Conference on Applied Statistics in Agriculture 2015 - 27th Annual Conference Proceedings

Statistical Methods in Topological Data Analysis for Complex, High-Dimensional Data

Patrick S. Medina
Purdue University, medinap@purdue.edu

R W. Doerge
Purdue University - Calumet Campus, doerge@purdue.edu

Recommended Citation

Medina, Patrick S. and Doerge, R W. (2015). "Statistical Methods in Topological Data Analysis for Complex, High-Dimensional Data," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1130

STATISTICAL METHODS IN TOPOLOGICAL DATA ANALYSIS FOR COMPLEX, HIGH-DIMENSIONAL DATA

PATRICK S. MEDINA & R.W. DOERGE

Abstract. The utilization of statistical methods an their applications within the new field of study known as Topological Data Analysis has has tremendous potential for broadening our exploration and understanding of complex, high-dimensional data spaces. This paper provides an introductory overview of the mathematical underpinnings of Topological Data Analysis, the workflow to convert samples of data to topological summary statistics, and some of the statistical methods developed for performing inference on these topological summary statistics. The intention of this non-technical overview is to motivate statisticians who are interested in learning more about the subject.

1. Introduction

Suppose one is asked to analyze the sample data in Figure 1. What could be said? Obviously, it is two dimensional, and it appears to be circular. However, it may be that these data were sampled from a distribution whose support is in the shape of a circular coil. Toward this end, how could a sample of data points be used to study the overall shape of the support of the distribution from which they were sampled? Furthermore, is it possible to learn this if these data are high-dimensional?

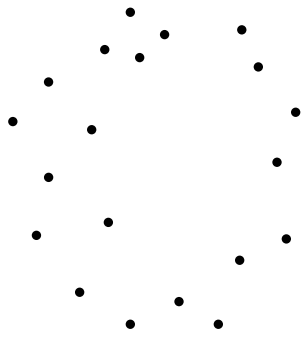

Figure 1. A random sample of 20 points sampled from some distribution $\mathcal{F}$ with unknown support $\mathcal{X} \subseteq \mathbb{R}^n$.

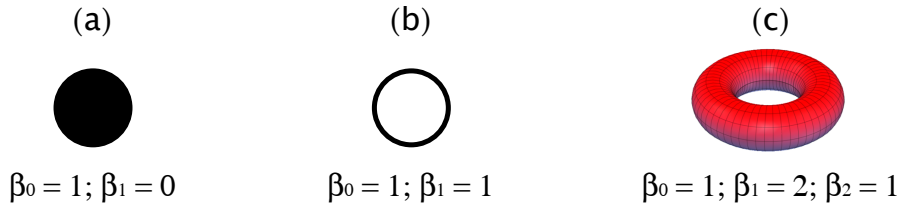

FIGURE 2. Three different mathematical spaces and their Betti numbers. Betti numbers quantify the distinct number of shape features that appear in a mathematical space. Specifically, connected components are H_0, loops by H_1, voids by H_2, and so on for higher dimensional analogues. **(a)** The image of a disc embedded in $\mathbb{R}^2$. If any two points are chosen in the disc then a line may be drawn connecting the two points. Hence, there is one connected component and the Betti number for H_0, β_0 is one. Further, since there are no holes in the disc then the Betti number for H_1, β_1 is one. **(b)** The image of a circle embedded in $\mathbb{R}^2$. If any two points are chosen along the circle then they may be connected by an arc between them. Also, there is a large hole inside of the circle. Hence, β_0 and β_1 are one. **(c)** The image of a Torus embedded in $\mathbb{R}^3$. If any two points are chosen along the exterior, then they may be connected by drawing a path between them, hence β_0 is one. A torus has two holes: the first is the large one in the center, and the second is seen by cutting the torus in half. Hence, β_1 is two. Finally, the inside of the torus is hollow, which means it has one void and the Betti number for H_2, β_2 is one.

Topological Data Analysis (TDA) has emerged as a branch of computational topology that enables researchers to study the shape properties of a mathematical space based on a representative sample taken from that space. The information is then used to learn how the original mathematical space is organized. TDA uses ideas from algebraic topology to quantify distinct shape characteristics of mathematical spaces. These concepts are general enough that they extend to data that are high-dimensional and very complex; i.e. they reside in spaces where traditional linear statistical methods or manifold learning techniques may fail to adequately capture properties of the underlying space.

