Multiple Hypothesis Testing with Persistent Homology

Mikael Vejdemo-Johansson
Department of Mathematics, CUNY College of Staten Island
Computer Science, CUNY Graduate Center
mvj@math.csi.cuny.edu

Sayan Mukherjee
Department of Statistics
Duke University
sayan@stat.duke.edu

Abstract

Multiple hypothesis testing requires a control procedure: the error probabilities in statistical testing compound when several tests are performed for the same conclusion. A common type of multiple hypothesis testing error rates is the Family-Wise Error Rate (FWER) which measures the probability that any one of the performed tests rejects its null hypothesis erroneously. These are often controlled using Bonferroni's method or later more sophisticated approaches all of which involve replacing the test level α with α/k, reducing it by a factor of the number of simultaneous tests performed. Common paradigms for hypothesis testing in persistent homology are often based on permutation testing, however increasing the number of permutations to meet a Bonferroni-style threshold can be prohibitively expensive. In this paper we propose a null model based approach to testing for acyclicity (ie trivial homology), coupled with a Family-Wise Error Rate (FWER) control method that does not suffer from these computational costs.

1 Introduction

Hypothesis testing in the based on topological summaries of data has been an area of Topological Data Analysis (TDA) that has seen growth recently as both applied and mathematical statistics have been developed using TDA. Almost of all the current literature on hypothesis testing in TDA has focused on two sample tests [16] or extensions to analysis of variance (ANOVA) settings [5] where differences across more than two conditions are considered. Both of these papers work on constructing a statistic from in-group and out-group distances and comparing this statistic for the observed diagram collections to collections of label permutations. Neither of these papers take into account multiple testing because the number of hypotheses tested is small, for example one in two sample tests. However, as the number of groups in an ANOVA increase mutiple testing is a concern, in addition there are many applications where TDA can be applied to many subsets of features of coordinates in a two sample test with the goal of finding those subsets which are significantly different between the two groups. When the number of subsets of features are in the hundreds or thousands correction for multiple hypothesis testing is crucial.

Additionally to the two sample and ANOVA style tests, work has also been done in the field on one sample testing through creating confidence regions using stability of persistence combined with bootstrap sampling of the underlying point cloud to derive bounds for bottleneck distances [8].

In all of these setups, the noise model is taken to be intrinsic to and represented by the diagrams themselves. This approach is too flexible for currently known limit theorems [11], which makes the construction of efficient Family-Wise Error Rate (FWER) control procedures with our approach difficult.

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In this paper, we propose a one sample hypothesis test for persistence diagrams that incorporates an explicit choice of noise model and a Family-Wise Error Rate control procedure compatible with the structure of the test.

Noise model: Our main candidate for a noise model first draws a sample from an stationary and ergodic point process, conditioned to be similar to the observed point cloud. Next, persistent homology using some construction of a filtered simplicial complex – such as Čech, Vietoris-Rips or alpha-complexes – is computed, together with some interesting statistic of the resulting persistence diagram.

Hypothesis test: Our hypothesis test using a noise model is essentially a simulation test. $N - 1$ noise samples are drawn, conditioned to be similar to the point cloud to be tested. For the resulting N point clouds – 1 observed and $N - 1$ simulated – the same pipeline with persistent homology and an interesting statistic of the persistence diagram is applied. The resulting statistic values are ranked, and if the value corresponding to the observed value is sufficiently extreme, the test rejects a null hypothesis of *the observed point cloud could have been produced by the noise model.*

FWER control: We assume that the noise model used has a limit theorem similar to the ones provided in [11]. The limit theorem of interest to us says that up to a scaling factor, persistence diagrams of filtered simplicial complexes generated from "square" sampling windows from a stationary and ergodic point process converge to a distribution on the persistence diagram half plane that is determined only by the underlying point process.

Since the persistence diagrams follow almost the same distribution, the statistics calculated from them will also be very similar in distribution. **Hence, up to rescaling, the values extracted from these diagrams will be comparable**.

