SIGNATURE PAGE

This thesis for honors recognition has been approved for the

Department of Mathematics .


Director

May 1 2017
Date


Reader

5-1-17
Date


Reader

5-1-17
Date

Topological Data Analysis: Giving Data Shape

Dylan Allen

May 1, 2017

Abstract

Topological Data Analysis (TDA) combines topology and data analytics which offers a new perspective when analyzing data. More so, TDA is capable of providing shape to data that otherwise may be difficult to visualize. In this thesis we provide a brief overview of an algorithm called MAPPER. We analyze two data sets, using statistical techniques and TDA. In the first data set TDA provides a summary where areas with high crime rate are noticeably separate from low crime rates. In the second data set TDA correctly diagnosed benign and malignant tumors in subsets of patients with 100% accuracy. In addition we note a subset of patients that seem to be related, but differ in a given attribute; which changes the model's diagnosis.

Contents

1 Introduction to Topological Data Analysis

Topology is the branch of mathematics that focuses on properties that are preserved through any type of deformation. Take for example some connected wire that is in the shape of a circle. We can deform this wire to take on the shape of a rectangle. Thus, shapes that are not equivalent in Euclidean geometry may be equivalent in topology. This concept in topology is referred to as a homeomorphism. Note that a *homeomorphism* is a function $f : X \to Y$ between two topological spaces X and Y that: *is a continuous bijection, and* and *has a continuous inverse function* f^{-1}. By the process described above there exists a continuous bijection from the circle to the rectangle. However, no such homeomorphism exists between a disk and an annulus thus they are not homeomorphic, see Figure 1.

Topological data analysis (TDA) is a combination of topology and data analytics. The field of topology can be described as "the study of geometric properties and spatial relations unaffected by the continuous change of shape or size of figures"(Oxford Dictionaries, 2017). That is, stretching or pulling of an object has no affect on an objects topological properties. Data analytics is also incorporated in TDA which is a broad field of mathematics incorporating statistics, computer science, and other forms of analysis to understand patterns in data.

TDA began in the early 1970's when Gunnar Carlsson, Harlan Sexton, and Benjamin Mann were Ph.D. students at Stanford University (Ayasdi, 2017). During their time at Standford they pondered the idea of applying topological theory to understand data. This concept of TDA gave rise to many interesting discoveries including persistence diagrams. In 2007 Gurjeet Singh invented an algorithm called MAPPER; which is capable of summarizing large datasets into a connected graph (Singh et al, 2007).

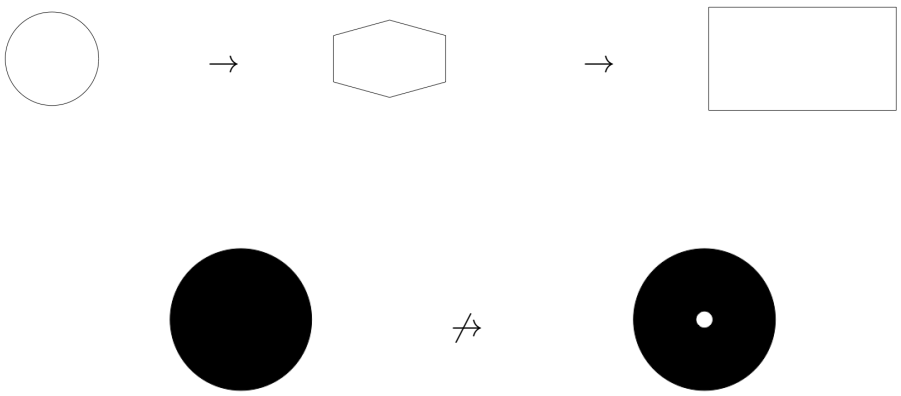

Figure 1: *Top*: Homeomorphism between a circle, hexagon, and rectangle *bottom*: a non-example of a homeomorphism.

1.1 Three Noteworthy Properties of TDA

The idea behind Topological Data Analysis is that all data has shape, and we can summarize this shape into a simplicial complex called a topological summary. This process can be justified with three fundamental observations (Kraft, 2016):

1. *Coordinate Invariance*: The topological properties of a shape are the same regardless of what coordinate system it is in. That is, we can move a shape or scale the coordinates and the object is exactly the same.

2. *Deformation Invariance*: Stretching, pulling, twisting a shape does not affect its underlying attributes. However, tearing or breaking a shape will affect its original shape.

3. *Compressed Representation*: We can simplify a large data set that has many points into a much simpler object with nodes and edges.

1.2 MAPPER

There exist data visualization methods which allow us to understand data through images and figures. Some methods include, but are not limited to scatter plots, heat maps, tree diagrams, and even multidimensional visualization techniques such as star plots, Chernoff faces, and spider plots. For the rest of this thesis we will focus on a different data visualization tool called MAPPER which uses topological theory to understand data. The output of MAPPER is a graph, with nodes and edges, which is a topological summary of the point cloud data being analyzed. More importantly it creates a visual representation of an entire data set in one connected graph.

In order to create a topological summary of the data there are several paramaters that must be determined:

Definition 1.1. MAPPER parameters

i *Filter function*: A function that maps a data frame X to the real numbers, that is $f : X \to \mathbb{R}$.

ii *Clustering technique*: A technique which allows the grouping of data entries into meaningful subsets of the data.

iii *Number of Intervals*: the total number of subsets which create an open cover of $f(X)$.

iv *Percent Overlap*: The percent overlap of each open cover in $f(X)$.

A more in depth explanation of each of the parameters will be discussed in Sections 2.1, 2.2, and 2.3. Nonetheless, a general understanding of the MAPPER algorithm will be presented here. Suppose we have some data set X that resembles a torus. We begin by coloring the data points in X by intervals from some filter function f. In addition some of these colored points will overlap with the other intervals and therefore be a part of two intervals. The points in each interval are then clustered by a

chosen clustering algorithm. The connected graph is then generated where the edges represent nodes that share a common data entry. Example 1 provides a visualization of what MAPPER is doing on a torus in $\mathbb{R}^3$.