In the context of a statistical problem, it is assumed that a sample of data $\mathcal{P}$ is drawn randomly from some distribution $\mathcal{F}$ whose support, $\mathcal{X} \subseteq \mathbb{R}^n$, is unknown. Based on this setting, tour motivation is to provide statisticians with a very friendly introduction to the mathematical underpinnings of Topological Data Analysis. Specifically, we provide an overview of the process by which data, $\mathcal{P}$, are converted to different topological summary statistics. The statistical methods developed for inference on a sample of topological summary statistics are introduced. Finally, the application of some of these statistical methods in analyzing differences between two structures of the maltose binding protein are examined.

2. Overview of Topological Data Analysis

TDA combines methods from the fields of algebraic topology and computational geometry for the purpose of studying key shape features of a mathematical space from which a sample of data may have been drawn. Two key tools for achieving this are homology and simplicial complexes.

2.1. **Homology.** Homology is a subject in algebraic topology that provides tools for computing the number of distinct shape features within a mathematical space. The shape features of interest are connected components - notated by H_0, loops - notated by H_1, voids - notated by H_2, and their higher dimensional analogues. For each of these shape features a Betti number quantifies the distinct number of shape features that appear in a mathematical space [1]. An example of different mathematical spaces and their Betti numbers is illustrated in Figure 2. Specific details on homology can be found in Munkres [2].

Betti numbers are typically used to understand the shape characteristics of a mathematical space. However, since data are a collection of discrete points, using the data directly as a representation of the underlying mathematical space will not, in general, capture interesting features that may exist. Hence, in order to learn about the shape features of the mathematical space, a geometric representation, known as a simplicial complex, of the space needs to be constructed from the data.

2.2. **Simplicial Complexes.** The main tool for constructing a geometric representation of a mathematical space from data is the simplicial complex. These are topological structures constructed by attaching simplices along their faces. A simplex is a general term for triangles in a Euclidean space. The 0-simplex, or vertex, is a point. The 1-simplex is a line segment. The 2-simplex is a triangle. Higher dimensional equivalents of these objects follow directly. Examples of the first three simplices are illustrated in Figure 3. Computational methods have been developed to construct a simplicial complex using a sample of data as the vertices. The details of simplicial complexes can be found in Munkres [2].

2.2.1. *Computational methods for constructing simplicial complexes.* One of the most common approaches for constructing a simplicial complex is the Vietoris-Rips complex [3]. It is constructed by examining all pairwise distances between points, $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$. Since the data are from a subset $\mathcal{X} \subset \mathbb{R}^n$, numerous notions of distance can be used. Two common metrics are the p-distance,

$$d_p(x, y) = \Big(\sum_{i=1}^{n} |x_i - y_i|^p \Big)^{\frac{1}{p}},$$

and the maximum distance,

$$d_\infty(x, y) = \max \Big\{ |x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n| \Big\}.$$

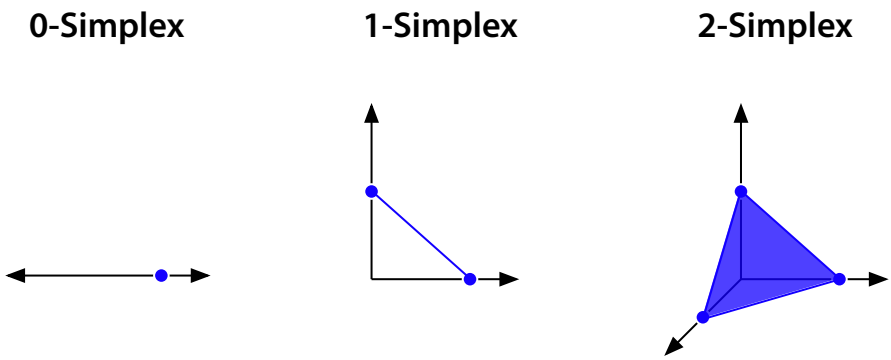

FIGURE 3. Examples of the first three simplices. The 0-simplex embedded in $\mathbb{R}$ is a point. The 1-simplex embedded in $\mathbb{R}^2$ is a line segement. The 2-simplex embedded in $\mathbb{R}^3$ is a triangle.

When $p = 2$, the p-distance is the Euclidean distance that is used in many applications, such as regression. While these metrics are commonly used in the construction of the complex, other metrics may certainly be considered.

The Vietoris-Rips complex is most easily understood by considering a subcollection of points of $\tau = \{p_{i_1}, \ldots, p_{i_m}\} \subseteq \mathcal{P}$. τ is a simplex in the simplicial complex if the points are all relatively close to each other. Specifically, the concept of closeness is conceived via a fixed scale parameter $\epsilon > 0$, and τ is a simplex in the simplicial complex if $d(p_{i_j}, p_{i_k}) < \epsilon$ for all i_j and i_k. Once this complex is constructed, it is possible to use the tools developed from homology to compute the number of distinct shape features present.

