Statistical Inference for Persistent Homology applied to fMRI

Hassan Abdallah[1], Adam Regalski[2], Mohammad Behzad Kang[2], Maria Berishaj[2], Nkechi Nnadi[2], Asadur Chowdury[3], Vaibhav A. Diwadkar[3], Andrew Salch[2]

[1]Dept. of Biostatistics, University of Michigan, Ann Arbor, MI
[2]Dept. of Mathematics, Wayne State University, Detroit, MI
[3]Dept. of Psychiatry & Behavioral Neuroscience, Wayne State University, Detroit, MI

October 2020

1 Introduction

Functional magnetic resonance imaging (fMRI) is a tool that provides a rich avenue to study brain activity via haemodynamic response. Making sense of the complex spatio-temporal relationships in fMRI data can give insight into the functional and structural organization of the brain. A common goal in experiments with fMRI involves attempting to establish associations between the fMRI signal and a given task. Statistical methods play a pivotal role in identifying and evaluating the validity of such associations. Statistical methods typically fall into two categories: parametric and non-parametric. Parametric methods make assumptions about the distribution of the test statistic and the error terms of a model and, when assumptions are met, provide a powerful tool for analyzing data. When assumptions are not met, however, they can suffer from inflated false-positive and family-wise error rates [5]. Non-parametric methods, on the other hand, make more limited assumptions and use strategies such as permutation testing to conduct an analysis.

Statistical methods for fMRI also differentiate themselves by the type of association they evaluate and the manner in which they incorporate spatial information. The general linear model(GLM), a predominant tool in this area, mostly evaluates associations on the basis that the relationship between the signal and covariates is linear. Spatial information can be encoded by summarizing sub-regions(such as spheres) by its average or maximum signal value and performing a cluster-level analysis. Some techniques combine the GLM with more sophisticated approaches to encoding spatial information [6]. Even so, there is room for improvement in terms of techniques that can characterize spatial relationships in fMRI data and leverage that information to identify associations, especially associations that are not necessarily linear. In this vein, topological data analysis has emerged as a viable option for exploring associations related to the topological or geometric characteristics of fMRI.

Topological data analysis(TDA) is a technique that leverages topology to uncover higher-dimensional geometric patterns in data via methods such as persistent homology. Persistent homology provides a quantified summary of the global topological profile of data given some distance-metric in the form of a persistence diagram. Persistence diagrams contain information related to the presence of topological features such as loops and geometric information such as curvature [1]. In recent years, frameworks have been developed to harness such information for statistical inference, in particular by Robinson and Turner [8].

In this paper, we give a method for evaluating whether fMRI data exhibits differentiating topological structure across different labels, such as those in a task-based experiment. We develop an extension of existing methods for statistical inference using TDA that accommodates the auto-correlative structure in fMRI data. The method is further explored to determine optimal conditions for performance considering topological nuances unique to fMRI data, including the presentation of a spatial-normalization process distinctive to this setting. Empirical power estimates are then computed via simulated fMRI data and various combinations of parameters and simulation characteristics. The R package, NeuroTDA, is provided on GitHub.

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

2 Background Knowledge on Topological Data Analysis (TDA)

Topological data analysis involves computations of *homology* and *persistent homology*, which we will discuss later on in this section. One should note that calculating persistent homology occurs over a field. In our work, and throughout this discussion, the appropriate field here is $\mathbb{Z}_2 = \{0, 1\}$, the group of integers modulo 2 equipped with the addition operation, where the set of even integers is represented by 0 and the set of odd integers is represented by 1.

We begin with a set of points $v_0, v_1, ..., v_k$ in $\mathbb{R}^k$ such that $v_1 - v_0, v_2 - v_0, ..., v_k - v_0$ are linearly independent. Taking the convex hull $[v_0, v_1, ..., v_k]$ of this set, we form its *k-simplex.* A *face* of the k-simplex is then the convex hull, as described before, of a proper subset of $\{v_0, v_1, ..., v_k\}$. Next, consider a countable set K of simplices such that, for each simplex in K, each face of the simplex is contained in K, and the intersection of two simplices in K is either a face of both simplices or empty. The set K is known as a *simplicial complex.*