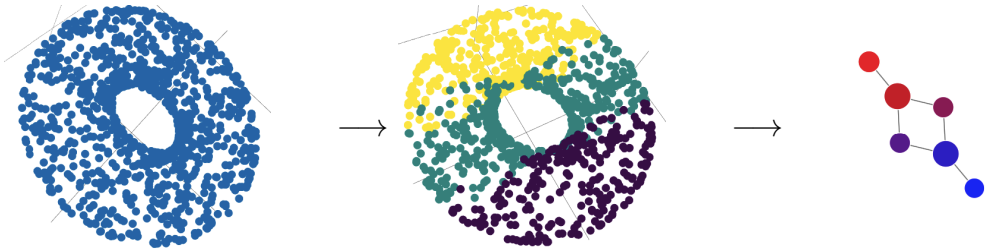

Figure 2: *left:* unfiltered torus. *center:* colored torus after applying filter function (y). *right:* topological summary of the torus

Example 1. In Example 1 we take some data set X that resembles a torus (donut) in $\mathbb{R}^3$. We then filter it, in this case we mapping points to their y values. We set the number of intervals to three and the percent overlap to 10%. After running MAPPER we see that the topological summary in Figure 2 gives the fundamental shape of the original data set X, in that it retains the hole of the donut.

Now suppose we set the filter function to the z values associated with each point $n \in X$. Then the topological summary is something fairly different. We see a straight line in the topological summary rather than a hole as in Figure 2. An example of this progression can be seen in Figure 3.

We see in Example 1 that the choice of filter function will change the topological summary. Thus changing the interpretation of the data being analyzed. Therefore choosing a filter function that explains the data is an important decision when choosing the MAPPER parameters.

2 MAPPER Inputs

As we will see in the upcoming sections MAPPER is robust and is capable of accepting different inputs depending on the users interest. The different methods presented in

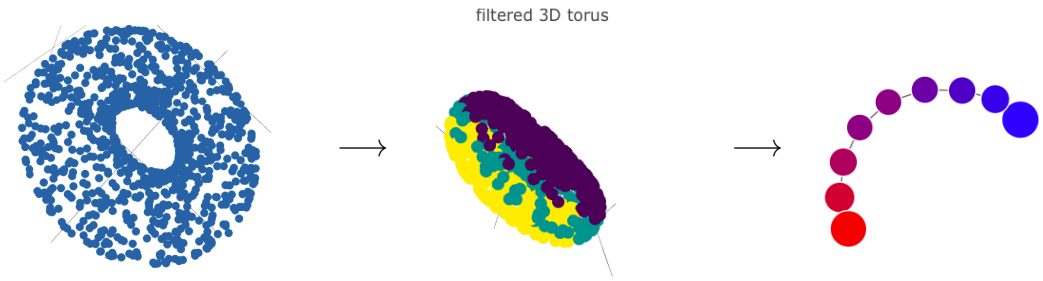

Figure 3: *left* unfiltered torus. *center* colored torus after applying filter function (z). *right* topological summary of the torus

this section are not a complete list of possible input parameters for MAPPER, but it will illustrate to the reader the adaptiveness of the algorithm. Moreover we will consider the inputs of interest that can be varied, for example: filter function, metric or distance function, and clustering technique. The number of intervals and percent overlap are also interesting and adaptable; however, we will not discuss them in upcoming sections.

2.1 Filter Functions

Choosing an appropriate filter function is a crucial yet difficult task when running MAPPER as we saw in Example 1. Some filter functions work better than others; however, there is no formal process in determining the correct one. In the this section we present a valuable filter function Singular Value Decomposition (SVD) that can be used when running MAPPER.

Singular Value Decomposition

The SVD has been shown to have many applications including image compression, signal processing, and principle component analysis. In addition to SVD's other numerous applications it has been shown to be a valuable filter function for MAPPER.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{bmatrix} \xrightarrow{f} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}$$

Figure 4: The matrix on the left represents X, After applying the filter function we have a $m \times 1$ vector which summarizes the data.

Definition 2.1. SVD (Muller et al, 2004)

The SVD of a matrix X of size $m \times n$ is

$$X = U\Sigma V^T, \tag{1}$$

where U and V are $m \times m$ and $n \times n$ orthonormal matrices. Σ is an $m \times n$ diagonal matrix with the nonnegative singular values $\sigma_j, j = 1, \dots, \min(m, n)$, arranged in nonincreasing order along the diagonal. The columns of U and V are denoted by the vectors $\mathbf{u_j}, j = 1, \dots, m$, and $\mathbf{v_j}, j = 1, \dots, n$.

Remark (Validity of SVD)**.** SVD is a matrix decomposition technique which is capable of converting a matrix X into its best rank k approximation for any $k \in [1, \text{rank}X]$ where the rank$X \in \mathbb{Z}$. Also, the vectors that comprise U and V are ranked. For example, the first vector $\mathbf{u_1}$ and $\mathbf{v_1}$ used in the best rank one approximation of the matrix X. Therefore we can use the vectors in U as a filter function.

We choose X to be some point cloud data, and the function f is described as $f(i) = i^{th}$ entry of $\mathbf{u_i}$ is the first column in U. Notice that the matrix U is comprised of the left singular vectors (the unit eigenvectors of XX^T). In practice any of the $\mathbf{u_i}$ vectors in the left singular matrix U can be used for MAPPER, see Figure 4. However, because U is formated such that the "best" singular vector is in the first column, $\mathbf{u_1}$ is typically a sound option.

Recall the Torus example given in Example 1. We chose the filter function $f : X \to Y$ to be defined as $f(x, y, z) = y$ for all $(x, y, z) \in X$. Here, X is a 1000×3

matrix and is mapped to a 1000×1 vector containing the y values of X. Now suppose instead we use the SVD to define a filter function form the vector $\mathbf{u_1}$, the first left singular vector of X. A graphical representation of this process can be seen in Example 2.