2.2.2. *Choosing an appropriate scaling parameter.* For a fixed value of ϵ, the Vietoris-Rips complex is an approximation of the underlying structure. As ϵ increases, more simplices are added to the simplicial complex, until it contains all possible simplices that can be generated from $\mathcal{P}$. Because of the dependency on the different values of the scale parameter, information about the shape of the underlying mathematical space varies, see Figure 4.

Instead of using a specific scale parameter ϵ, Topological Data Analysis uses a range of scale parameters to dynamically keep track of when distinct shape features appear and disappear from the simplicial complex. The concept is similar to the idea of hierarchical clustering, which tracks cluster membership over a specified scale parameter, and is known in TDA as "persistent homology."

2.3. **Persistent Homology.** As mentioned, persistent homology [4] tracks the appearance and disappearance of distinct shape features in a simplicial complex via a changing scale parameter. It allows users to understand the number of features that appear in their data,

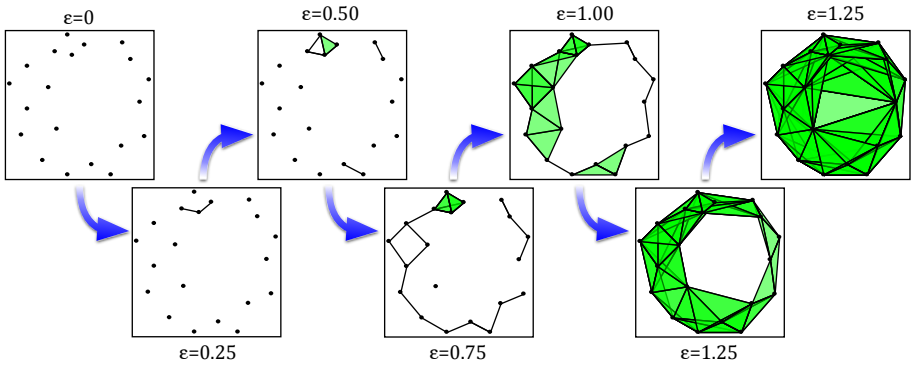

FIGURE 4. The evolution of the simplicial complex at different values of the scaling parameter. As the scaling parameter increases, more simplices are added to the simplicial complex.

but also, it evaluates how long these features exist. The details of persistent homology may be found in Edelsbrunner et al. [4] and Edelsbrunner and Harer [5]. Reviews of the Topological Data Analysis workflow may be found in Carlsson [1], Ghrist [6], and Nanda and Sazadanović [7].

3. Topological Summary Statistics

Topological summary statistics allow the quantification and visualization of persistent homology. The main summary statistics used in Topological Data Analysis are the persistence diagram, barcode, and the persistence landscape. Persistence diagrams and barcodes are closely related and, as such, the focus here will be on the persistence diagram and persistence landscape.

3.1. **The Persistence Diagram.**

Persistent homology tracks the existence of distinct shape features over a range of values for a scaling parameter. If a shape feature appears at a parameter value of ϵ_a, and disappears at a parameter value of ϵ_b, then then the persistence diagram retains this information through the point (ϵ_a, ϵ_b) in $\mathbb{R}^2$. The length of the shape feature's existence can be measured by the vertical distance between this point and the diagonal $(y = x)$. Hence, a persistence diagram is a multiset (a set that allows the number of elements to be repeated) of points in $\mathbb{R}^2$, and the diagonal, where the diagonal has infinite multiplicity. Figure 8 illustrates a persistence diagram and explains its interpretation. Details and methods for computing persistence diagrams are covered in Edelsbrunner and Harer [5].

3.1.1. *Mathematical properties of the persistence diagram.* In order to make the space of persistence diagrams amenable to statistical inference, the definition of a persistence diagram is limited to a finite multiset of points in $\mathbb{R}^2$, and the diagonal $(y = x)$, where each point on the diagonal has infinite multiplicity [8].

The set of persistence diagrams that satisfy this definition, $\mathcal{D}$, equipped with the Wasserstein metric is considered a metric space. The p^{th} Wasserstein distance between two persistence diagrams $d_1, d_2 \in \mathcal{D}$ is defined as

$$W_p(d_1, d_2) := \Big(\inf_{\gamma} \sum_{x \in d_1} ||x - \gamma(x)||_\infty^p \Big)^{\frac{1}{p}},$$

where γ varies across all one-to-one and onto functions from d_1 to d_2. When $\mathcal{D}$ is restricted to the set of persistence diagrams whose p^{th} Wasserstein distance from itself to the diagonal is finite, Mileyko et al. [8] show that the space of previous persistence diagrams is a Polish space [9]. This result gives rise to the statistical construct of a mean and variance for persistence diagrams (see Section 4.1).