Let K be finite, and consider the vector space $V_k(K)$ of formal linear combinations with coefficients in $\mathbb{Z}_2$ 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. These boundary maps lead to the *chain complex* $\cdots \xrightarrow{\delta_{k+1}} V_k(K) \xrightarrow{\delta_k} V_{k-1}(K) \xrightarrow{\delta_{k-1}} V_{k-2}(K) \xrightarrow{\delta_{k-2}} \cdots \xrightarrow{\delta_3} V_2(K) \xrightarrow{\delta_2} V_1(K) \xrightarrow{\delta_1} V_0(K) \xrightarrow{\delta_0} 0$. In particular, the boundary maps satisfy $\delta_{k+1} \circ \delta_k = 0$. Thus, with the kernel $Ker(\delta_k)$ of δ_k known as the set of *cycles*, and the image $Im(\delta_{k+1})$ of δ_{k+1} called the set of *boundaries*, we observe that $Im(\delta_{k+1})$ is a subset (and, specifically, a normal subgroup) of $Ker(\delta_k)$, and we may form the quotient group $H_k(K) := Ker(\delta_k)/Im(\delta_{k+1})$. This is known as the k^{th} *homology group* of K. This construction gives the k-cycles of K up to a boundary of a $(k+1)$-chain. Such an equivalence defines the different homology classes in the quotient group.

Now, consider the family $\{K_a : a \in \mathbb{R}\}$ of simplicial complexes that occur as sub-complexes of K such that $K_m \subseteq K_n$ whenever $m \leq n$. This is known as a *filtered simplicial complex.* In this paper, the filtrations we apply are always Vietoris-Rips filtrations. For $m \leq n$, denoting the boundary maps on $V_k(K_m)$ and $V_k(K_n)$ by δ_k^m and δ_k^n, respectively, we naturally have inclusion maps $\iota : K_m \longrightarrow K_n$, which, in turn, gives inclusion maps $\iota : Im(\delta_{k+1}^m) \longrightarrow Im(\delta_{k+1}^n)$ and $\iota : Ker(\delta_k^m) \longrightarrow Ker(\delta_k^n)$. The image of the homomorphism $\iota_k^{m,n} : H_k(K_m) \longrightarrow H_k(K_n)$ induced from these inclusion maps, then, is one of the k^{th} *persistent homology groups* of K. In particular, one achieves a sequence of homomorphisms of successive k^{th} homology groups, i.e., the k^{th} *persistence module.*

We are now prepared to define the *persistence diagram.* The k^{th} *persistence diagram* of the k^{th} persistence module is a multiset of points in $(\mathbb{R} \cup \{\infty\})^2$. Each point in the diagram represents a homology class and corresponds to a particular topological feature; the x-coordinate of this point, known as the *birth radius*, represents the time at which the topological feature was born, and the y-coordinate, known as the *death radius*, represents the death time of the feature. Explicitly, in the relevant persistence module, these radii are given by the indices at which a feature appears and disappears. By convention, we include (with infinite multiplicity) all points such that $x = y$ (that is, the points lying along the diagonal) - points along the diagonal represent homology classes that are not persistent for any amount of time. The further a point is from the diagonal of a persistence diagram, the longer the homology class persisted. Points close to the diagonal, therefore, are considered topological noise.

For a more comprehensive introduction to topological data analysis, we refer the reader to [4].

3 Hypothesis Testing for Topological Data Analysis

The purpose of this section is to discuss (and extend via [2]) the methods used in [8] which we will apply to our fMRI data.

Methods Discussed by Robinson & Turner

The goal of Robinson and Turner's work is to use hypothesis testing to compare two groups of persistence diagrams. The null hypothesis is that, if one splits the set of persistence diagrams into two groups using a chosen labeling of each diagram, then the difference observed between the sets of diagrams based on this labeling would be just as likely as the differences observed between sets of diagrams with the labels chosen randomly. 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. The hypothesis testing is carried out by obtaining a p-value that measures the strength of evidence against the null hypothesis compared to a chosen significance level. To clarify, all methods discussed in this subsection are due to Robinson and Turner.