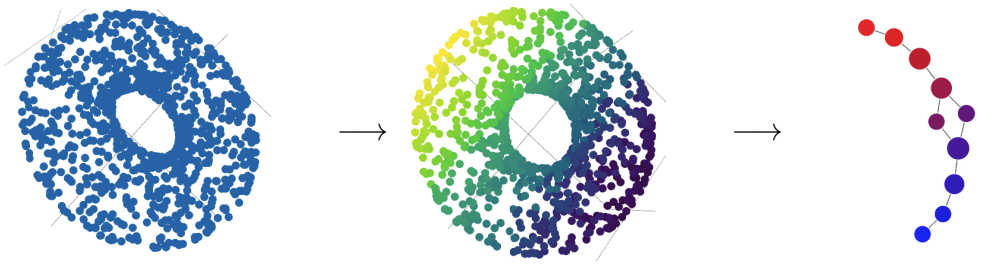

Figure 5: *Left* unfiltered torus. *Center* colored torus after applying filter function. *Right* topological summary of the torus

Example 2. *MAPPER on Torus Using SVD.* On the left we have the graphical representation of the data frame X. We use the SVD as a filter function using the left singular vector $\mathbf{u_1} \in U$. Now let $a_i \in \mathbf{u_1}$ where $i \in [1, 1000]$. We then create a mapping $f : X \to \mathbb{R}$ where $f(i) = a_i$. and color these observations by the respective a_i values, as in Figure 5. After choosing a clustering technique we can generate the topological summary seen in the right image in Figure 5.

2.2 Metric Space

Another component of MAPPER is the clustering algorithm which partitions data into a discrete number of groups. Fortunately, MAPPER does not place any conditions on the clustering algorithm (Singh, Carlsson 2007). All clustering algorithms require some sort of distance measure or metric to run the clustering algorithm. In the next section we will examine a few clustering algorithms that can be used to create a topological summary.

Before we examine the multiple distance functions we must give the following well known definition of a metric space.

Definition 2.2. Metric Space

A metric space (X, d) *consists of a non-empty set* X *and a function* $d : X \times X \to [0, \infty)$ *such that:*

i. (Positivity) For all $x, y \in X$, $d(x, y) \geq 0$ *with equality if and only if* $x = y$.

ii. (Symmetry) For all $x, y \in X$, $d(x, y) = d(y, x)$.

iii. (Triangle Inequality) For all $x, y, z \in X$ $d(x, y) \leq d(x, z) + d(z, y)$.

A function d *satisfying conditions (i)-(iii), is called a metric on* X.

Each of the following functions presented satisfy conditions (i)-(iii) of Definition 2.2 and are therefore considered metrics on X, the point cloud data.

Metrics

In this section we are going to consider several metrics useful for clustering. This is by no means a comprehensive list of metrics; however, the metrics given have been shown to be useful in a variety of applications.

Euclidean Metric: For some data set $X \in \mathbb{R}^n$ we let the metric

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where x and y are points in X.

Manhattan Metric: For some data set $X \in \mathbb{R}^n$ we let the metric

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

where x and y are points in X.

Chebyshev Metric: For some data set $X \in \mathbb{R}^n$ we let the metric

$$d(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$$

where x and y are points in X. This is also referred to as the maximum metric.

Example 3. Suppose we have two vectors:

$$\overline{\mathbf{x}} = [-0.68, -0.85, -0.42]$$
$$\overline{\mathbf{y}} = [-1.00, 2.31, -0.85]$$

sampled from the torus in Example 2. Applying the Euclidean (d_e), Manhattan(d_m), and Chebyshev (d_c)metric on x_1 and x_2 we see that

$$\begin{aligned} d_e(\overline{\mathbf{x}}, \overline{\mathbf{y}}) &= \sqrt{(-.68 + 1.00)^2 + (-0.85 - 2.31)^2 + (-0.42 + 0.85)^2} \\ &= \sqrt{10.27}. \\ d_m(\overline{\mathbf{x}}, \overline{\mathbf{y}}) &= |-.68 + 1.00| + |-0.85 - 2.31| + |-0.42 + .85| \\ &= 3.91 \\ d_c(\overline{\mathbf{x}}, \overline{\mathbf{y}}) &= \max(|-.68 + 1.00|, |-.85 - 2.31|, |-.42 + .85|) \\ &= 3.16. \end{aligned}$$

This is an example of two points sampled from a Torus X in $\mathbb{R}^3$; however in MAPPER, the metric is computed for all pairs of points. Therefore when determining the distance matrix for the torus with 1000 data entries we see there are

$$\frac{1000!}{2(1000-2)!} = 499,500$$

combinations.

We will use this idea of a metric when talking about clustering techniques in Section 2.3. We have shown in Example 3 that the Euclidean metric can be applied to a torus in $\mathbb{R}^3$; however, this can be expanded to $\mathbb{R}^n$. To learn more about distance metrics see (Rokach, Maimon, Chapter 15) for a thorough examination.

2.3 Clustering

Clustering in data analysis is the process of separating a set of points into subsets, or clusters, with similar characteristics. However, data points are not always uniformly similar; two points may have one attribute that is highly similar while another attribute is completely different. Therefore, clustering techniques are valuable in partitioning the dataset into subsets of similar points. We examine single-linkage clustering; however, other clustering techniques are useful in the MAPPER algorithm.

Single Linkage Clustering

Single-linkage clustering is a form of hierarchal clustering that starts with each data point in its own set and recursively joins similar points. The clustering of points are determined by a chosen metric. Points that are determined to be close by the metric that is chosen will be clustered in the same group. This process is repeated until all points have been grouped into one set

To begin, we choose a metric to use in the clustering algorithm for the data set X with m observations with n variables. We then partition the entries into a total of m subsets. That is, each set contains only one data entry. The distance between sets is determined by

$$D(A, B) = \min\{d(a, b) \mid a \in A,\ b \in B\} \tag{2}$$

where d is the metric and A, B are sets consisting of points from X. The two sets that contain points with the smallest distance are united. This process is continued until all points are combined into one set.

Example 4. Take the small data set of 6 points taken randomly from a circle, seen in Figure 6. In this example we used the euclidean metric to measure the distance between each point.