3.1.2. *Comparison to the barcode.* Persistence diagrams are closely related to the summary statistic referred to as the barcode [10]. A barcode is simply a multiset of intervals (ϵ_a, ϵ_b) in $\mathbb{R}^2$, with $\epsilon_a < \epsilon_b$, that keeps track of the appearance and disappearance of shape features across different scaling parameters. Although the barcode is similar to that of the persistence diagram it does not require information about the diagonal. Metrics and other properties of barcodes may be found in Carlsson et al. [11], and Zomorodian and Carlsson [12]. An example of how barcodes are constructed is found in Figure 8.

3.2. **Persistence Landscapes.** Persistence landscapes are a new concept [13] and are an alternative measure of persistent homology. Persistence landscapes, while a statistic, in fact resides in a nicer mathematical space than persistence diagrams. As such, they are more amenable to existing statistical methods (fully discussed in Section 4.4). The interpretation of the persistence landscape is more subtle than the interpretation of either the persistence diagram and the barcode. Rather than directly encoding information about the number of shape features in a homology group and their length of existence, the persistence landscape gives a measure of the number of features that simultaneously exist at a particular scaling parameter. An example of a persistence landscape is in Figure 5.

3.2.1. *Mathematical properties of a persistence landscape.* A persistence landscape, Λ, is a sequence of piecewise continuous functions $\lambda_k : \mathbb{R} \to \mathbb{R}$. Distance measures may be defined for persistence landscapes by integrating the functions, λ_k, and summing the result across all values of k. Persistence landscapes can be treated as a random variable that takes values in a Banach space, which gives rise to the use of probabilistic results [14]. Details and other results are covered in Bubenik [13].

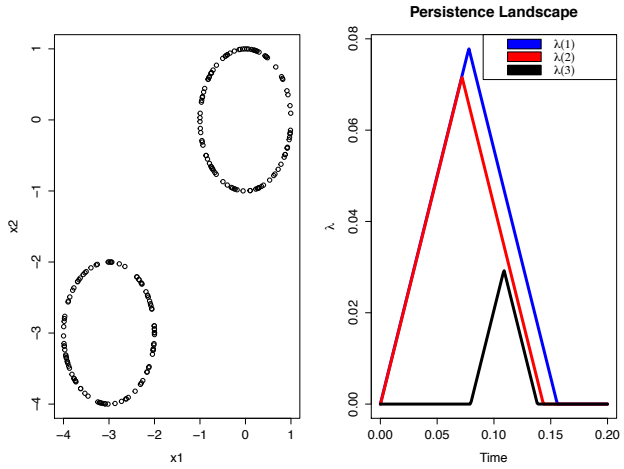

FIGURE 5. **Left:** A random sample of 200 points uniformly drawn from two different circles; 100 points were sampled from the circle centered at $(0, 0)$ and 100 points were uniformly sampled from the circle centered $(-3, -3)$. For these samples the Betti numbers for the connected components and holes are both two. **Right:** The persistence landscape for the number of connected components. The persistence landscapes are such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Since the first two persistence landscapes are large, this gives evidence for the number of connected components being two, which is consistent with the example on the left.

4. Statistical Inference with Topological Summary Statistics

A significant amount of work has been performed for the purpose of extending statistical methods to Topological Data Analysis. We are specifically interested in defining a clear notion of the mean and variance for persistent homology, developing testing methods that distinguish between the distributions from which topological summary statistics may have been sampled, and methods for determining which shape characteristics are statistically significant and which are topological noise. Here, we review some of the statistical methods developed for both the persistence diagram and the persistence landscape.

It is generally assumed that the points of the sample $\mathcal{P} = \{X_1, \ldots, X_n\} \subseteq \mathbb{R}^n$ are identically and independently distributed through some random process with distribution $\mathcal{F}_X$. Applying the Topological Data Analysis workflow to $\mathcal{P}$ results in a topological summary statistic, $TS(\mathcal{P})$. However, the distribution of our topological summary statistic may not follow the same distribution as the original sample. Hence, in order to understand the distribution of the topological summary statistic it is necessary to understand the effect of the topological transformation on the distribution of the original data. As a simple example of this effect, suppose U is a random sample drawn from Uniform$(0, 1)$ distribution,

then $-\ln(U)$ is distributed as an exponential(1) random variable. In general, it is not clear what influence the topological transformation has on the distribution of the data.