We've stated that the null hypothesis pertains to differences between distributions of persistence diagrams determined by the original data. Recall that our original fMRI data is gathered by sampling m points from a subset of $\mathbb{R}^4$ with some noise. In this particular case, the initial chosen labeling on our set of persistence diagrams would split our set of persistence diagrams into two groups, each representing a particular process and coming from a subset of $\mathbb{R}^4$ with some noise. Drawing from Robinson and Turner's discussion in Section 2.3 of [8], then, if we have two subsets of $\mathbb{R}^4$ from which we sample m points with some noise, and we wish to know whether the two subsets are different, it is legitimate to examine this difference by studying the respective distributions of persistence diagrams determined by the underlying topology of each point cloud.

Thus, 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. Not only that, but the randomization test that we may apply omits the need to hypothesize a distribution model from which persistence diagrams are drawn under the null hypothesis, as mentioned by Robinson and Turner.

Let's now consider 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 will 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 ϕ is a bijection 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. What does one mean here by an optimal bijection? One can compute a cost given by assigning the element $x \in X$ to $y \in Y$ given by the square of the L^2-norm, or the square of the Euclidean distance, $||x-y||_2^2$ between x and y. From this set of costs, we consider the bijection $\phi : X \longrightarrow Y$ that minimizes the total cost $\sum\limits_{x \in X} ||x-\phi(x)||^2$ – this is the optimal bijection. This bijection is 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, section III of [**?**] discusses how to apply the Hungarian algorithm to A to find the optimal bijection between elements of S and elements of T. How does one construct the appropriate matrix to carry out this Hungarian algorithm with two persistence diagrams X and Y ? 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.

We begin the hypothesis test with a set of n persistence diagrams divided into two groups $\beta_1 = \{X_{1,1}, X_{1,2}, ..., X_{1,n_1}\}$ and $\beta_2 = \{X_{2,1}, X_{2,2}, ..., X_{2,n_2}\}$ containing n_1 diagrams and n_2 diagrams respectively. Again, this division is done according to a systematic labeling scheme that we initially choose. A necessary assumption to apply the methods that we'll discuss here (and, the analogous methods in the next subsection that we're really interested in, generalizing this work to multiple groups of persistence diagrams), is independence of our observations (respectively, persistence diagrams). Since the fMRI data that we analyze is *non-independent* time series data, in the next section, we apply

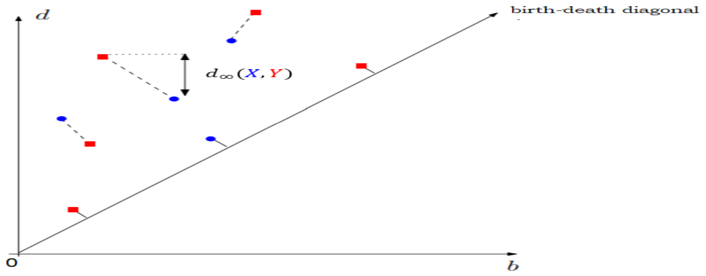

Figure 1: A visual representation of bottleneck distance $d_\infty(X, Y)$ between two persistence diagrams X and Y. The blue points represents points in persistence diagram X, while red points represents points in persistence diagram Y. A line drawn from a point $x \in X$ (respectively, $\phi(x) \in Y$) 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].

a multi-level block sampling framework to satisfy the exchangeability criteria for our permutation test, thereby removing the requirement of independence. The null hypothesis, recall, is that the initial choice of labeling is no less likely than labeling obtained by randomly permuting the labels, i.e., that the two processes determining the groups β_1 and β_2 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.

The test statistic is the joint loss function given by

$$F_{p,q}(\{X_{1,i}\}, \{X_{2,i}\}) := \sum_{m=1}^{2} \frac{1}{2n_m(n_m - 1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} d_p(X_{m,i}, X_{m,j})^q,$$

where $p \in [1, \infty)$, $q \in [1, \infty)$. We take p to be infinity and q to be 1. Since the groups β_1 and β_2 are determined by a choice of labeling L, we will use the notation $F(L)$ to mean the joint loss function computed on the two groups of persistence diagrams determined by L, and $F(L_{observed})$ to mean the joint loss function computed on the two groups of diagrams determined by the initial choice of labeling. This test statistic is a generalization of the test statistic

$$\sigma^2_{\beta_{12}}(L) = \sum_{m=1}^{2} \frac{1}{2n_m(n_m - 1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} d_2(X_{m,i}, X_{m,j})^2$$

The statistic $\sigma^2_{\beta_{12}}(L)$, the joint loss of a particular labeling scheme, is given by the sum of variances within the groups of persistence diagrams determined by L, and such that 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[1]. This is because the latter consideration is very computationally expensive.

