TOPOLOGICAL DATA ANALYSIS
AND THE MAA NATIONAL STUDY OF COLLEGE CALCULUS

by

IRMA EMMA STEVENS

Under the Direction Of:

Noah Giansiracusa

Abstract

In this study, topological data analysis techniques were used to analyze a discrete dataset. The dataset is the results of a national study done on college calculus students in the United States. The goal of the study was to create and implement techniques to analyze the data by considering the persistent homologies up to the second dimension. The resulting persistence diagrams were then interpreted by considering the context of the data analyzed, and in some cases, these results were compared to the findings in the report on the calculus study.

Index words: Topological Data Analysis, Persistent Homology, Mathematics Education, Calculus Students

TOPOLOGICAL DATA ANALYSIS
AND THE MAA NATIONAL STUDY OF COLLEGE CALCULUS

by

IRMA EMMA STEVENS

B.S., The University of North Carolina at Charlotte, 2013

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the Requirements for the Degree

MASTER OF ARTS

Athens, Georgia

2016

TOPOLOGICAL DATA ANALYSIS
AND THE MAA NATIONAL STUDY OF COLLEGE CALCULUS

by

IRMA EMMA STEVENS

Major Professor:	Noah Giansiracusa
Committee:	Jason Cantarella Edward Azoff

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

Topological Data Analysis and the MAA National Study of College Calculus

Irma Emma Stevens

May 2, 2016

Acknowledgments

The dataset under consideration for this study is from the Characteristics of Successful Programs in College Calculus (CSPCC) of the Mathematical Association of America (MAA), David Bressoud, PI. This material is based upon work supported by the National Science Foundation under grant DRL REESE #0910240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Thank you to my committee members. Thank you Dr. Noah Giansiracusa for supporting me throughout this process of attempting to merge two of my passions, mathematics and mathematics education. Thank you for pushing me to learn and for guiding me along the way. Thank you to Dr. Jason Cantarella for your conversations that sparked many great ideas and for keeping me passionate about the work that I am doing. Thank you Dr. Edward Azoff for your continual support in both my mathematics and mathematics education programs. I always look forward to hearing your thoughts about what I am up to and listening to your stories about academia.

Thank you to my family. Thank you, Dad, for always making me feel like I can accomplish anything I set my mind to do. Thank you, Mom, for being an incredible source of comfort and support when whatever I set my mind to do becomes difficult. Thank you to my favorite brother for reminding me not to take life too seriously.

To avoid missing anyone, a big thank you to all my friends who have been there for me. Whether it was helping me through homework assignments and projects, visiting me from other states, or offering words of support and encouragement, I am incredibly grateful for you all. I could not have accomplished this goal without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data of various kinds are being produced at an unprecedented rate. However, obtaining data is only the first part of the process of research. The next step is to analyze the data. Over the past few decades, various powerful statistical methods have been used to sort through and determine the significant components of datasets. In this thesis, I will explore a newly developing way of analyzing data, topological data analysis. According to Offroy and Duponchel [16], topological data analysis used in spectroscopy is capable of detecting sup-populations which are not observed with PCA or HCA), is less sensitive to noise, spectral resolution and spectral shift, and can handle large data sets. The tool has shown promise in analyzing data sets that have very low signal to noise ratio, variable shifts and missing data. The dataset I will be considering comes from a large scale study across the United States on college Calculus students. I will provide a summary of the Calculus report and its findings, outline the goals, theory, and some of the previous research done using topological data analysis, explain the methods that were used to adapt the dataset so that it would be suitable for running TDA, describe the results of the analysis, and provide some future directions for research.

This study aims to address the following research questions:

(1) How can topological data analysis inform analysis of a discrete dataset? What adaptations to datasets need to be made and what patterns emerge in the resulting analysis of the dataset?

(2) How can 3-D plots of variables aid in the interpretation of persistence diagrams produced from three variables? What conclusions can be made about Calculus I students by considering these persistence diagrams and their respective 3-D plots, and how do these results compare with previous findings of the dataset?

(3)What considerations need to be made when using high dimensional input for topological data analysis?

Chapter 2

MAA Calculus Report Data and Analysis

2.1 The Dataset

The dataset under consideration for this study is from the Characteristics of Successful Programs in College Calculus (CSPCC) of the Mathematical Association of America (MAA), David Bressoud, PI. The goal of the report was (1) to establish a "base of knowledge of who takes Calculus I and why, what their preparation has been, what they experience in the classroom, and how this affects their confidence, enjoyment of mathematics, and intention to persist in the study of mathematics" and (2) to identify institutional practices that contribute to the retention of STEM students [4, p. v].

