a space would result in an abundance of features with persistence $\sqrt{(2)}/2$ ($\approx$0.707). It is then reasonable to infer that features at that persistence and below are likely more related to small-scale "topological noise" rather than large-scale, meaningful topological organization within the data. Thus, our initial cutoff for minimum persistence threshold is 0.8. In our results section, we determine whether increasing that cutoff gives a more powerful test or not.

5.3 fMRI Data Simulation

Here, we discuss how we generated the results of a simulated fMRI experiment in order to test the power, accuracy, and reliability of the proposed method. We used the R package neuRosim [16] to simulate the data.

5.3.1 Experimental Design

We generated simulated fMRI data with a repetition time (TR) of two seconds in a spatial region (i.e., a region of stereotactic space) in the shape of a standard fMRI mask of the hippocampus. Throughout each simulated run, the signal amplitudes in this spatial region are first given by a standard simulation of physiological noise. Physiological noise is intended to mimic noise caused by heart beat and respiratory rate. It is modelled by sine and cosine functions with the addition of Gaussian noise to increase variability across voxels.

The simulated data is structured so that, in each simulated run, there are six "epochs," consisting of 20 seconds each. At the onset of each epoch, the signal amplitudes are increased in a sphere-shaped region within the hippocampus-shaped region, depicted in the figures below. Activation is greatest at the beginning of each epoch and fades throughout.

5.3.2 Simulation Characteristics

With this experimental design we varied the characteristics of both noise and signal in the interest of deciding what, if any, topological structure this method

might detect. By varying noise and signal characteristics, the topology of the data will vary with it.

Each simulated data set was generated by the following process:

1. For each time index t from $t = 1$ to $t = 120$, set the signal amplitude in each voxel in a standard hippocampal mask to the values given by simulated physiological noise. Spatial (i.e., stereotactic) points outside the hippocampal mask are not considered at all, i.e., they are not set to have zero signal amplitude but are instead entirely absent from the data.

2. Choose a radius r (we considered the values $r = 1, 3, 5, 7$, and 15, in separate runs) and a point p in the mask. In a spherical region of radius r (measured in voxel edge lengths) with center p, replace the physiological noise signal with an "activated signal" of high amplitude at the start of each epoch, and decaying in amplitude throughout the epoch. We used a standard amplitude curve for simulated fMRI provided by neuRosim, depicted in Figure 3.

   It is necessary to choose the initial effect size (which can be thought of as a measure of the magnitude of activation) of the activation in the sphere. We generated data with initial effect sizes 2, 5, 10, and 20, to compare the results.

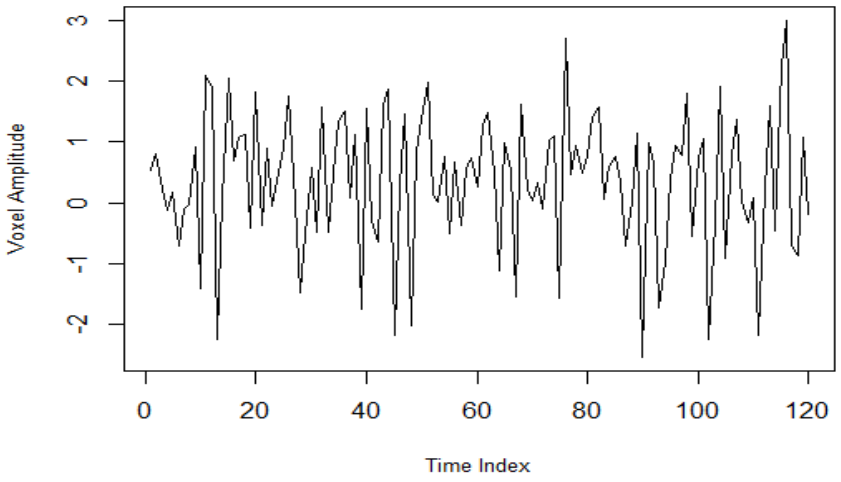


Figure 2: This shows the amplitude of a voxel outside of the embedded sphere that does not respond to the experimental task and has been simulated with physiological noise. Simulated with effect size = 5.