By calculating z-scores, the appropriate scaling constant can be derived from the statistic values directly. These considerations lead to our suggested FWER controlled procedure for multiple one sample hypothesis tests with a specified noise model:

For each of K observed diagrams, draw $N - 1$ noise model samples. Compute filtered complexes, persistence diagrams and statistics for all the $K \cdot N$ point clouds. We can organize all these statistic values in a $K \times N$ matrix with one row for each observation and its associated simulations. Use the $N - 1$ simulated values in each row to estimate mean and standard deviation of the statistic corresponding to that specific observed diagram and its simulated repetitions. Compute z-scores for each statistic value using the row-wise means and standard deviations. At this point we can treat each column as a separate simulation run for the entire collection of diagrams, and pick the column wise most extreme statistic. This produces a total N extreme values – rank them and check if the rank assigned to the most extreme statistic from the observed point clouds is sufficiently extreme to reject a null hypothesis of *all of the observed point clouds could have been generated by the noise model*.

The idea that topological summaries such as persistence diagrams form a probability space for which formal statistical analysis is well defined was developed in [14]. Further developments on defining useful summary statistics within persistent homology and considering means, medians, and variances of persistence diagrams was pursued in several papers [15, 17, 19]. The main challenge in considering persistence diagrams as a probability space was pointed out in [17, 18]—the space of persistence diagrams is positively curved which results in non-unique geodesics. As a result the mean of a set of diagrams need not be unique which complicates data analysis. To avoid this issue persistence landscapes were introduced in [4], persistence landscapes are functions so they can be considered as random functions in a Banach space, a construction that admits central limit theorems, unique means and medians. Further examinination of bootstrap properties of persistence based summaries as well as a notion of confidence intervals for points in a diagram was developed in [8, 10]. An alternative approach was considered in a series of papers where instead of considering a persistence diagram as a summary a probability density was used as a topological summary, an approach called distance to measure [6, 7, 8]

In the context of hypothesis testing [2] proposed using goodness of fit statistics – Kolmogorov-Smirnov, χ^2 or Mann-Whitney – to test compare empirical distributions from two samples of persistence diagrams. The ideas most closely related to the procedures we develop in this paper was to define hypothesis testing procedures directly on persistence diagrams using permutation testing and barcode distances [5, 16]. In this paper we will extend two sample single hypothesis testing and ANOVA procedures to the multiple hypothesis test setting.

2 Noise model and one-sample hypothesis test

We construct a one-sample hypothesis test that explicitly encodes a noise model in the test. Our test will test the null hypothesis *the observed point cloud could have been produced by the noise model* against an alternative of *there are homological obstructions to consistency between the observation and the noise model*. Our construction works with a wide range of possible noise models – however, for the test to fit into the FWER control procedures that we introduce in Section 2.1, we rely on limit theorems to ensure that the statistics used are comparable between different hypothesis tests.

Definition We take *noise model* to refer to a method that given a point cloud generates new point clouds.

Ideally, the point clouds generated by a specific noise model will share some properties with the input point cloud, and share some properties with our expectations of a lack of homological structure.

We propose as one widely applicable noise model to use uniformly distributed points in a bounding shape for the observed point cloud. Following conventions in spatial statistics, we suggest to condition the noise model sample on the size of the observed point cloud. This idea produces two concrete noise models:

Uniform bounding box noise model Given point cloud X

1. Estimate a bounding box of X
2. Sample $|X|$ points uniformly in this bounding box

Given points $x_1, \dots, x_n \in \mathbb{R}^d$, we can produce a uniformly minimum variance unbiased estimator of the bounding box by estimating a bounding interval $[\hat{a}_i, \hat{b}_i]$ for each coordinate separately. Writing $\min_i$ for the smallest value of the ith coordinate and $\max_i$ for the corresponding largest value, this estimator is given by

$$\hat{a}_i = \frac{N+1}{N}(\text{min}_i - \text{max}_i) \qquad \hat{b}_i = \frac{N+1}{N}(\text{max}_i - \text{min}_i),$$

With a noise model in place, simulation testing produces a hypothesis test as specified. For the simulation test we assume that some statistic $\gamma(\mathcal{D})$ of a persistence diagram $\mathcal{D} = \{(b_1, d_1), \dots\}$ has been chosen. One example of such statistics is $\max_i(d_i - b_i)$, but there are plenty other possible choices of useful statistics that can be used.