The data collection occurred in two phases, the first of which is the focus of the MAA Calculus report considered here. In 2010, the researchers involved in the study selected a stratified random sample of non-profit colleges and universities offering a degree in mathematics (Associate's, Bachelor's, Master's, or Doctoral). The institutions were stratified by highest degree offered by the department and size of the undergraduate population. Over the summer and fall of 2010, the researchers contacted the selected institutions' mainstream Calculus I coordinators, where "Mainstream Calculus I" is defined as any first course in calculus that can be used as part

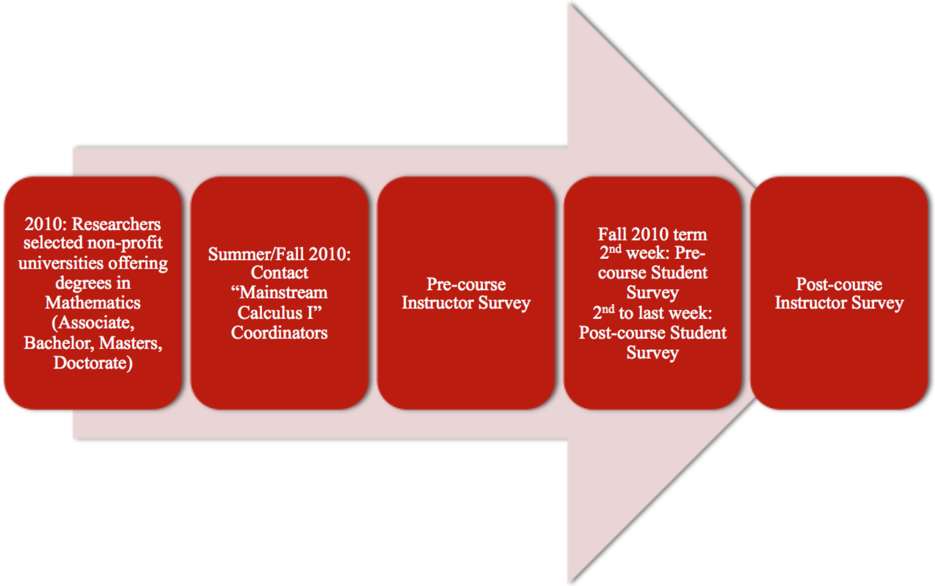

2010: Researchers selected non-profit universities offering degrees in Mathematics (Associate, Bachelor, Masters, Doctorate)
Summer/Fall 2010: Contact "Mainstream Calculus I" Coordinators
Pre-course Instructor Survey
Fall 2010 term 2nd week: Pre-course Student Survey 2nd to last week: Post-course Student Survey
Post-course Instructor Survey

Figure 2.1: CSPCC data collection timeline.

of the calculus prerequisite for higher level mathematics courses [4, p. vi]. The coordinators gave basic information about the course and the contact information of the institutions' instructors of the identified course. The instructors were surveyed before and immediately after their first fall term, and the students in the courses were surveyed in the second and the second-to-last weeks of class (Figure 2.1). This data was organized into a file and had identifiers removed; the resulting file is the dataset used for this study.

According to the 2010 Conference Board of the Mathematical Sciences (CBMS) [3], there were about 300,000 mainstream calculus students across the United States (110,000 from PhD-granting universities, 41,000 from MA-granting universities, 82,000 from BA-granting four-year colleges, and 65,000 from AS-granting two-year colleges) [4, p. 1] in 2010 (Figure 2.2). These numbers are the most recently updated as of April 2016.

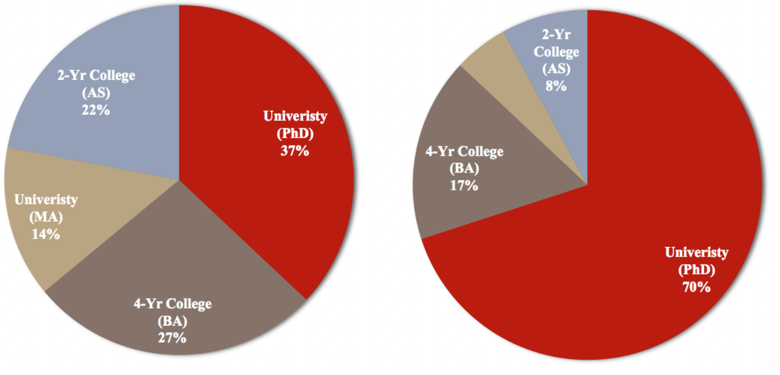


Figure 2.2: (left) CBMS 2010 report on distribution of Calculus I students by institution type [3] (right) CSPCC student participants.

In comparison, the CSPCC collected data from 10,114 mainstream Calculus I students (7,086 from PhD-granting universities, 535 from MA-granting universities, 1,742 from BA-granting four-year colleges, and 751 from AS-granting two-year colleges (Figure 2.2) from 213 colleges and universities [4, p. 135]. Note, however, that the percentages reported by the MAA Calculus Report were first calculated for each institution type and then combined with a weighted average determined by the number of Calculus I students at each type of institution [4, p. 1]. The CSPCC information reported on in this report comes from the 7,260 students who responded to both the start and end of term surveys [4, p. 136] (Figure 2.3).