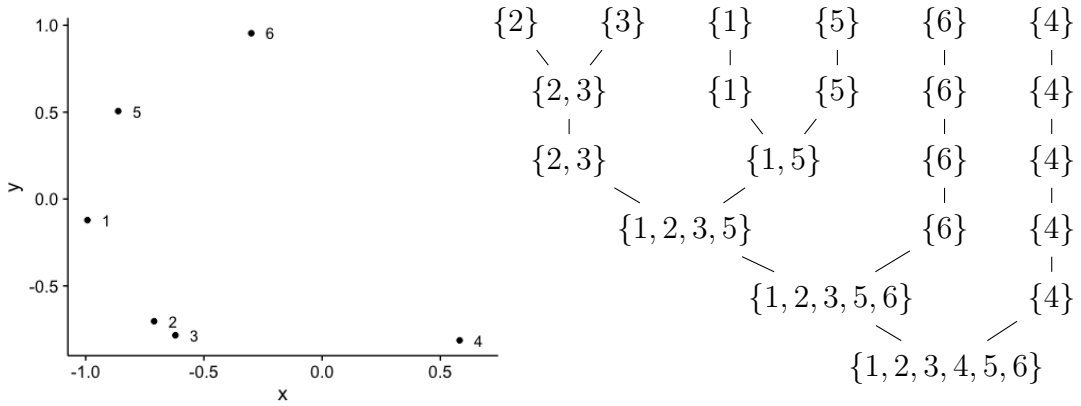

Figure 6: *left* points in $\mathbb{R}^2$. *right* Single Linkage Clustering Process

We see from Figure 6 that the two points with the shortest distance are points 2 and 3. Therefore using single linkage we union the two sets together. Next, points 1 and 5 are the next minimum distance so we union those two sets together. At this stage the sets containing $\{2, 3\}$, and $\{1, 5\}$ have values which have the minimum distance, that is points 1 and 2. Therefore we merge the two sets. We continue this process until every point is contained in one set.

We have given a brief example of single linkage clustering in Example 4; however, there are many other clustering algorithms such as complete linkage and average linkage that can be used as well. It is also important to note when the user chooses the clustering technique for MAPPER, the number of intervals are also chosen by the user. For a more comprehensive study on the subject we refer the reader to (Tan et. al 2015).

3 MAPPER Description

Now that we have a more comprehensive exploration of the parameters let's reconsider the algorithm. A formal definition is given in Definition 3.1; however, for the scope of this paper we will not go into finite covers, simplicial complexes or topological spaces and instead point the reader to (Carlsson, 2009) or (Kraft, 2016) to learn more.

Definition 3.1. MAPPER (Dey et. al 2015)
Let X and Z be topological spaces and let $f : X \to Z$ be a well-behaved and continuous map. Let $U = \{U_\alpha\}_{\alpha \in A}$ be a finite open cover of Z. The MAPPER construction arising from these data is defined to be the nerve simplicial complex of the pullback cover: $M(U, f) := N(f^*(U))$.

Assume we have some dataset X generated by a torus, as in Example 2. We first decide the filter function that will take $X \subseteq \mathbb{R}^3$ to $\mathbb{R}$. Next we apply the filter function on the data set, creating a new array of points, call it $\overline{\mathbf{v}}$. We then apply the paramaters from Definition 1.1. In addition, we determine the number of intervals when filtering and the percent overlap of each interval. The chosen clustering technique is then applied to each interval from $\mathbf{v}$. When running the clustering algorithm we must determine how many iterations until the process is terminated. Too many will result in just a single node while too little will result in many tiny nodes with few data points. The percent overlap will determine the edges in the connected graph from the output of MAPPER. If two or more more clusters share a data entry in common then the nodes are connected.

4 Applications of MAPPER

4.1 Boston Housing Data

We now examine the effectiveness of MAPPER on a known data set. The data set being analyzed originates from [12]. In the Boston Housing Data set there are a total of 506 entries with 14 variables, which can be seen in Table 1. In (Harrison & Rubinfield, 1978), they used the data to estimate how willing people are to pay for air quality improvements. In this analysis we are going to use the data with a different objective. That is, we are going to determine the per capita crime rates using all the variables in Table 1. We will begin by looking at the data using statistical techniques such as linear regression and exploratory data analysis. We will then examine the effectiveness of TDA on the same data set using all the predictor variables.

	Variable	Description
1.	CRIM	per capita crime rate by town
2.	ZN	proportion of residential land zoned for lots over,25,000 sq.ft.
3	INDUS	proportion of non-retail business acres per town
4.	CHAS	Charles River dummy variable (= 1 if tract bounds,river; 0 otherwise)
5.	NOX	nitric oxides concentration (parts per 10 million)
6.	RM	average number of rooms per dwelling
7.	AGE	proportion of owner-occupied units built prior to 1940
8.	DIS	weighted distances to five Boston employment centres
9.	RAD	index of accessibility to radial highways
10.	TAX	full-value property-tax rate per $10,000
11.	PTRATIO	pupil-teacher ratio by town
12.	B	1000(Bk - 0.63)ˆ2 where Bk is the proportion of blacks,by town
13.	LSTAT	% lower status of the population
14.	MEDV	Median value of owner-occupied homes in $1000's

Table 1: Variables collected in the Boston Housing Dataset

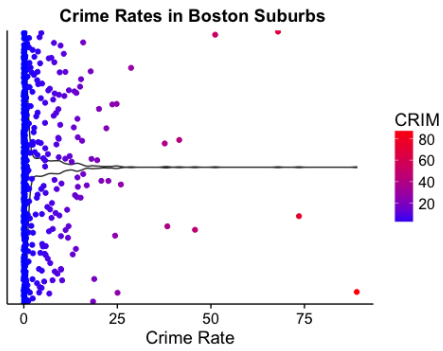

Figure 7: Violin Plot of Crime Rate

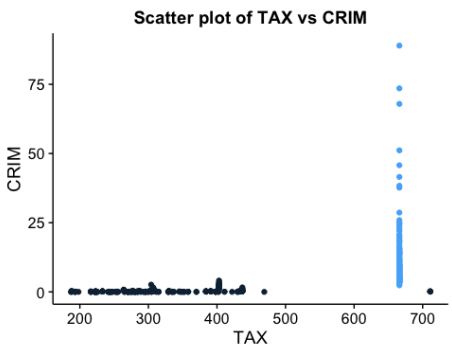

Figure 8: CRIM colored by TAX = 666

4.1.1 Statistical Analysis

Before beginning any sort of statistical analysis it is necessary to examine the data through exploratory data analysis (EDA). This is useful to find general trends in the data that otherwise would have gone unnoticed. Since we are interested in predicting crime rate, we will especially focus to understand this variable and its interactions with the predictors.