One-sample persistent homology simulation test Given a point cloud X, an integer N

1. Generate noise model samples $M_2, \dots, M_N$
2. Calculate persistence diagrams $\mathcal{D}_1$ of X and $\mathcal{D}_j$ of each M_j
3. Sort the diagram statistics $\gamma(\mathcal{D}_1), \dots, \gamma(\mathcal{D}_N)$ and reject the null hypothesis if the rank r of $\gamma(\mathcal{D}_1)$ is sufficiently extreme

We may then reject the null hypothesis at a level of $p = (N - r + 1)/N$.

2.1 Family-wise error rates

When using either Bonferroni, Holm or Hochberg's FWER control procedures [3, 12, 13] in a permutation or simulation setting, the number m of simultaneous tests can drive up the number of permutations or simulations required for an acceptable test level dramatically. If computations are expensive – such as with persistence diagrams or with bottleneck distances – then this quickly becomes prohibitive.

Most interesting persistence statistics vary with the overall scale of the point cloud; different point clouds produce statistics that usually are not immediately comparable. If they were, however, we could detect a deviation from the null model behaviour through the existence of a particularly large value for corresponding barcode-based statistics. We can produce a joint test by first making the statistics comparable, and then performing a simulation test where in each simulation step the largest statistic value across simulated representatives for all the point clouds is extracted.

The approach is rooted in the observation that, having computed test statistics $t_1, \ldots, t_m$ from each test separately, the probability of any one of the test statistics exceeding a threshold c is equal to the probability of the maximum among them exceeding that threshold:

$$\alpha_{\mathrm{FWER}} = \mathbb{P}(\{t_1 > c\} \cup \cdots \cup \{t_m > c\} | H_0) = \mathbb{P}(\max_i t_i > c | H_0)$$

We don't need distributional assumptions as long as the null hypothesis sampling distributions are comparable across all test cases.

Based on this we propose the following approach

Family-wise error rate controlled test for acyclicity Given a family of point clouds $X_1, \ldots, X_K$; a method for constructing filtered simplicial complexes and calculating persistence diagrams from them; an invariant $\gamma : \{\text{Point clouds}\} \to \mathbb{R}$; and a null model $\mathcal{M}$ of random point clouds, we may reject the null hypothesis of acyclicity in favor of non-acyclicity by:

1. Draw $M_1^2, \ldots, M_K^N$ from $\mathcal{M}$.
2. Calculate persistence diagrams $\mathcal{D}_j^1$ from X_j and $\mathcal{D}_j^i$ from M_j^i.
3. Compute all $\tilde{y}_i^j = \gamma(\mathcal{D}_i^j)$.
4. For each $i \in [1, K]$ and $j \in [2, N]$, use $\tilde{y}_i^j$ to create a standardization method, (ie to calculate mean and standard deviation for the studentization, or to calculate the empirical CDF for histogram equalization) and standardize all $\tilde{y}_i^j$ to y_i^j.
5. For each $j \in [1, N-1]$ calculate $y_i = \max_j y_i^j$.
6. Compute the rank r of y_1 among all the y_i.

We may then reject the null hypothesis at a level of $p = (N - r + 1)/N$.

3 Experiments

To validate our suggested FWER method and evaluate its performance we perform simulation tests on null model data input to verify the level, and with a single noisy circle input together with null model data input for a power analysis of each method. The level is measured by generating sets of point clouds from the null model, and measuring how often the joint null hypothesis is rejected. The power is measured by generating sets of point clouds both from the null model and using the noisy circle input, ensuring that exactly one of the point clouds is drawn from a noisy circle. The power is the rate of rejection of the null hypothesis for these cases. Examples of these noisy circles can be seen in Figure 1.