Overall, five major surveys were constructed: one for the coordinators, two for the instructors (pre- and post-course), and two for the students (pre- and post-course). Based on their survey of the literature, the researchers decided to analyze six dependent variables: confidence, enjoyment, desire to continue studying mathematics, intention to continue calculus, increased interest (end of term survey only), and final grade of C or higher (end of term survey only) [4, p. 134]. Confidence, enjoyment, and increased interest were assessed using a Likert scale (e.g., 0=Strongly Disagree;

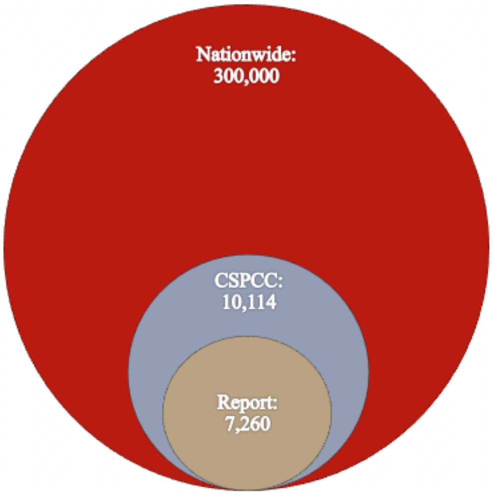


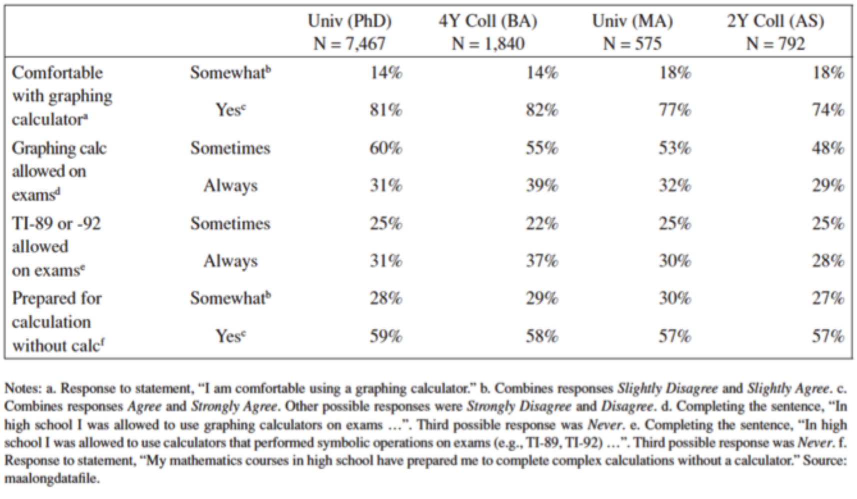
Figure 2.3: Calculus I students nationwide, in the study, and included in the CSPCC (not to scale).

1= Disagree; 2=Slightly Disagree; 3= Slightly Agree; 4=Agree; 5=Strongly Agree). All of the survey questions received answers that were discrete in nature (e.g., yes/no, select all that apply, and letter grade).

2.2 Analysis

This section describes the methods of analysis used on this data in order to provide insight into the research that has already been done on this particular dataset. Furthermore, examples of the results of these forms of analysis provide insights into the findings of the study, some of which will be considered when interpreting the results of the topological data analysis.

Table 2.1: CSPCC report of students' survey responses on graphing calculator use.



		Univ (PhD) N = 7,467	4Y Coll (BA) N = 1,840	Univ (MA) N = 575	2Y Coll (AS) N = 792
Comfortable with graphing calculator[a]	Somewhat[b]	14%	14%	18%	18%
	Yes[c]	81%	82%	77%	74%
Graphing calc allowed on exams[d]	Sometimes	60%	55%	53%	48%
	Always	31%	39%	32%	29%
TI-89 or -92 allowed on exams[e]	Sometimes	25%	22%	25%	25%
	Always	31%	37%	30%	28%
Prepared for calculation without calc[f]	Somewhat[b]	28%	29%	30%	27%
	Yes[c]	59%	58%	57%	57%

Notes: a. Response to statement, "I am comfortable using a graphing calculator." b. Combines responses *Slightly Disagree* and *Slightly Agree*. c. Combines responses *Agree* and *Strongly Agree*. Other possible responses were *Strongly Disagree* and *Disagree*. d. Completing the sentence, "In high school I was allowed to use graphing calculators on exams …". Third possible response was *Never*. e. Completing the sentence, "In high school I was allowed to use calculators that performed symbolic operations on exams (e.g., TI-89, TI-92) …". Third possible response was *Never*. f. Response to statement, "My mathematics courses in high school have prepared me to complete complex calculations without a calculator." Source: maalongdatafile.

2.2.1 Percentages

Oftentimes, the authors reported percentages. For example, while 80% of 7,089 students in PhD-granting universities reported that their home supported their studying mathematics and 40% reported that no one encouraged them to study mathematics, 67% of 7,089 students in 2-year AS-granting colleges reported that their home supported their studying mathematics and 49% reported that no one encouraged them to study mathematics [4, p. 4]. The reported percentages are often displayed in a table format. For example, Table 2.1 shows the percentages of students reporting various aspects of graphing calculator use [4, p. 9]. Depending on the focus of the study, the percentages of different survey responses are split by gender, ethnicity, and/or type of institution.