From the EDA we were able to gain some useful insights about the data. First, we see that crime rate is skewed right, therefore most of the neighborhoods from the Boston housing data have low crime rates (see Figure 7). To see how the predictors relate to crime rate, a correlation test was run on the data. We determined that the predictors RAD, and TAX are highly correlated with crime rate and may be useful when generating the model. However, after closer examination it was determined that high crime rates corresponded to a $TAX = 666$, see Figure 8. We suspect that this is a defect of the data collection and may be a default value. The data was clustered into two new data tables one in which corresponded to TAX values equal to 666 and TAX values not equal to 666.

We ran a new linear regression model on each of the new datasets. First, we examined the data set with TAX values of 666. This data set consisted of 132 observations with 13 variables. We used forward and backward selection techniques

and determined the adjusted R^2, and BIC values corresponding to the appropriate model. We determined that

$$\log(CRIM) = 5.08126 - 1.16487 \cdot \log(DIS) - 0.75396 \cdot \log(MEDV) \tag{3}$$

is an appropriate model when that TAX is equal to 666. From equation (3) a 1% increase in DIS, distance for radial highways decreases $CRIM$, crime rate in neighborhood by 1.16%, and a 1% increase in $MEDV$, median value of home decreases $CRIM$ by .75396%. The corresponding adjusted R^2 value for (3) is 0.5536. This model was then verified and plotted against the actual values, see Figure 9. A similar analysis was run on the neighborhoods where TAX was not equal to 666. This data set contained 374 observations with 13 variables. Again, we use forward and backward selection techniques and determined the adjusted R^2 and BIC values corresponding to the appropriate model. For this dataset we determined that

$$\log(CRIM) = 1.38009 - 0.87270 \cdot NOX^{-2} + .28192 \cdot \log(DIS) \tag{4}$$

is a sufficient model given that TAX is not 666. From equation 4 a 1% increase in DIS increases CRIM by .28% and as NOX, nitric oxide concentrations decrease CRIM decreases proportionally. The corresponding adjusted R^2 value for (3) is .5391. The fitted values of equation (4) were plotted against the transformed CRIM values, see Figure 9.

From this statistical analysis it appears that the models are valid and are sufficient at predicting crime rates in communities around Boston. We saw that EDA was a important step in the analysis of the data set as it revealed that some clustering of data points was necessary before a valid model was created. Further the TAX value of 666 seemed to either be a default data entry or corresponded to a distinct subsection of Boston neighborhoods.

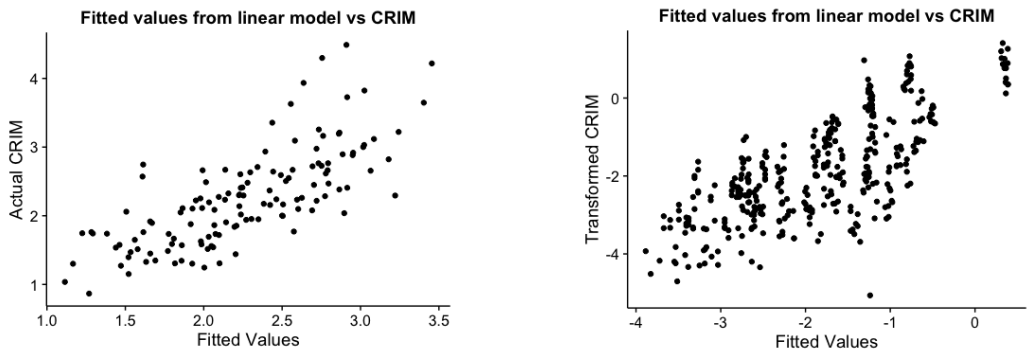

Figure 9: *left*: Actual vs fitted model from equation (3). *right*: Actual vs fitted model from equation (4)

Next, We are going to use the TDA MAPPER algorithm on the Boston Housing Data set and examine its effectiveness.

4.1.2 Topological Data Analysis

We will now analyze the data set using every predictor presented in Table 1; we will also include the outcome variable CRIM in the model. For the MAPPER parameters we used the Chebyshev metric with single linkage cluster and SVD as the filter function. We set the number of intervals to 7 and overlap was set to 40% with the number of bins per cluster set to 7.

From the parameters described above, a topological summary was generated with eleven nodes. The topological summary had a flare of three, a loop of four, and a small flare of two. To understand the relation between the fourteen variables we colored the graph by the mean value of each variable independent of all other variables. We should note a few properties about the topological summary: first, the size of the node corresponds to the number of observations contained in the node. Second, the color of the node represents the mean value of all entries in that node with respect to the chosen variable.

To get a better understanding of the graph generated from MAPPER we begin by coloring the topological summary by mean crime rate, see Figure 10. From this

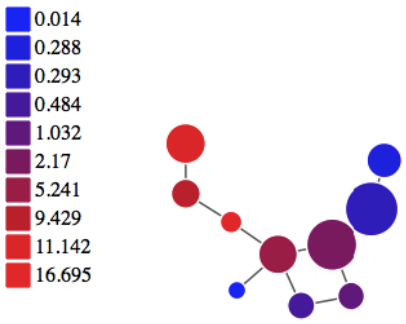

Figure 10: Topological summary of colored by crime rate.

coloring we see that their is an unusually high crime rate in the flare of three. It also appears that the mean crime rate decreases from left to right when looking at Figure 10. To see if there is a relationship between crime rate and any other variable we colored by the other fourteen variables as a form of exploratory data analysis. We saw that crime rate seemed to be somehow related with the variables DIS, MEDV, and AGE, the proportion of owner occupied homes built prior to 1940, see Figure 11. Examining this trend more closely we see that as crime rate increases it appears that the distance from radial highways decreases and the value of the home decreases.