We use the null model of uniformly distributed points in a plane rectangle, and for computational expediency we restrict our testing to two ambient dimensions.

Our simulations test for all combinations of:

- $N \in \{100, 500\}$ (number of point clouds for each test)
- $K \in \{5, 10, 50\}$ (number of simultaneous tests to control)

For each box, we draw uniformly at random

- Box side lengths in $\{0.1, 1, 10\}$
- Point counts for a box in $\{10, 50, 100, 500\}$

Table 1: Rejection rates for null model and noisy circle data using the FWER control method described in Section 2.1.

$p <$	null	$\sigma = 0.1$	$\sigma = 0.25$
0.01	0.04	0.88	0.37
0.05	0.10	0.90	0.54
0.10	0.13	0.93	0.62

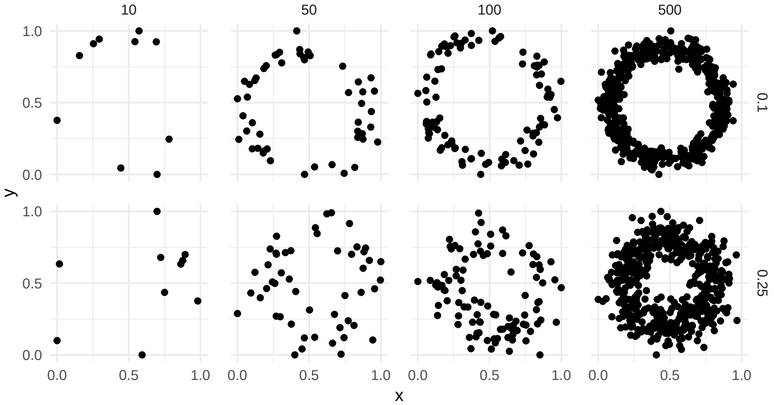

Figure 1: Noisy circles as used by the power calculation. Top row, $\sigma = 0.1$ and bottom row $\sigma = 0.25$. The plots have, from left to right, 10, 50, 100 and 500 points.

- For the power test: in one of the boxes, points on a circle with added multivariate isotropic Gaussian noise with variance from $\{0.1, 0.25\}$ fitted in a square box with side lengths 1×1.

The α-complex construction is topologically equivalent to Čech complexes [1], and for speed in our simulations we choose to use the α-complex persistent homology calculation in the R package TDA [9]. With simulations in place we perform bootstrap evaluations of level and power of our methods.

We will use the invariant $\gamma(X) = \sqrt{2}\|X\|_{\mathcal{B}} = \max t_d - t_b$ of maximum bar length.

We validate the FWER control procedures by estimating the probability of false discovery on null model data and we analyze the power of the proposed methods by attempting to detect a single noisy circle in a family of null model data samples.

For the experiments, we precomputed 160 000 point cloud invariants. Since we are working with point clouds in the plane, we computed in homological dimensions 0 and 1, and for each combination of box shapes and point counts as well as for each noise level and point count combination, we generated 5 000 point clouds.

3.1 Validation and Power estimation

We evaluate the empirical level of our proposed methods. From 100 simulations drawing from precomputed barcode sizes, the null rejection rates for null model data for our methods are summarized in Table 1. For each of the simulations, a random number, between 2 and 50 of point cloud invariants were drawn from the precomputed data. To each point cloud invariant, another 99 point clouds with matching box sizes and point counts are drawn as a simulation test. These 100 batches of 100 point clouds go through each of our proposed methods, and rejection rates at confidence levels of 0.1, 0.05 and 0.01 are calculated.

For the power analysis we picked pre-calculated invariants from circles with a 1×1 bounding box, with additive multivariate Gaussian noise with a standard deviation of 0.1 and 0.25 respectively – see Figure 1 for examples of the generated point clouds. For each of 100 simulations, one circle invariant was picked, and another random number (between 1 and 49) of null model point cloud invariants added. This collection of point clouds go through the same process of generating 99 null model invariants for each, and run the collections through the described methods. The result of 100 simulations each at the two noise levels is shown in Table 1.