2.2.2 Means and Standard Deviation

Means and standard deviation are a common way to report the results. For example, the average SAT Math score (and standard deviation) from University (PhD) to 4-year College (BA) to University (MA) to 2-year (AS) went from 663 (71) to 632 (72) to 616 (81) to 589 (95) [4, p. 5] (Figure 2.4). The means and standard deviations are reported by institution type for instructor-student reported graphing calculator use during class and class size (i.e., for the institutions selected for clinical interviews, four-year BA-granting colleges 25.46 (4.457) (N=14); MA-awarding universities 26.50 (5.334) (N=12); PhD-granting universities 54.70 (57.045) (N=47). In their analysis of the Calculus I Curriculum, the results focus primarily on the problem types faculty included on assignments and exams (e.g., [4, p. 48]). However, some means and standard deviations are reported on the survey questions. For instance, 421 faculty members have a mean of 4.90 (1.03) in response to the survey question: "My primary role as a Calculus instructor is to: 1 (work problems so students know how to do them) to 6 (help students reason through problems on their own)" and 90% of faculty responded 4 or higher [4, p. 49].

2.2.3 Frequency

Sometimes, only raw numbers are reported. For example, from the 4,828 students from PhD-granting universities who had taken Precalculus by 11^{th} grade, only 7 students did not go on to take calculus in high school [4, p. 6].

2.2.4 Factor Analysis

The most rigorous method of analysis used in the report was factor analysis. Factor analysis is a branch of multivariate analysis that assumes the approximate linearity

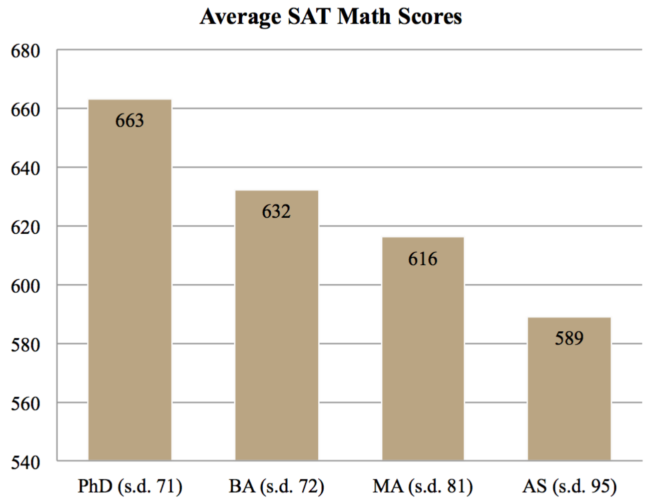


Figure 2.4: SAT Math scores for Calculus I students by institution type [4, p. 5].

of variables and then considers the relationships between the variables. According to Costello and Osborne [5, p. 17], factor analysis aims “to reveal any latent variables that cause the manifest variables to covary. During factor extraction, the shared variance of a variable is partitioned from its unique variance and error variance to reveal the underlying factor structure” [5, p. 2]. As a result, only shared variance is considered. In other words, let X_1,. . . ,X_p be a set of p observed correlated variables, the goal of factor analysis is to account for the correlations using a smaller number, k, of hypothetical variables. In order to do this, the correlation matrix of the variables is compared to the unit diagonal matrix to see if any correlation exists. If a correlation between variables is found, the goal is to see if there exists an F_1 such that, when its effect is eliminated, the correlations between the X variables are zero. If not, the new partial correlations between the variables are considered, and the process continues until all partial correlations between the X variables are zero. Thus, unlike the independently observable variables from which they came, factors are an “unobservable variable derived from *internal* analysis of the X-variables themselves” [13, p. 7].

This analysis is in the section on the impact of instructor and institutional factors on students’ activities. The researchers used factor analyses to develop a composite measure (of “students’ attitudes towards mathematics, three composites of pedagogical features, and four composites of institutional characteristics” [4, p. 17]), and multivariate regression analyses were used to model the impact of the factors onto students’ attitudes while accounting for control variables.

Before conducting factor analysis, the researchers prepared the dataset. These researchers used the pre- and post- completed surveys from 3,103 students in 308 classrooms. According to the data, the variables indicating attitude (confidence, enjoyment, and persistence) were highly correlated (from r=0.52 to r=0.70) and so the authors decided to combine the three variables to create a dependent variable called “mathematics attitude” by standardizing each of the three variables from the post survey (because two survey responses used a 6-point scale and one used a 4-point scale), averaging them, adjusting the standard deviation of the resulting composite to the standard deviation of the initial mathematics attitude composite, and then centering it on the average decline in the mathematics attitude score between the beginning and end of the semester [4, p. 17]. By going through this standardizing, averaging, and re-standardizing process in the beginning-of-the-semester surveys as well, the researchers created the same composite of initial mathematics attitude. The results of the analysis show that, on average, student attitudes toward mathematics declined from beginning to end of a college calculus course [4, p. 18] (close to a third of a standard deviation).