Next, we evaluated the flare of three by running a regression on the points within each node. It turned out to be nearly identical to the statistical analysis run in Section 4.1.1. Moreover, when running regression on the flare of three the predictor variables, DIS and MEDV were identical to that in model (3). It was also noted that as the proportion of owner occupied units built prior to 1940 increased so did the crime rate. We determined after running the analysis on the flare of three that the appropriate model is

$$\log(CRIM) = 2.14662 - 1.08057 \cdot \log(DIS) - 0.73828 \cdot \log(MEDV) \qquad (5)$$

with a corresponding adjusted R^2 of 0.559.

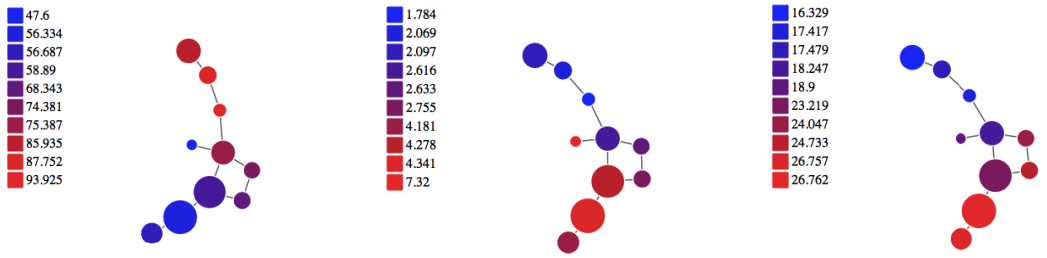

Figure 11: *left*: Colored by AGE. (3). *center*: Colored by DIS. *right*: Colored by MEDV. (4)

4.1.3 Housing Conclusion

The housing dataset was an interesting data set to analyze due to its repeated values in the variable TAX; which were highly correlated with higher CRIM rates. Through EDA we were able to filter the TAX values by evaluating the scatterplot of CRIM vs TAX. From this we were able to generate two linear models after filtering by TAX. We also saw that TDA was a useful tool at filtering the data set. In addition we were able to gain insights about the data that were not obvious through traditional regression techniques, that is it appeared that AGE was related to DIS, MEDV, and CRIM.

4.2 Wisconsin Cancer Data

Topological Data Analysis and MAPPER has been shown to be useful in some binary variables. The most well known study and breakthrough for TDA was done by Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. In this study they were able classify a subgroup of breast cancer patients that survived with 100% accuracy (Nicoulau, Levine, Carlsson 2011). We implemented a similar analysis on the Breast Cancer Wisconsin data set. The data consisted of 569 variables observations with 11 variables. The goal of this study is to understand how the 11 predictor variables relate with the outcome variable diagnosis. The variables in the data set are listed in Table 2. We begin in a similar manner as Section 4.1.1. We are going to examine

	Variable	Description
1.	Diagnosis	(M = malignant, B = benign)
2.	Radius.mean	mean of distances from center to points on the perimeter
3	texture.mean	standard deviation of gray-scale values
4.	perimeter.mean	perimeter of tumor
5.	area.mean	area of tumor
6.	smoothness.mean	local variation in radius lengths
7.	compactness.mean	$(perimeter^2/area - 1.0)$
8.	concavity.mean	severity of concave portions of the contour
9.	concavepoints.mean	number of concave portions of the contour
10.	symmetry.mean	symmetry of tumor
11.	fractaldimension.mean	"coastline approximation" - 1

Table 2: Variables collected in the Breast Cancer Wisconson Dataset

the data set through EDA and statistical techniques.

4.2.1 Statistical Analysis

For this analysis it is appropriate to use logistic regression techniques to determine the probability of having a malignant tumor. We examined the variables that appeared to be highly correlated to having a malignant tumor. There were four predictor variables that appeared to be most related to determining the correct diagnosis: mean radius, mean perimeter, mean area, and mean concave points. Two of the box-plots, area.mean and concave.mean can be seen in Figure 12.

Before starting the statistical analysis the outcome variable diagnosis was converted to a dummy variable where Malignant = 1 and Benign = 0. We use forward and backward elimination techniques to determine the appropriate logistic regression model. We determined the logistic regression model to be,

$$P = \frac{1}{1 + e^{-(-16.7481 + 101.6037x_1 + 0.3255x_2 + 0.0078x_3)}} \tag{6}$$

where $x_1 = concavepoints.mean$, $x_2 = texture.mean$, $x_3 = area.mean$. The residual deviance and null deviance for model (6) were 161.70 and 751.44 respectively. It is