4.1. **Fréchet mean and variance of persistence diagrams.** The Fréchet mean and variance of persistence diagrams are discussed in [8]. Let $(\mathcal{D}_\mathcal{P}, \mathcal{B}(\mathcal{D}_\mathcal{P}), \mathcal{F}_{\mathcal{D}_\mathcal{P}})$ be a probability space on the space of persistence diagrams as defined in Section 3.1.1, where $\mathcal{B}(\mathcal{D}_\mathcal{P})$ is the Borel σ-algebra on $\mathcal{D}_\mathcal{P}$, and $\mathcal{F}_{\mathcal{D}_\mathcal{P}}$ is a probability measure on this space. In order to define the Fréchet mean it is required that $\mathcal{F}_{\mathcal{D}_\mathcal{P}}$ have a finite second moment. That is,

$$M_{\mathcal{D}_\mathcal{P}}(d) = \int_{\mathcal{D}_\mathcal{P}} W_p(d, e)\, d\mathcal{F}_{\mathcal{D}_\mathcal{P}}(e) < \infty\,,$$

for a fixed diagram $d \in \mathcal{D}_\mathcal{P}$. The Fréchet variance is defined as

$$\mathrm{Var}_{\mathcal{F}_{\mathcal{D}_\mathcal{P}}} := \inf_{d \in \mathcal{D}_\mathcal{P}} \left\{ M_{\mathcal{D}_\mathcal{P}}(d) < \infty \right\},$$

and the Fréchet mean by

$$\mathrm{E}_{\mathcal{F}_{\mathcal{D}_\mathcal{P}}} := \left\{ d | M_{\mathcal{D}_\mathcal{P}}(d) = \mathrm{Var}_{\mathcal{F}_{\mathcal{D}_\mathcal{P}}} \right\}.$$

Since the definition of a Fréchet mean is an infimum over a space, in general, the mean may not be unique. To our knowledge, this is the only definition of a mean and variance for persistence landscapes.

4.1.1. *Algorithm for computing the Fréchet mean and variance.* An algorithm for computing the Fréchet mean is given in Turner et al. [15] for the special case of the L^2-Wasserstein metric and the distribution of the sample of persistence diagrams is a combination of Dirac masses [16]. In this setting, the authors provide a law of large numbers, however they can only ensure that their algorithm converges to a local minimum.

4.2. **Hypothesis Testing for Persistence Diagrams.** A problem of particular interest in Topological Data Analysis is determining whether two subsets of $\mathbb{R}^n$ are the same. Robinson and Turner [17] present an argument that a necessary, but not sufficient, condition is that the underlying distribution of their persistence diagrams are the same. To accomplish this, they develop a nonparametric permutation test to test for differences in the distributions of two different samples of persistence diagrams. Rejecting a null hypothesis, H_0, that the two distributions are the same provides evidence that the two subsets, themselves, are different. Details of the joint loss function that is employed, the reasons for developing a permutation test, and the justification for the output of their method being a p-value are in Robinson and Turner [17].

4.3. **Confidence Sets for Persistence Diagrams.** As with the majority of statistical applications, there is an interest in separating signal from noise. As such, an interesting question in persistent homology is how to distinguish important shape features from topological noise. The general working hypothesis in Topological Data Analysis is that features which exist (i.e., persist) over large intervals of the scaling parameter are significant. However, it is not always clear what constitutes a large interval, or if features that exist over a small interval are truly noise or of interest. This first issue is addressed by Fasy et al. [18] who originated methods for computing a $1-\alpha$ confidence set for an estimated persistence diagram. For instance, suppose $\mathcal{P}$ is a single sample taken from some distribution $\mathcal{F}$, then a persistence diagram $\hat{d}$ may be constructed from the data. A value, c_n, is computed from the data so that when the vertical distance between a point and the diagonal is less than $\sqrt{2}c_n$ the lifespan of the feature is not different from zero.

4.4. **Persistence Landscapes.** Persistence landscapes that are assumed to be random variables that take values in a Banach space are more amenable to the classical statistical theory of hypothesis testing and confidence intervals. In Bubenik [13] the vector space structure of the underlying L^p space is used to construct a pointwise mean for the persistence landscape Λ. That is, if we have samples $\mathcal{P}_1, \ldots, \mathcal{P}_n$, with corresponding persistence landscapes $\Lambda^1, \ldots, \Lambda^n$, then the mean landscape for the k^{th} sequence is given by

$$\bar{\lambda}_k(t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_k^i(t)\,.$$

Using results from [14], Bubenik [13] shows that the Strong Law of Large Numbers holds for the mean persistence landscape and that they obey the Central Limit Theorem. Chazal et al. [19] show that the convergence of the Central Limit Theorem is uniform.