The researchers controlled for student, instructor, and institutional level variables in order to account for alternative hypotheses by “using variables that control for differences in students, classroom, and institutions, especially those differences that do not represent decisions that can be made or conditions that can be modified" [4, p. 18]. They considered information such as students' grades in high school mathematics classes, the students' year of college (using dummy variables), size and status of the instructor, and three dummy variables representing institutional type. Recall that dummy variables are created in order to include qualitative (or categorical) variables by arbitrarily assigning numbers to the various levels of the qualitative variable. [14, p. 213]

The variables of interest were instructor pedagogy and institutional characteristics. Since a total of 121 survey items characterized these two categories, the researchers decided to use exploratory factor analysis to help group items that might be indicators of the same underlying feature. The results of the factor analysis reveal that student responses to questions about instructor characteristics factored into three clusters: Good Teaching, Technology, and Ambitious Pedagogy. Of these three, only Good Teaching had a positive effect on the change in students attitudes towards mathematics (a composite of three outcomes: mathematics confidence, enjoyment, and persistence) [4, p. 83]. A similar process was used for institutional characteristics that resulted in four different factors.

In order to explain the variance in students' attitude at the end of the semester, the researchers decided to employ hierarchical linear modeling (HLM) which considered course, instructional, and institutional levels [4, p. 25]. The following statements are two examples of conclusions drawn from both the resulting model. The main effects model indicated that students' initial attitudes "powerfully predicted their attitudes at the end of the semester" and male students scored significantly higher than females on the mathematics attitude composite [4, p. 26]. Furthermore, in the main effects model, Ambitious Teaching had a negative effect, although the relative effect size of Good Teaching far outweighed that of ambitious teaching. Characteristics of Ambitious Teaching include 14 survey items such as the use of group projects, the inclusion of unfamiliar problems both in homework and on exams, requirements for students to explain how they arrived at their answers, and a decreased reliance on lecture as the primary mode of instruction [4, p. 93], whereas Good Teaching come from the 22 survey items which contains questions about the instructor providing understandable explanations, making the students feel comfortable to ask questions, grading fairly, etc. [4, p. 21].

2.2.5 *t*-Tests

Some of the analysis done on the results of the survey was done using Welch's t-test. Welch's t-test is for cases in which the sample populations are both assumed to come from Gaussian populations but not necessarily have the same standard deviation. Since it is a t-test, it is used to test the hypothesis that two populations have the same mean. For example, according the CSPCC, "Calculus I students at BA-granting four-year colleges are significantly less concerned about finances than those at PhD-granting universities (Welch's t-test, p<0.001) whereas those at AS-granting two-year colleges are significantly more concerned about finances (Welch's t-test, p<0.001)" [4, p. 3].

2.2.6 Chi-Square Tests with *t*-Tests

Further analysis involved conducting either t-tests or Chi-square tests of independence to explore differences between faculty in institutions selected for the case study (N=104) and those not selected for case study (N=399). For instance, there iss a statistically significant difference in the two groups in their use of projects (χ^2(1, N=364)=7.54, p<.006). More specifically, 20% of faculty in selected institutions

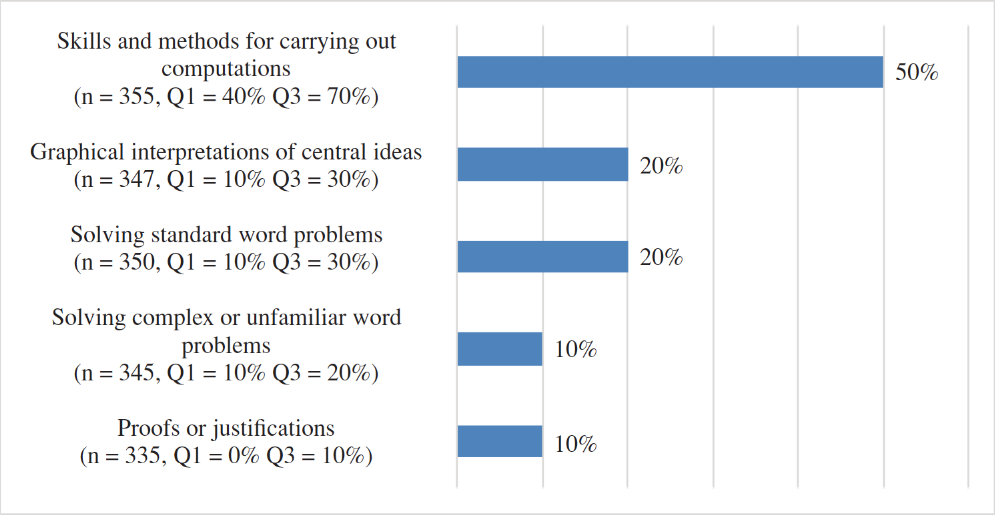


Figure 2.5: Median responses of faculty reported responses to the percentage (0%to 100%) of different problem types faculty include on assignments and exams [4, p. 48].

assigned two or more projects while only 9% of non-selected institutions' faculty did [4, p. 50].

2.2.7 Medians

The researchers also created Figure 2.5 to show the median responses to the percentage of different problem types faculty included on assignments and exams; they also included the first and third quarter percentile responses. The faculty included in this analysis are the 503 faculty who responded to pre- or post-course instructor surveys.