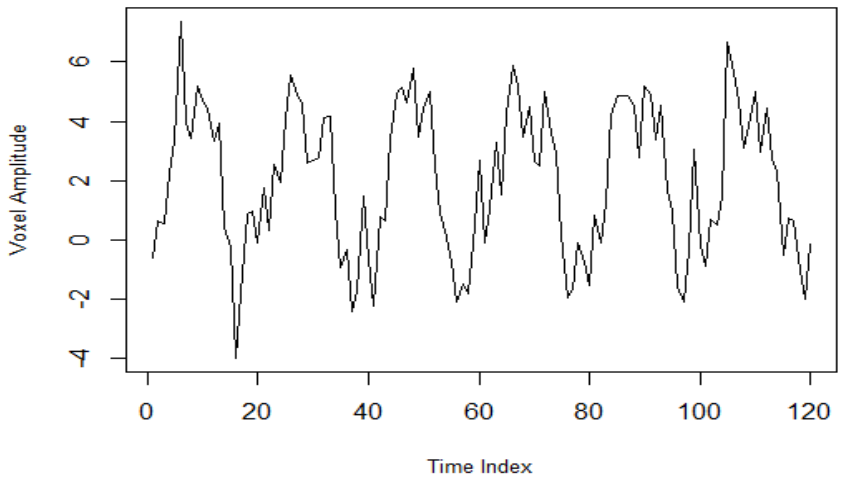
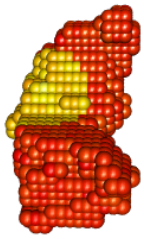


Figure 3: This shows the amplitude of a voxel within the embedded sphere that does respond to the periodic experimental task. Simulated with effect size = 5.

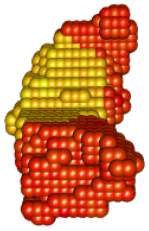
Below are images of the resulting mask, with the spheres indicated in yellow color:

(a) Radius = 1

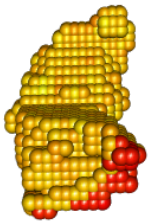
(b) Radius = 3



(c) Radius = 5



(d) Radius = 7



(e) Radius = 15

Figure 4: Simulation Volumes: Spheres of various sizes embedded in a mask of the right hippocampus.

We see from Figure 4 that, in the simulation volumes, the activated regions (pictured in yellow) do not form tunnel or ring-like shapes. In particular, if we regard the red regions of the simulation volumes pictured in Figure 4 as subsets of $\mathbb{R}^3$, the classical singular homology group H_1 is trivial. Consequently one expects to find that the persistent H_1 of these data sets consists of relatively low-persistence features. This expectation about the simulated data pictured in Figure 4 is borne out: see below, in Figure 6. (See [1] for an influential study of the sensitivity of low-persistence features in persistent homology to geometric

structure in a data set.)

Signal to noise ratio (SNR) is the magnitude of the signal over the magnitude of the noise. The SNR establishes the rough amplitude of noise only after the amplitude of the non-noise signal has already been established. An SNR of greater than one indicates that the signal has a greater magnitude than the noise. Our simulations included SNR values of 2, 5, 10, and 20.

Minimum persistence was also investigated at values of .8, 1, and 1.2. Recall from section 5.2 that, for fMRI data, we see .8 as a canonical choice to remove noise from the persistence diagrams.

The two normalization functions discussed earlier were also compared, with results explained in the Results section, below.

5.4 Results

Our method was evaluated on its ability to identify the task-based activation of embedded spheres of various radii. This was accomplished by calculating statistical power. Statistical power is the probability that a method rejects the null hypothesis when the alternative hypothesis is correct. In this case, the null hypothesis is that the persistence diagrams of observations during the "resting" phases of our simulated experiment are no different than the persistence diagrams of those during the "task" phases. Power was empirically estimated by first simulating each set of parameters 500 times and conducting the permutation test with 2000 permutations for each simulation. The proportion of tests that rejected the null hypothesis is then our empirical estimate for power. The figures at the end of this section summarize the empirical power estimates across sphere radius, minimum persistence threshold, and effect size.