4.4.1. *Hypothesis testing for persistence landscapes.* An advantage of the persistence landscape is that it allows for hypothesis testing when samples are high-dimensional and nonlinear. In order to use persistence landscapes for hypothesis testing one has to use a functional - functions that map $\mathbb{R}^n$ to $\mathbb{R}$ - on the persistence landscapes. When functionals satisfy certain conditions, Bubenik [13] proved the Central Limit Theorem remains for the transformed persistence landscape. Under this framework classical hypothesis testing procedures exist, and can be employed to distinguish between objects. In other words, for a large number of samples, it is possible to use common t-tests or Hotelling's T^2 to distinguish between two subsets of $\mathbb{R}^n$.

It should be noted that applying a functional in fact, projects all of the information in a persistence landscape to a single point. While this may result in a loss of information, this approach makes it easier to directly apply classical statistical tests, such as the t-test.

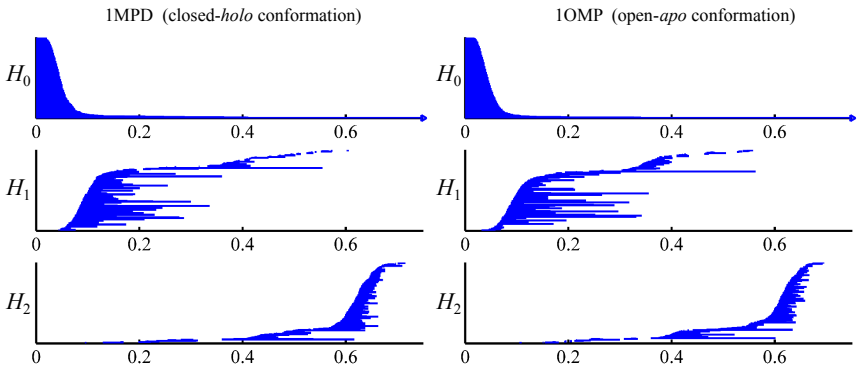

FIGURE 6. Barcodes for the first three homology groups of the closed conformation structure (left) and the open conformation structure (right) [20]. Slight differences between the two structures can be seen in the first homology group. Differences between the other homology groups are more subtle.

5. Application of Topological Data Analysis for Studying the Conformation Space of the Maltose Binding Protein

Technologies exist to capture proteins in a way that enables the study of their physical structure in three-dimensional space. However, when a protein is captured its physical structure represents one of many possible shapes. Many factors, including environmental or biological function, may influence the overall shape of a protein. Kovacev-Nikolic et al. [20] consider the conformation space, or the space of possible shapes, of the maltose binding protein was studied using topological methods. It is known that the maltose binding protein makes conformational changes when a ligand attaches. If a protein is closed it is always prone to having a ligand attached. One objective of Kovacev-Nikolic et al. [20] was to determine if statistically significant differences exist between the structures of the open and closed conformation space.

A sample of seven open structures and seven closed structures were obtained from the protein data bank [21]. The proteins were converted from their physical coordinates to a structure that considers the energy relationship between residues using the elastic network model. Reasons for this conversion are discussed in [20]. Persistent homology is then computed on each sample using the Vietoris-Rips complex, discussed in Section 2.2.1. Figure 6 illustrates one of the computed barcodes for the first three homology groups, and Figure 7 illustrates the mean persistence landscapes for the first two homology groups. In order to perform a hypothesis test, a functional is applied to each persistence landscape, resulting in a single value

$$X_{i,h}^{j} = \sum_{k=0}^{\infty} \int_{\mathbb{R}} \lambda_k(t)\, dt\,, \tag{1}$$