2.2.8 Intra-category Comparison T-tests

When analyzing academic and social supports, the researchers used intra-category comparison *t*-tests for the selected and non selected institutions for interviews. They found that none were significant at the .05 level; for example, selected and non-selected institutions were just as likely to offer tutoring by full time mathematics faculty [4, p. 70].

2.2.9 Cronbach's α

The mean, standard deviation, and reliability of the good teaching factors based on the results of the factor analysis were reported (Table 2.2). Reliability was measured using the Cronbach α. According to Tavakol and Dennick [17, p. 54], "The number of test items, item interrelatedness and dimensionality affect the value of α." Moreover, acceptable values of α range from 0.70 to 0.95. "A low value of α could be due to a low number of questions, poor interrelatedness between items or heterogeneous constructs [17, p. 54]". Tavakol and Dennick note that high α values may suggest that some items are redundant as they are testing the same question but in a different guise. They also report that a maximum α value of 0.90 has been recommended.

2.2.10 Limitations of Analysis

The CSPCC self-reported limitations of their study and these limitations should also be kept in mind when reading the results of the topological data analysis performed in this study. First, the data is self-reported. Second, several institutions had low response rates. Third, there is an over-representation of PhD institutions. Lastly, the researchers were hesitant to compare students' responses at the start and end of the term because "[s]tudents who answered the survey at the end of the term were, for the most part, those who had successfully negotiated the course" [4, p. 13]. According to the report, however, the only statistically significant difference between the responses before the term and after the term (only including students who responded to both surveys) was "in answer to the question about hours working

Table 2.2: Mean, standard deviation, and Cronbach α on good teaching factors.

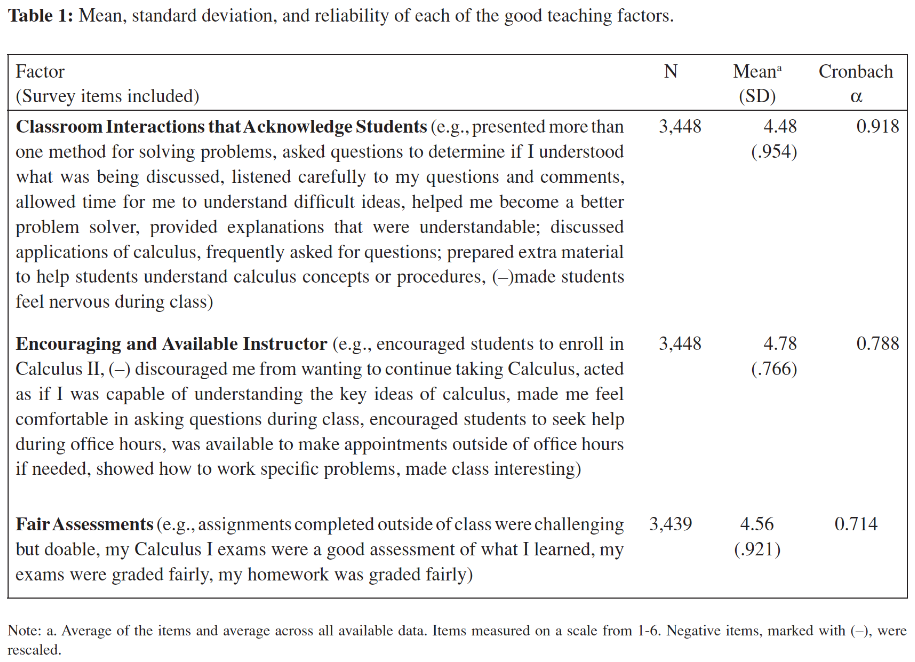


Table 1: Mean, standard deviation, and reliability of each of the good teaching factors.

Factor (Survey items included)	N	Mean[a] (SD)	Cronbach α
Classroom Interactions that Acknowledge Students (e.g., presented more than one method for solving problems, asked questions to determine if I understood what was being discussed, listened carefully to my questions and comments, allowed time for me to understand difficult ideas, helped me become a better problem solver, provided explanations that were understandable; discussed applications of calculus, frequently asked for questions; prepared extra material to help students understand calculus concepts or procedures, (–)made students feel nervous during class)	3,448	4.48 (.954)	0.918
Encouraging and Available Instructor (e.g., encouraged students to enroll in Calculus II, (–) discouraged me from wanting to continue taking Calculus, acted as if I was capable of understanding the key ideas of calculus, made me feel comfortable in asking questions during class, encouraged students to seek help during office hours, was available to make appointments outside of office hours if needed, showed how to work specific problems, made class interesting)	3,448	4.78 (.766)	0.788
Fair Assessments (e.g., assignments completed outside of class were challenging but doable, my Calculus I exams were a good assessment of what I learned, my exams were graded fairly, my homework was graded fairly)	3,439	4.56 (.921)	0.714

Note: a. Average of the items and average across all available data. Items measured on a scale from 1-6. Negative items, marked with (–), were rescaled.

at a job, and this only held among students at PhD-granting universities" [4, p. 13]. One of the solutions to this limitation was to corroborate some of the trends identified in the various forms of analysis through in-depth site visits to the selected institutions.

