Multiple Hypothesis Testing with Persistent Homology

Mikael Vejdemo-Johansson
CUNY College of Staten Island
CUNY Graduate Center

Sayan Mukherjee
Duke University

Why?

Without multiple hypothesis testing, error probabilities compound when performing multiple hypothesis tests.

There are Data Mining applications that require hundreds or thousands of persistence calculations.

How?

We propose a simulation-based approach to testing with persistent homology, with the advantage that we can produce a multiple hypothesis correction framework for this test.

Choose:

- A *null model M* for the data. Example: Uniformly distributed points in the bounding box.
- A *persistence invariant* γ. Example: Length of the longest bar in Vietoris-Rips persistent homology of degree 1.

Given observed point clouds $X_1, \ldots, X_n$, a test procedure that can reject the hypothesis *All of the point clouds were drawn from M.* proceeds with:

- For each X_i, generate $M_i^2,\ldots,M_i^N$ from M.
- For each X_i, calculate $Y_i^1 = \gamma(X_i)$
  For each M_i^j, calculate $Y_i^j = \gamma(M_i^j)$
- For $Y_i^2,\ldots, Y_i^N$, estimate mean μ_i, std.dev. σ_i.
- For each Y_i^j, calculate $Z_i^j = (Y_i^j - \mu_i)/\sigma_i$.
- For each j, calculate $Z^j = \max_i \{Z_i^j\}$.
- Rank the Z^j, reject the null at $p=(N-\text{rank}(Z^1)+1)/N$

Why does this work?

Theorem (Hiraoka et al):
If M is an ergodic stationary point process in $\mathbb{R}^d$, and X_i a sample from the box $[-i/2, i/2]^d$, then the persistence diagrams generated by the homology of a large variety of filtered simplicial complex constructions converges to a distribution that depends on M and the simplicial complex construction as i grows, up to a multiplicative constant that depends on i.

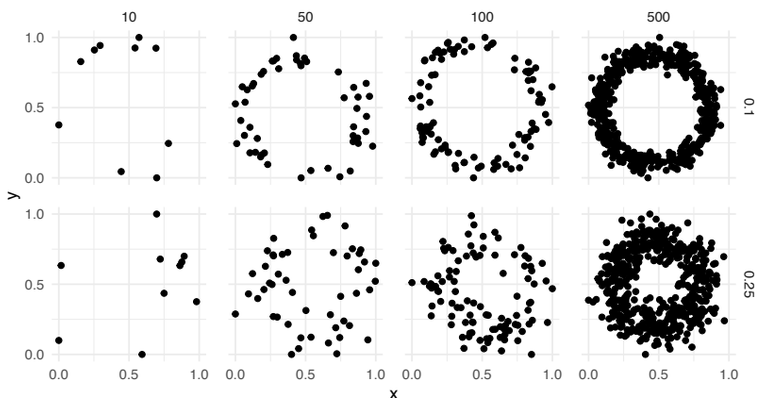
Convergence means that for large enough windows into a good null model M, the distribution (up to a constant factor) approximates a fixed distribution.

If we can find this multiplicative constant, we can compare point clouds drawn from different-sized windows.

The Z-score procedure estimates the multiplicative constant directly on barcode invariants.

The normalized Z-scores can be taken to represent N draws - each drawing point cloud collections similar to the $\{X_i\}$.

Maximizing the Z-scores picks a most extreme representative for each draw. Ranking picks the extent to which the $\{X_i\}$ yield extreme values.



How well does this work?

We measure rejection rates for our example null model as well as for two levels of isotropic Gaussian additive noise added to samples from a circle.

Using our example γ, we generated 5000 instances each for every combination of

N in {100, 500}
n in {5, 10, 50}
Box side lengths drawn from {0.1, 1, 10}
Point counts for a box from {10, 50, 100, 500}

These were rejection rates using our approach. In the cases σ=0.1 and σ=0.25, a single point cloud drawn from a circle was introduced alongside n-1 null model point clouds.

p<	null	σ=0.1	σ=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

Example point clouds at 10/50/100/500 points. Top row: σ=0.1, bottom row: σ=0.25.

arXiv preprint: https://arxiv.org/abs/1812.06491