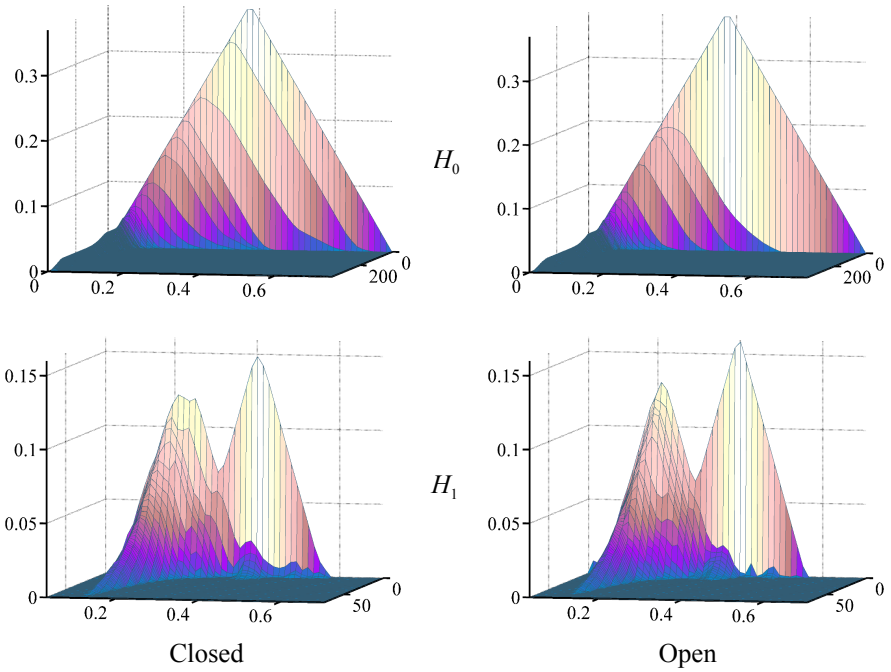

FIGURE 7. Persistence landscapes for the number of connected components, H_0 (top) and holes, H_1 (bottom) for the closed conformation structure (left) and the open conformation structure (right) [20]. (top) Observable differences between the closed and open persistence landscape are observed in the first homology group, H_0. The second persistence landscape of the closed structure attains a peak value of more than 0.30 and attains that value at about 0.40, while the second persistence landscape of the open structure attains a peak value at around 0.20 and attains that value at about 0.30. For the first homology group, H_1, the initial peak of the first few persistence landscapes of the closed structure are uneven, while the same persistence landscapes in the open structure are smooth.

where $1 \leq j \leq 7$, $i \in \{\text{open}, \text{closed}\}$, and $0 \leq h \leq 2$ is the homology group. This functional is simply the total area area under all of the persistence landscapes in the k^{th} homology group. Using these values, a permutation test is performed to compare the mean value of each homology group $H_0 : \mu_{h,\text{open}} = \mu_{h,\text{closed}}$ against $H_a : \mu_{h,\text{open}} \neq \mu_{h,\text{closed}}$. The p-values for homology in both degree zero and one were computed to be 5.83×10^{-4} giving evidence of a significant difference in the number of connected components and holes, while the p-value for the second homology group is 0.0396. No corrections on the significance level were considered for multiple-testing, however the p-values for homology in degree

zero and degree one are small enough to conclude differences in the means exist for these homology groups.

6. Discussion

Improvements in biotechnology have resulted in a massive growth of the amount and complexity of the data used in every field. In this work we have introduced the untapped power of tools provided by Topological Data Analysis. The TDA framework is well situated for studying datasets that are high-dimensional and complex. Although there is evidence of success in applying these methods to both visualize high dimensional data and for classification, further extensions of statistical methods to Topological Data Analysis are needed.