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, Robinson and Turner show that we can take α to be a true p-value by Lemma 1 of [8]. Robinson and Turner do so with the assumption that the persistence diagrams are drawn independently of each other and are identically distributed. As we mentioned before, our time series data is not independent, and our observations occurring as persistence diagrams are hence not independent (which remains true when we generalize these methods in the next subsection to multiple groups of persistence diagrams). However, again, we're able to correct for this using our methods in the next section.

Robinson and Turner develop an algorithm to compute the proportion α described above to be taken as the p-value. However, one cannot take that exact proportion as the true p-value; the algorithm could obtain a proportion of 0, and by [**?**], a p-value of 0 is not permissible for permutation tests. Thus, they modify the proportion by adding 1 to its numerator and denominator, and develop an algorithm to compute this instead to be taken as a p-value. They show that this is indeed a true p-value by Theorem 1 of [8]. This updated algorithm, taken from [8], that we apply is shown below, where the output Z is the p-value.

Robinson & Turner's Methods Applied to Multiple Groups of Persistence Diagrams

Via [2], we now extend the methods discussed in the last subsection to use hypothesis testing to compare multiple groups of persistence diagrams, which is what we will be truly interested in moving forward for applying to our fMRI data. The need for us to extend the comparisons between persistence diagrams discussed in the last subsection to 3 or more groups of persistence diagrams comes from the multi-level block sampling framework that we apply to our fMRI 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. Thus, our null hypothesis is that, if one splits a set of persistence diagrams into multiple groups using a chosen labeling of each diagram, then the

[1] There is not a clear notion of the mean of a set of persistence diagrams.

difference observed between the sets of diagrams based on this labeling would be just as likely as the differences observed between sets of diagrams with the labels chosen randomly. More generally, the null hypothesis is that the multiple processes determining the original labeling scheme one has chosen (and, hence, the initial split into multiple groups of persistence diagrams) are not significantly different. The alternative hypothesis, then, is that at least two of the processes determining the original labeling scheme are significantly different. We obtain a p-value measuring the strength of evidence against the null hypothesis compared to a chosen significance level to carry out our hypothesis test. We may assess the strength of evidence against the claim that the multiple processes are the same by studying the distributions of persistence diagrams from each process, without the need to hypothesize a distribution model from which persistence diagrams are drawn under the null hypothesis.

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 a systematic labeling scheme that we initially choose. Again, we forgo the necessary assumption that observations (respectively, persistence diagrams) are independent in the next section by applying a multi-level block sampling framework to satisfy the exchangeability criteria for our permutation test. The null hypothesis, recall, is that the initial choice of labeling is no less likely than labeling obtained by randomly permuting the labels, i.e., that the multiple processes determining the groups β_1, β_2, ... , β_s, 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.

The test statistic is the joint loss function is 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 $p \in [1, \infty)$, $q \in [1, \infty)$. 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. This test statistic is a generalization of the test statistic

$$\sigma^2_{\beta_{123\cdots s}}(L) = \sum_{m=1}^{s} \frac{1}{2n_m(n_m-1)} \sum_{i=1}^{n_m} \sum_{j=1}^{n_m} d_2(X_{m,i}, X_{m,j})^2$$

The statistic $\sigma^2_{\beta_{123\cdots s}}(L)$, the joint loss of a particular labeling scheme, is given by the sum of variances within the s groups of persistence diagrams determined by L, and such that pairwise distances between persistence diagrams are only computed once and stored in a table.

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, analogous to the last subsection, we can take α to be a true p-value, again forgoing the assumption that persistence diagrams are drawn independently of each other and are identically distributed by our multi-level block framework described in the next section.

Referring to the algorithm mentioned in the last subsection, our algorithm to compute the proportion α described in the last paragraph as our true p-value is largely the same. 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.