Chapter 3

Overview of Topological Data Analysis

This section providea an overview of topological data analysis and some of the previous analysis done on real-world systems. The main appeal of topological data analysis is its robustness against noise in data and its ability to be a coordinate-free method of analysis. This section describes some of the different ways to use and visualize the results of topological data analysis by considering examples in research.

3.1 Topological Data Analysis

One of the beauties of computational algebraic topology is its ability to provide insights into high-dimensional data. Data analysis in general generally aims to address two fundamental tasks: inferring higher dimensional structure from lower dimensional representations and assembling discrete points into a global structure [8, p. 61] According to Ghrist [8], researchers such as Carlson, de Silva, Edelsbrunner, Harer, Zomorodian, and others use the following principle themes:

> (1) It is beneficial to replace a set of data points with a family of **simplicial complexes**, indexed by a proximity parameter. This converts the data set into global topological objects.
>
> (2) It is beneficial to view these topological complexes through the lens

of algebraic topology – specifically, via a novel theory of **persistent homology** adapted to parameterized families.
(3) It is beneficial to encode the persistent homology of a data set in the form of a parameterized version of a Betti number: a **barcode**. (p. 61-62)

Keeping these principle themes in mind, the following discussion outlines how the CSPCC data was collected in a way that is suitable for topological data analysis. First, consider the dataset for this study: survey responses from calculus students, instructors, and coordinators. Consider this data as unordered sequences of points in a Euclidean n-dimensional space $\mathbb{E}^n$. The goal of topological data analysis is to see if the global "shape" of this data can provide some insight into information about the underlying phenomena which the data represents [8, p. 62].

Consider the case where n=3 for this dataset. One example of this case would be choosing to represent each individual student's response to three different survey questions as a point in 3-D space. Such an example is called point cloud data and is easily representable using 3-D plotters on a computer. In this setting, noting some of the global features of the data is still possible, and this situation will be explored in more detail in Chapter 5. However, this study is also interested in the question of how to analyze the data in higher dimensions. The data in these cases are still referred to as point cloud data, and the goal of topological data analysis is still to extract course features from these high-dimensional sets using algebraic topology. Overall, this method of extraction differs from methods in non-linear statistics, which is rarely topological [8, p. 63].

The next step is to convert the point cloud into a complex. One way to accomplish this goal is to use the point cloud as the vertices of a combinatorial graph. In this case, the edges are determined by proximity using some specified distance

ϵ. In order to perceive higher-order features beyond clustering, one can think of the aforementioned combinatorial graph as a scaffold for a higher-dimensional object by completing the graph to a simplicial complex – "a space built from simple pieces (simplices) identified combinatorially along faces" [8, p. 63] or a space with a triangulation [15, p. 7] (Definition 2). Two methods of filling in higher-dimensional simplices of the proximity graph are the Čech complex and the Vietoris-Rips complex (heretofore referred to by its other name, the Rips complex). This study used the Rips complex, but both methods will be described below to highlight the difference between them and to highlight the advantages/disadvantages of each method.

Definition 1. *The join of n points is a convex polyhedron of dimension n-1 called a* ***simplex****. [9, p. 9]*

Definition 2. *A* ***simplicial complex*** *can be described combinatorially as a set X_0 of vertices together with sets X_n of n-simplices, which are (n+1)-element subsets of X_0. [9, p. 107]*

Ghrist [8, p. 63] offers a definition for the Čech complex (Definition 3).

Definition 3. *Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the* ***Čech complex****, C_ϵ, is the abstract simplicial complex whose k-simplices are determined by unordered (k+1)-tuples of points $\{x_\alpha\}_0^k$ whose closed $\epsilon/2$-ball neighborhoods have a point of common intersection.*

The idea is that there is a point set X in a metric space and a number ϵ >0. Then, using a subset $S \subset X$, we form an $\epsilon/2$-ball around each point in S. S is included as a simplex if there is a common point contained in all of the balls in S. The $\epsilon/2$ balls are illustrated in Figure 3.1, and a k-simplex is formed whenever there is a subset of k points with a common intersection [10]. Thus, the Čech complex has the homotopy type of the $\epsilon/2$ cover.