The *lower* the minimum persistence threshold for features considered, the more powerful the method became (see Figure 4). This indicates that the incorporation of lower persistence features provides information useful for identifying the activity of the embedded sphere. Performance also improved when increasing maximum radius of homology computed from 3 to 4, informing us that more information improved results rather than overwhelming the method. Additionally, the canonical normalization scheme in Definition 4.1 outperformed the Definition 4.2 normalization scheme for constants 10 and 100, and performed comparably with constant 50. Thus, normalizing the signal to have a similar spread to the spatial coordinates performs better than either having a smaller variation in the signal or larger variation in the signal relative to the spatial coordinates.

For effect size 5 and above, our method displayed power > 0.85 for all radii except r=15 (see Figure 5). The sensitivity of our method to task-activated spheres as small as radius 1 without sub-setting the data is evidence that, even for more subtle patterns of activity, persistence diagrams record differentiating topological structure. The drop-off in power for radius=15 is likely because, as the embedded sphere at that radius made up most of the hippocampus-shaped data, it likely was not as detectable via one-dimensional homological features.

Perhaps including zero-dimensional homological features(which represent connected components) would improve sensitivity to larger clusters.

Our simulations demonstrate the efficacy of statistical inference using TDA to capture associations in task-based fMRI experiments.

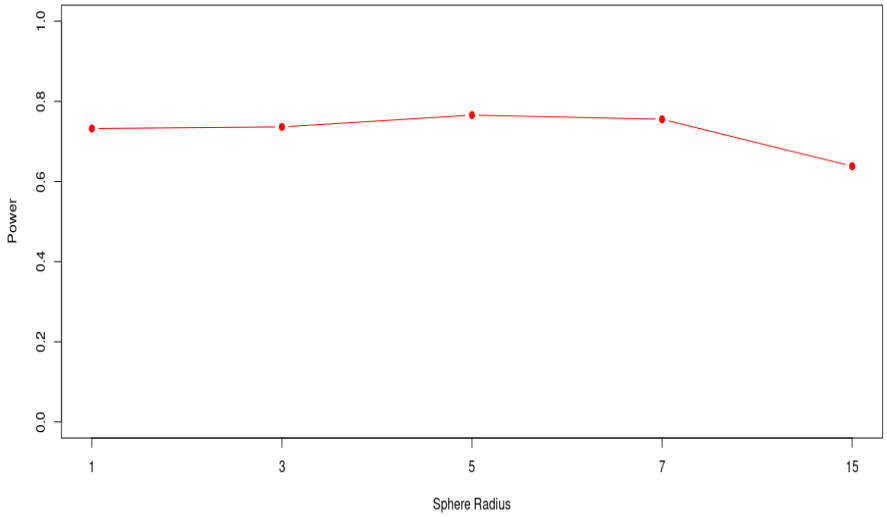


Figure 5: Empirical power estimates by radius of embedded sphere.

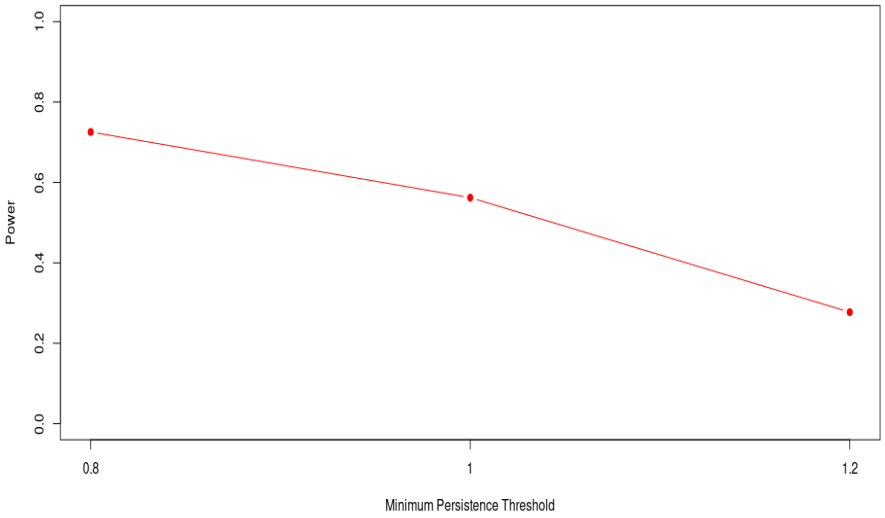


Figure 6: Empirical power estimates by minimum persistence threshold of homological features.