4 TDA Hypothesis Testing for fMRI

In order to make this statistical method viable for fMRI data, there are assumptions and parameters that need to be considered and optimized. Firstly, the independence assumption of persistence diagrams must be loosened to accommodate the dependence structure present in successive fMRI scans. Secondly, a suitable normalization technique for the BOLD signal needs to be identified such that 4-dimensional point clouds born from fMRI data are organized in a way that persistent homology is most sensitive to evolving topological structure. Finally, parameters related to persistence, such as maximum radius to compute persistence and cutoffs for persistence diagram feature selection, need to be explored in the context of fMRI data.

Exchangeability

The original hypothesis test described in Robinson Turner assumed that all the processes generating the persistence diagrams are independent and therefore retain *global exchangeablity*, meaning labels may be exchanged freely during any permutation test. In fMRI data, successive scans are not independent because of temporal autocorrelation in the BOLD signal. Freely exchanging labels in our permutation test while the independence assumption is violated makes the test vulnerable to a larger rate of Type I error [9]. For instance, in the presence of temporal autocorrelation, an observed test statistic in which the initial labelling scheme included mostly contiguous sets of labels may be more extreme than a simulated test statistic because the randomized set of labels may not exhibit similarly contiguous structure, rather than as a result of a true departure from the null hypothesis. To remedy this, a multi-level block shuffling scheme is implemented as in [9].

Instead of exchanging one datum 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 temporal dependence and experiment design.

For example, consider an fMRI experimental design that consists of 200 total scans. Suppose a stimulus or task is administered every 20 scans and lasts for 20 scans. Our first-level blocks in this scenario would be defined as each contiguous set of 20 scans starting at 1, accounting for 10 first-level blocks. These blocks would follow whole-block exchangeability, meaning they are shuffled as a whole when labels are permuted. Possible permutation of labels would include the set labels for scans 1-20 being switched with the set of labels for scans 181-200. 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 1-100 and the other is the data associated with 101-200. A distinguishing pattern across early versus late stages of fMRI experiments have been noted ([7] will need better cite here) so the test statistic computed for this set of labels could display an extremeness resulting from this temporal phenomena. As such, second-level exchangeability blocks are necessary.

In this example, a second-level exchangeability blocks structure would consist of two blocks, one covering scans 1-100 and another covering scans 101-200. Unlike the first-level EBS', these are not whole-block exchangeable and are instead within-block exchangeable. As a result, the first-level blocks present in labels 1-100 could only be exchanged amongst themselves and not with the first-level blocks in labels 101-200. 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.

The above described approach to constructing a multi-level block shuffling scheme for an fMRI experiment can be extended by adjusting first-level block sizes to whatever the experimental epoch sizes are for a given experiment.

Normalization

The structure of fMRI data consists of three spatial coordinates and a signal amplitude (with the additional component of time). 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 would rescale the three spatial dimensions but not the fourth (signal amplitude), changing the topological features and persistence diagrams acquired. Therefore, the necessity of a normalization technique arises in order to yield results that are invariant under changing units, specifically normalization applied to the fourth dimension that is signal amplitude. Recall, topological features are discovered through the persistence diagrams which arise from the filtrations of Vietoris-Rips complexes on the point-cloud data. Therefore, the scales in the different coordinates impact when and how the ϵ-balls will overlap and display features. Thus topological structure is impacted by how we normalize the data. Below, two methods of normalizing the signal amplitude are discussed.

Definition 4.1 (Normalization Scheme 1)**.** Define the following notation:

$$\begin{aligned}
S_{min} &= min\{\{min\{\text{x-coordinates}\}, min\{\text{y-coordinates}\}, min\{\text{z-coordinates}\}\} \\
S_{max} &= max\{\{max\{\text{x-coordinates}\}, max\{\text{y-coordinates}\}, max\{\text{z-coordinates}\}\} \\
A_{min} &= min\{\text{signal amplitude}\} \\
A_{max} &= max\{\text{signal amplitude}\}
\end{aligned}$$

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

$$\begin{aligned}
S_{min} &= \frac{min\{\text{x-coordinates}\} + min\{\text{y-coordinates}\} + min\{\text{z-coordinates}\}}{3} \\
S_{max} &= \frac{max\{\text{x-coordinates}\} + max\{\text{y-coordinates}\} + max\{\text{z-coordinates}\}}{3} \\
A_{min} &= min\{\text{signal amplitude}\} \\
A_{max} &= max\{\text{signal amplitude}\}
\end{aligned}$$

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 (i.e. the bold score) with
$\left[\frac{\epsilon - A_{min}}{A_{max} - A_{min}} \cdot (S_{max} - S_{min})\right] + S_{min}$

Other normalization techniques consist of scaling Normalization Scheme 1 by a factor of 10, 50, and 100.

