Exploring the Intersection of Statistics and Topological Data Analysis

By: Alexander Richardson ‘25

Topological Data Analysis (TDA) is an emerging field in data science that uses mathematical techniques to understand the shape and structure of complex, high-dimensional data. Despite its potential, TDA has been slow to gain traction among statisticians. One key reason is that TDA typically requires choosing a distance metric—a way of measuring how "far apart" data points are. This choice is often made arbitrarily, which is not ideal for statisticians who prefer metrics that are directly informed by the data itself. Our project seeks to address this gap by developing statistically grounded distance metrics for TDA, specifically by deriving these metrics from statistical models. We focus on mixture models, which are flexible tools that combine multiple probability distributions to represent complex, multi-modal data.

We first created synthetic data using a Gaussian Mixture Model (GMM), a common type of mixture model, and used this synthetic data to fit mixture models. To begin, I delved into the Expectation-Maximization (EM) algorithm and Markov Chain Monte Carlo (MCMC) methods, which are widely used in frequentist and Bayesian statistics, respectively, for fitting mixture models to data. After gaining a deep understanding of these models and how to train them, we began deriving new distance metrics. We define the distance between two data points within a mixture model as the probability that they belong to the same component distribution. This definition can vary depending on whether the model is fit using the EM algorithm or MCMC methods, and we provided detailed derivations for both approaches.

To test our method, we applied it to the synthetic data, which was designed to have clear topological features—such as clusters of points that correspond to different component distributions. After fitting the mixture model to the data, we used a TDA technique called persistent homology to measure these features. Persistent homology captures topological properties like connected components (representing clusters in the data) and holes, with longer persistence indicating more significant features. We then compared the results of persistent homology using our mixture model-based distance metric against the traditional Euclidean distance. Notably, in a three-component case, our GMM-based distance metric more accurately captured the number of components than the Euclidean distance. Since the models are already fit, the TDA analysis should be relatively quick. Looking forward, we are particularly interested in applying our method to Hidden Markov Models (HMMs), which are a type of mixture model specifically designed for time-series data. We plan to use Bowdoin’s meal swipe data to build an HMM that models evolving social networks, making this an ideal dataset for testing our approach.

While the work in this work will continue throughout the semester, we already have promising results that demonstrate the viability of our approach. We are currently drafting a paper to share our findings with both the statistics and topology communities, and we are excited to see how our work will be received.

Faculty Mentor: Jack O’ Brien
Funded by: Surdna Foundation Undergraduate Research Fellowship Program