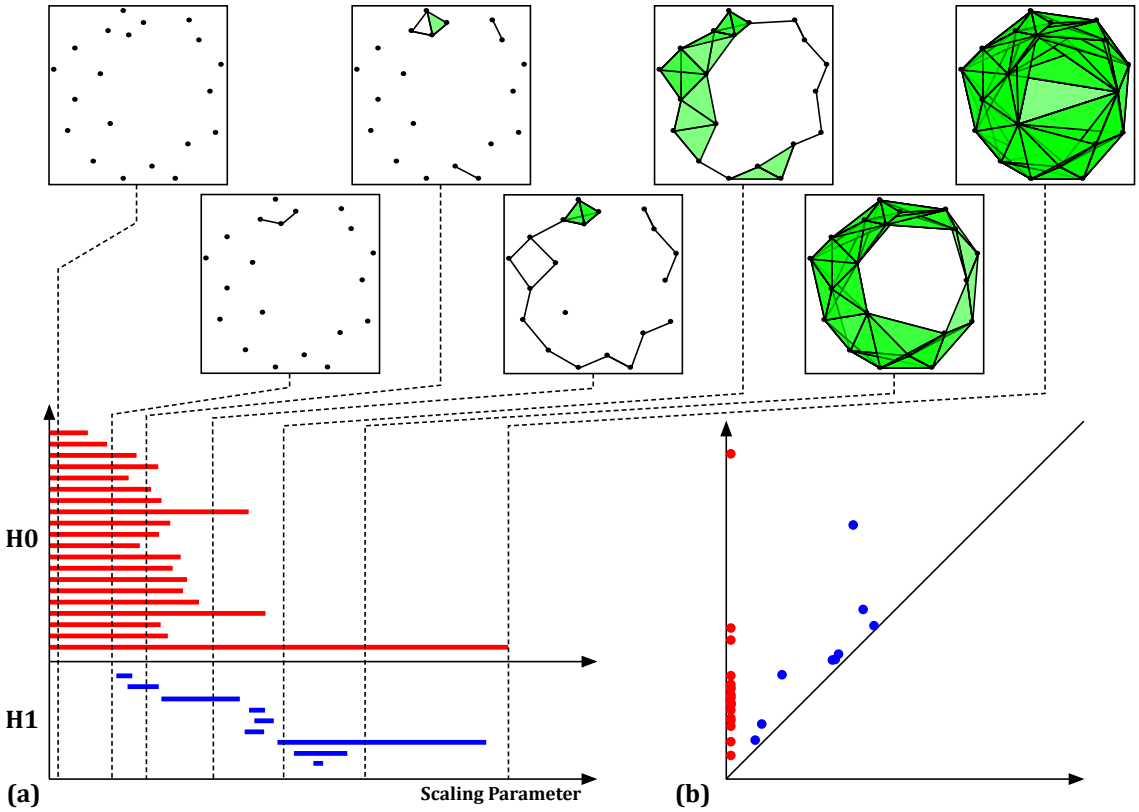

FIGURE 8. The top of the plots represent the simplicial complex at different values of the scaling parameter. **(a)** Illustration of the barcode summary statistic. The red and blue bars show the persistent homology of the connected components, H_0, and loops, H_1, respectively. The x-axis is the scaling parameter. At each value of the scaling parameter the number of distinct features that exist may be counted for each homology group by simply counting the number of bars at that value. For example, for the first simplicial complex there are twenty connected components and zero holes since the complex is starting with twenty vertices. At the fourth simplicial complex there are three connected components and one hole. The length of each bar indicates the length of each shape feature's existence. **(b)** Illustration of the persistence diagram. The red and blue dots indicate show the persistent homology of the number of connected components and holes respectively. For each dot, the x-coordinate indicates the value of the scaling parameter at which the shape feature appeared and the y-coordinate indicates the value of the scaling parameter at which the shape feature disappeared. The distance from each point to the diagonal indicates the length of each shape feature's existence. Since all features in H_0 existed at time zero, they are all above the point $x = 0$. The features for H_1 came into existence at later times, so they are scattered at different points in the x-axis.

References

[1] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009-04.
[2] James R. Munkres. *Elements of Algebraic Topology*. Perseus, Reading, MA, 1984.
[3] Afra Zomorodian. Fast construction of the vietoris-rips complex. *Computer and Graphics*, page 263271, 2010.
[4] Letscher, D. Edelsbrunner, H. and Zomorodian, A. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.
[5] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.
[6] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008-01.
[7] Vidit Nanda and Sazadanović Radmila. Simplicial models and topological inference in biological systems. In Nataša Jonoska and Masahico Saito, editors, *Discrete and Topological Models in Molecular Biology*, Natural Computing Series. Springer Berlin Heidelberg, 2014.
[8] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.
[9] Richard M. Dudley. *Elements of Algebraic Topology*. Chapman & Hall, New York, NY, 1989.
[10] Anne Collins, Afra Zomorodian, Gunnar Carlsson, and Leonidas Guibas, J. A barcode shape descriptor for curve point cloud data. *Computer & Graphics*, 28:881–894, 2004.
[11] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas J. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(2):149–187, 2005.
[12] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Computational Geometry*, 33:249–274, 2005.
[13] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015-01.
[14] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Classics in Mathematics. Springer-Verlag, 2011.
[15] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014-07.
[16] Jean Dieudonne. *Treatise on Analysis, Volume 2*. Pure and Applied Mathematics (Book 10). Academic Press, 1976.
[17] Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. *arXiv preprint arXiv:1310.7467*, 2013.
[18] Brittany T. Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.
[19] Frederic Chazal, Brittany T. Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *arXiv:1312.0308 [math.ST]*.
[20] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolic, and Giseon Heo. Using cycles in high dimensional data to analyze protein binding. *eprint arXiv:1412.1394*, page 21, 2014-12.
[21] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535–542, May 1977.