Normalization Scheme 2 is preferred due to the BOLD scores 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: Let $x \in (35, 57), y \in (65, 90), z \in (32, 63)$. To demonstrate the preference of an average, take the amplitude ϵ in the range $(32, 90)$ determined by finding the minimum and maximum values taken among the x, y, and z coordinates. In comparison, take the amplitude ϵ determined by the averages of the x, y, and z coordinates, yielding the range $(44, 70)$. The reason an average is preferred is due to the sizes of the ranges. Note the size of the ranges for the x, y, and z coordinates are 22, 25, and 31 (respectively). The range for the amplitude under the first method is 58 whereas the range for the amplitude under the second method is 26 which is closer to the ranges of the spatial coordinates. The properties of the BOLD score mirror the properties of the spatial coordinates.

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 radius in which to compute homology. The ideal choice for this parameter is half the distance of the two farthest points in a data set, since there are no nontrivial changes in the topology of the space beyond that radius(the topology is that of a single connected component). In most fMRI data sets, computing persistent homology up to that distance is not computationally feasible. Instead, a threshold is chosen that completes in a reasonable amount of time and has the potential to capture nontrivial topology. For example, using maximum radius 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, rather 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 note that making that tweaking this choice of parameter does not artificially alter results since as this parameter is changes, so does the hypothesis and statistical conclusion of our test. For example, rejecting the null hypothesis would indicate that 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 (0.707). It is then reasonable to infer that features at that persistence and below are likely more related to "topological" noise rather than topologically-interesting phenomena. 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.

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 to simulate the data.

Corresponding BibTeX entry:

@Article, title = neuRosim: An R Package for Generating fMRI Data, author = Marijke Welvaert and Joke Durnez and Beatrijs Moerkerke and Geert Verdoolaege and Yves Rosseel, journal = Journal of Statistical Software, year = 2011, volume = 44, number = 10, pages = 1–18, url = http://www.jstatsoft.org/v44/i10/,

Experimental Design

Over six epochs of 20 scans with TR=2 the experiment simulates a task that at the onset of each epoch a sphere of increased activation is embedded into a hippocampus of physiological noise. Activation is greatest at the beginning of each epoch and fades throughout.

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.

The radius of the sphere, measured in voxel edge lengths, takes on the possible values 1,3,5,7,and 15. Any points of the sphere that fall outside of the hippo-campus are removed. The initial effect size (which can be thought of as the magnitude of activation) of the activation in the sphere takes on the possible values 2,5,10, and 20. Effect size is a measure of the magnitude of the signal. These are the onset effect sizes and they decrease to zero as the epoch reaches its end.

Signal to noise ratio 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 the section on persistence thresholds that for fMRI data we see .8 as a canonical choice to remove noise from the persistence diagrams.

We found in practice by multiplying a normalized bold score by a scalar that topological features of fMRI data will reveal themselves in persistent homology at different scalars.

The two normalization functions discussed earlier were also compared.

5 Results

Each set of parameters was simulated 500 times and the permutation test was conducted with 2000 permutations each. For effect size 5 and above, our method displayed power ¿ 0.90 for all radii except r=15, which performed at 0.66. 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.

Furthermore, the *lower* the minimum persistence threshold for features considered, the more powerful the method became. 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.

Finally, the best performing normalization technique was the one that transformed the signal into a range that made it most similar to the range for the spatial coordinates. Overall, our simulations demonstrate the efficacy of statistical inference using TDA to capture associations in task-based fMRI experiments.

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] 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.

[6] Lacosse E Kumar VJ Mueller K Kuehn E Grodd W Scheffler K. Lohmann G, Stelzer J. Lisa improves statistical analysis for fmri. *Nature Communications*, 10 2018.

[7] 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.

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

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