Acknowledgments and Disclosure of Funding

On behalf of all authors, the corresponding author states that there is no conflict of interest.

MVJ would like to acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

SM would like to acknowledge partial funding from HFSP RGP005, NSF DMS 17-13012, NSF BCS 1552848, NSF DBI 1661386, NSF IIS 15-46331, NSF DMS 16-13261, as well as high-performance computing partially supported by grant 2016-IDG-1013 from the North Carolina Biotechnology Center

The simulated data used in this paper are available from Figshare, DOI 10.6084/m9.figshare.10262507

References

[1] Ulrich Bauer and Herbert Edelsbrunner. The Morse theory of čech and Delaunay filtrations. In *Proceedings of the thirtieth annual Symposium on Computational Geometry*, page 484. ACM, 2014.

[2] Andrew J. Blumberg, Itamar Gal, Michael A. Mandell, and Matthew Pancia. Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. *Foundations of Computational Mathematics*, 14(4):745–789, 2014. URL: http://link.springer.com/article/10.1007/s10208-014-9201-4.

[3] Carlo Bonferroni. Sulle medie multiple di potenze. *Bollettino dell'Unione Matematica Italiana*, 5(3-4):267–270, 1950.

[4] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015. URL: http://arxiv.org/abs/1207.6437.

[5] Christopher Cericola, Inga Johnson, Joshua Kiers, Mitchell Krock, Jordan Purdy, and Johanna Torrence. Extending Hypothesis Testing with Persistence Homology to Three or More Groups. *Involve, a Journal of Mathematics*, 11(1):27–51, January 2018. arXiv: 1602.03760. URL: http://arxiv.org/abs/1602.03760, `doi:10.2140/involve.2018.11.27`.

[6] Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884, 2017. URL: http://arxiv.org/abs/1412.7197.

[7] Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Subsampling methods for persistent homology. In *International Conference on Machine Learning*, pages 2143–2151, 2015. URL: http://arxiv.org/abs/1406.1901.

[8] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the Bootstrap for Persistence Diagrams and Landscapes. *arXiv:1311.0376 [cs, math, stat]*, November 2013. URL: http://arxiv.org/abs/1311.0376.

[9] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, and Vincent Rouvreau. Tda: statistical tools for topological data analysis. *Software available at https://cran.r-project.org/package=TDA*, 2014.

[10] Brittany Terese Fasy, A. Rinaldo, and L. Wasserman. Stochastic Convergence of Persistence Landscapes and Silhouettes. *Convergence*, (1/25), 2014.

[11] Yasuaki Hiraoka, Tomoyuki Shirai, and Khanh Duy Trinh. Limit theorems for persistence diagrams. *The Annals of Applied Probability*, 28(5):2740–2780, October 2018. `doi:10.1214/17-AAP1371`.

[12] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. URL: http://dx.doi.org/10.1093/biomet/75.4.800, `arXiv:/oup/backfile/content_public/journal/biomet/75/4/10.1093/biomet/75.4.800/2/75-4-800.pdf`, `doi:10.1093/biomet/75.4.800`.

[13] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[14] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011. URL: http://stacks.iop.org/0266-5611/27/i=12/a=124007.

[15] Elizabeth Munch, Katharine Turner, Paul Bendich, Sayan Mukherjee, Jonathan Mattingly, and John Harer. Probabilistic Fréchet means for time varying persistence diagrams. *Electronic Journal of Statistics*, 9(1):1173–1204, 2015.

[16] Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, November 2017. URL: http://link.springer.com/10.1007/s41468-017-0008-7, `doi:10.1007/s41468-017-0008-7`.

[17] Katharine Turner. Means and medians of sets of persistence diagrams. arXiv e-print 1307.8300, July 2013. URL: http://arxiv.org/abs/1307.8300.

[18] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

[19] Katharine Turner, Sayan Mukherjee, and Doug M. Boyer. Sufficient statistics for shapes and surfaces. *Annals of statistics*, 2013.