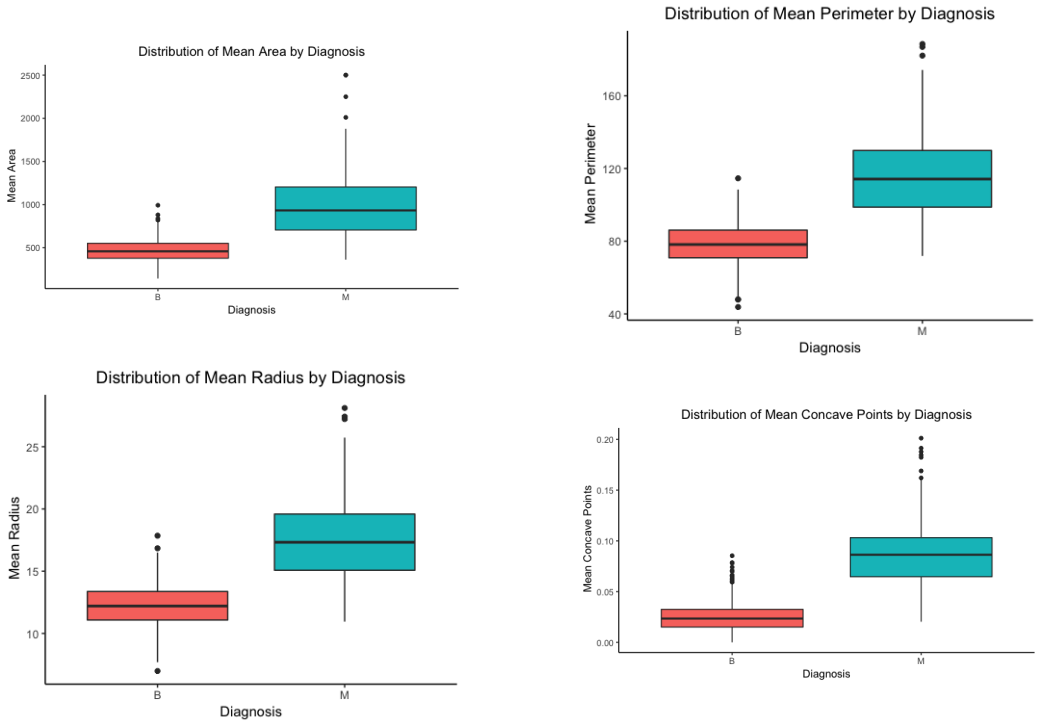

Figure 12: *Top from left to right:* boxplot for mean area of tumor split by diagnosis. boxplot for mean perimeter of tumor split by diagnosis. *Bottom from left to right:* boxplot for mean radius of tumor split by diagnosis. boxplot for mean concave of tumor split by diagnosis

important to note that logistic regressions do not have an R-squared value; however, an analogous computation can be done. An appropriate comparison to R-squared for the logistic regression model in equation (6) is, $1 - \frac{161.70}{751.44} = .785$.

Due to the importance of diagnosing a patients tumor correctly we examined the possibility of type 1 and type 2 errors, or false positive and false negative respectively. More specifically it could be detrimental to the patient if a malignant tumor goes untreated due to an analysis suggesting it is benign. Therefore a confusion matrix was generated depicting the type 1 and type 2 errors that occurred in the analysis. Three confusion matrices were generated to compare the different cutoff values of determining if a patients tumor is benign of malignant. The cutoff values were determined from the values assigned in equation (6) and compared against the actual diagnosis of the tumor. The three values examined were .5, .25, and .05, see Table 3. Looking at

		Predicted Value					
		$P < .5$	$P \geq .5$	$P < .25$	$P \geq .25$	P$< .05$	$P \geq .05$
True	Malignant	21	191	13	199	2	210
Diagnosis	Benign	345	12	322	35	278	79

Table 3: Confusion table with true diagnosis vs predicted diagnosis. The values in the table are the number of patients.

a cutoff value of .5 we see that there is a $10\% = \frac{21}{212}$ chance of a false negative result. It is important to note that false positives are an area of concern; however, false negatives are of utmost concern as we don't want to send a patient home thinking they have a benign tumor when it is actually cancerous. If we set the cutoff value for the model to .05, that is if model (6) outputs anything greater than .05 we consider the patient for further testing. Even with this cutoff we still have 2 patients or $1\% = \frac{2}{212}$ that are misdiagnosed. In addition, at this cutoff value the error rate is $\frac{81}{569} \approx 14.2\%$

Logistic regression seems to be an efficient way for determining benign and malignant tumors. However, misdiagnoses are still an issue and can be a problem when diagnosing a patient. In addition we are unable to determine if the interaction between variables affects the type of tumor. Next we will look at determining the type of tumor through a different lens, TDA.

4.2.2 Topological Data Analysis

After looking at the predicted value versus the true diagnosis there were still a few false negatives even with a small cutoff value (P = .05). In addition, the logistic regression model only took into account three variables: concavepoints.mean, texture.mean, and area.mean. Therefore we are going to examine the data set with MAPPER using the Chebyshev metric with single linkage cluster and SVD as the filter function. We set the number of intervals to 10 and overlap was set to 10% with the number of bins per cluster set to 10.

From the topological summary from MAPPER we began by coloring the nodes

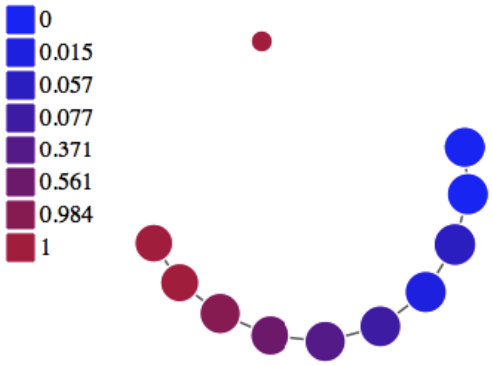

Figure 13: Topological Summary colored by diagnosis

by the mean outcome variable; diagnosis. Therefore the nodes with a mean value of 1 are all malignant tumors and the nodes with a mean value of 0 are all benign tumors. From Figure 13 we see there is a clear pattern in the distribution of patients in each node. The node on the left has a higher percentage of people with malignant tumors, the two leftmost nodes have patients with only malignant tumors and the two rightmost nodes have patients with only benign tumors.

From first appearances it appears that TDA is capable of sorting people with benign and malignant tumors. The two left nodes account for 100 of the 569 patients and the two right most nodes account for 136 of the 569 patients. This accounts for a little over 40% of the data set that was properly diagnosed. Next we examine the factors that are contributing to whether a person is sorted in the far left nodes or far right nodes. After coloring the graph we saw there were three variables that seem to be highly correlated with the diagnosis of the tumor. We saw that the radius, perimeter, and area of the tumor followed the same pattern as the topological summary colored by diagnosis, see Figure 14.