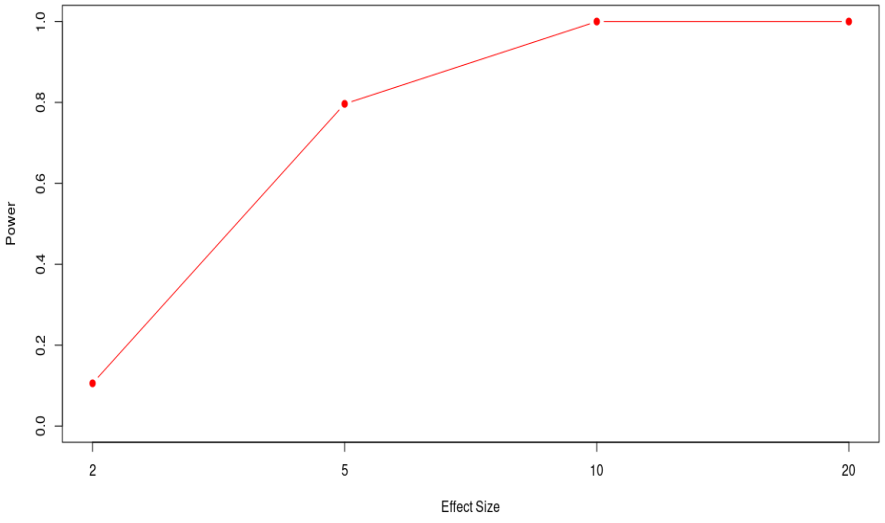


Figure 7: Empirical power estimates by effect size for embedded sphere's response to task.

References

[1] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. *Inverse Problems*, 36(2):025008, jan 2020.

[2] Christopher Cericola, Inga Jo Johnson, Joshua Kiers, Mitchell Krock, Jordan Purdy, and Johanna Torrence. Extending hypothesis testing with persistent homology to three or more groups. *Involve*, 11(1):27–51, 2018.

[3] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16(110):3603–3635, 2015.

[4] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. 10 2017.

[5] Herbert Edelsbrunner and John L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. An introduction.

[6] Anders Eklund, Thomas Nichols, and Hans Knutsson. Can parametric statistical methods be trusted for fMRI based group studies? *arXiv preprint arXiv:1511.01863*, 11 2015.

[7] Lohmann G, Stelzer J, Lacosse E, Kumar VJ, Mueller K, Kuehn E, Grodd W, and Scheffler K. LISA improves statistical analysis for fMRI. *Nature Communications*, 10 2018.

[8] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[9] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, 2:83–97, 1955.

[10] Bradley Macintosh, Richard Mraz, William McIlroy, and Simon Graham. Brain activity during a motor learning task: An fMRI and skin conductance study. *Human brain mapping*, 28:1359–67, 12 2007.

[11] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017.

[12] Jose A Perea. A brief history of persistence. *Morfismos*, 23(1):1–16, 2019.

[13] Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, 10 2013.

[14] Andrew Salch, Adam Regalski, Hassan Abdallah, Raviteja Suryadevara, Michael J. Catanzaro, and Vaibhav A. Diwadkar. From mathematics to medicine: a practical primer on topological data analysis (TDA) and the development of related analytic tools for the functional discovery of latent structure in fMRI data, 2021. accepted to PLoS ONE.

[15] Larry Wasserman. Topological data analysis. *Annu. Rev. Stat. Appl.*, 5:501–535, 2018.

[16] Marijke Welvaert, Joke Durnez, Beatrijs Moerkerke, Geert Verdoolaege, and Yves Rosseel. neuRosim: An R package for generating fMRI data. *Journal of Statistical Software*, 44(10):1–18, 2011.

[17] Anderson Winkler, Matthew Webster, Diego Vidaurre, Thomas Nichols, and Stephen Smith. Multi-level block permutation. *NeuroImage*, 62, 06 2015.