Even though the topological summary seemed to follow a pattern for the two variables area, and perimeter, we see smoothness appeared to be more disordered and followed no discernible pattern as in Figure 14. More-so we saw that two of the

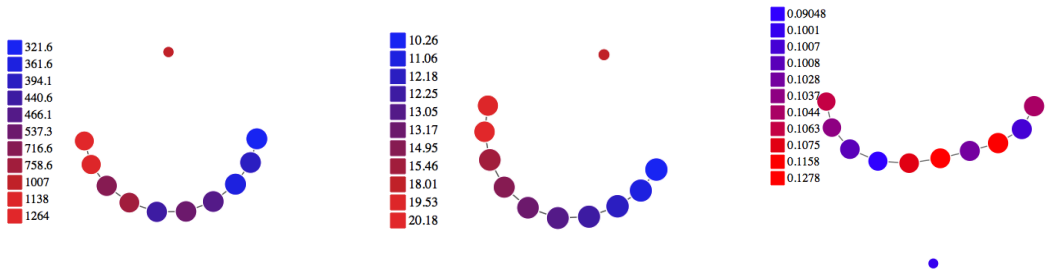

Figure 14: *left to Right:* Topological Summary for mean area, perimeter, and smoothness.

nodes in the center were unable to determine the diagnosis of the tumor. From Figure 13 two of the nodes have a mean diagnosis value of 0.371, node 6 and 0.561, node 7. The diagnosis in both nodes were nearly evenly split and it was difficult to determine which patient will have a benign or malignant tumor. Therefore another logistic regression was run on nodes six and seven to see if we could predict the diagnosis.

Following TDA we used an exhaustive search, which determine the variables that contribute most favorably to the outcome variable. The exhaustive search determines an appropriate model for nodes six and seven independently. The logistic regression model for node six is

$$P = \frac{1}{1 + e^{-(-11.9321 - 1.4998x_1 + 0.5179x_2 + 148.7470x_3)}}, \tag{7}$$

where $x_1 =$ radius_mean, $x_2 =$ texture_mean, and $x_3 =$ smoothness_mean. The logistic regression model for node seven is

$$P = \frac{1}{1 + e^{-(-35.5810 + 75.6029x_1 + 0.7425x_2)}}, \tag{8}$$

where $x_1 =$ concavity_mean, and $x_2 =$ texture_mean.

We saw that texture smoothness and radius are effective predictors for patients in node 6 while concavity and texture are effective for node 7. From the topological summary smoothness appeared to be much different in node 6 and node 7. We can

see this from the variation between the blue and red color assigned to each node. Following the same process as Section 4.2.1 we plot the fitted vs actual values from models (7), and (8). As we saw in the statistical analysis section logistic regression was prone to false negatives and false positives. However, after running logistic regression again on the new data sets, that is nodes 6 and 7; we saw that the cutoff values were much larger. Setting P = .15 from equation (7) we were able to classify every malignant tumor in node six. Setting P=.10 from equation (8) we were able to classify the malignant tumors in node seven. However, what is more important is it appears there is a subset of patients that share similarities in many attributes such as mean area and perimeter, but differ wildly in the smoothness of the tumor, and this observation leads to a different model and consequently a different diagnosis.

5 Conclusion

TDA is a beneficial tool when trying to visualize high dimensional data sets. We saw a subset of patients that seemed to be related in many attributes, but differed in another, which resulted in a different diagnosis. In addition, TDA sorted neighborhoods around Boston from low to high crime rates. With this filtration we could then compare what variables seemed to be related with high, or low crime rates. We also noticed TDA can be useful as an initial filter function to create subsets of the original dataset that can later be used with traditional statistical techniques.

5.1 Strengths

The purpose of TDA is to give data shape and a gain insight from this shape. Therefore TDA, and more specifically MAPPER is useful in summarizing large datasets into a graph that can be further examined. It's beneficial in that it allows for a visualization of variables that otherwise may be left unnoticed through statistical

techniques.

5.2 Weaknesses

As we discussed earlier, MAPPER has several inputs that contribute to the overall summary of the data. Slight changes in the MAPPER inputs may result in a different summary then the previous one, and therefore different conclusions from the same data set may arise. However, these weaknesses in MAPPER can be addressed by understanding how the MAPPER inputs affect the overall summary.

5.3 Further Research

The MAPPER inputs, percent overlap, clustering algorithm, filter function, and number of intervals are well understood; however, further research on how the inputs interact would be useful. We also suggest researching different filter functions as SVD was the primary interest in this paper. Finally, examining whether alterations of the current MAPPER algorithm is beneficial given different scenarios.

References

[1] Muller, Neil, Loureno Magaia, and B. M. Herbst. "Singular Value Decomposition, Eigenfaces, and 3D Reconstructions." SIAM Review 46.3 (2004): 518-45. Web. 16 Mar. 2017.

[2] Singh, Gurjeet, and G. Carlsson. Algorithms for Topological Analysis of Data. Diss. Standford U, 2007.

[3] "Metric Spaces." (2006): 1-2. Metric Spaces. University of Oslow. Web. 19 Mar. 2017.

[4] Maimon, Oded, and Lior Rokach. Chapter 15 CLUSTERING METHODS. Department of Industrial Engineering Tel-Aviv University. Web. 20 Mar. 2017. https://www.cs.swarthmore.edu/ meeden/cs63/s16/reading/Clustering.pdf.

[5] Nicolau, M., A. J. Levine, and G. Carlsson. "Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival." Proceedings of the National Academy of Sciences 108.17 (2011): 7265-270. Web. 12 Jan. 2017.

[6] KRAFT, RAMI. Illustrations of Data Analysis Using the Mapper Algorithm and Persistent Homology. Diss. Royal Institute of Technology, 2016.

[7] "Topology." Oxford Dictionaries. Oxford Dictionaries, Web. https://en.oxforddictionaries.com/definition/topology.

[8] Dey, Tamal K., Facundo Mmoli, and Yusu Wang. "Multiscale Mapper: Topological Summarization via Codomain Covers." Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (2015): Web.

[9] Carlsson, Gunnar. "Topology and Data." Bulletin of the American Mathematical Society 46.2 (2009): 255-308. Web. 8 Feb. 2017.

[10] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. Dorling Kindersley: Pearson, 2015. Print.

[11] Ayasdi. "The Leader in Topological Data Analysis." Ayasdi. Ayasdi, Web. 11 Apr. 2017. https://www.ayasdi.com/company/.

[12] Harrison, D. and Rubinfeld, D.L. "Hedonic prices and the demand for clean air", J. Environ. Economics & Management, vol.5, 81-102, 1978.