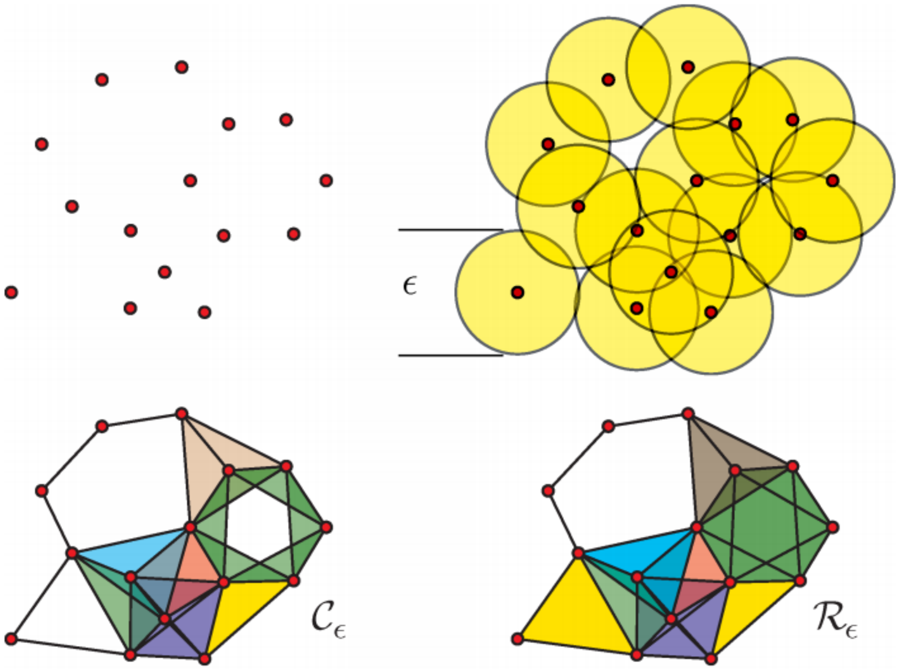


Figure 3.1: A fixed set of points [upper left] converted into a Čech complex (bottom left) and Rips complex (bottom right) (image courtesy of Ghrist [8, p. 64]).

Next, consider the Čech Theorem (or the "Nerve Theorem") [8, p. 63]. Following the definition of homotopy, this theorem implies that C, an abstract simplicial complex of potentially high dimension, behaves like a subset of $\mathbb{E}^n$ [8, p. 63].

Čech Theorem. *C_ϵ has the homotopy type of the union of closed radius $\epsilon/2$ balls about the point set $\{x_\alpha\}$.*

The difficulty with the Čech complex is that it is difficult to compute. For instance, to check to see if there exists any 5-simplices, one would need to inspect all the subsets of size 5. According to Kun [10], computing the entire complex requires exponential time in the size of the metric space. Therefore, the attention will now turn to the Rips complex (Definition 4) [8, p. 63].

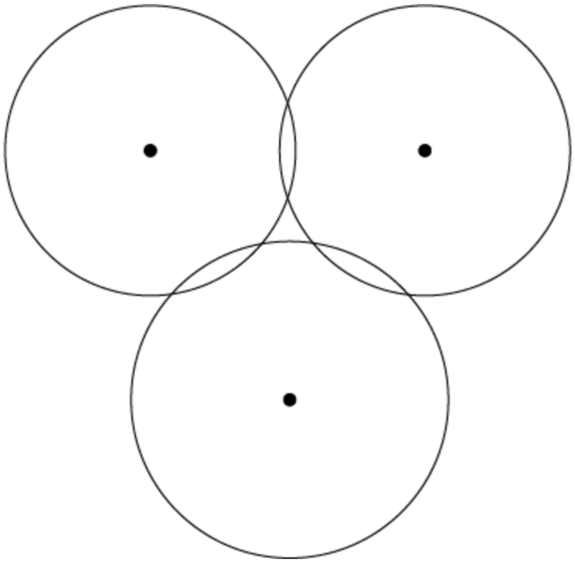


Figure 3.2: A case using the vertices of an equilateral triangle where the Rips complex is a graph with a 2-simplex but the Čech complex is simply a graph [10].

Definition 4. *Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the* ***Rips complex****, R_ϵ, is the abstract simplicial complex whose k-simplices correspond to unordered (k+1)-tuples of points $\{x_\alpha\}_0^k$ that are pairwise within distance ϵ.*

The Rips complex works in the same way as the Čech complex, but instead of adding a d-simplex when there is a common point of intersection of all the $\epsilon/2$-balls, simplices are added when all the balls have pairwise intersections [10]. See Figure 3.1 to compare the different simplices that result from the Čech complex and the Rips complex. Also, consider the following example from Kun [10]. Given three points that are vertices of an equilateral triangle where the lengths of the sides of the triangle are all one, consider drawing (1/2)-balls around each point (Figure 3.2). Then, each of the balls intersects the other two balls, but there is no point of triple intersection. In this example, the Rips complex is a graph with a 2-simplex, but the Čech complex is just a graph.

Unfortunately, the aforementioned Čech theorem does not hold for Rips complexes. The previous example is an obvious counterexample because the two complexes have different topology. Note that even though the Rips complex generally has more simplices, this complex is still considered to be less computationally expensive than the Čech complex. According to Ghrist [8, p. 63], the reason for this result is because the Rips complex is a flag complex – "it is maximal among all simplicial complexes with the given 1-skeleton," where the 1-skeleton is the topological graph. That is, an edge is included between two points in the point cloud if the corresponding $\epsilon/2$ balls overlap (Figure 3.3). The combinatorics of the 1-skeleton is stored as a graph and thus provides a rough sense of the proximity of the nodes. This process is much simpler than storing the boundary operator needed for the Čech complex. However, computing both quantities still requires exponential time [10].

One of the difficulties of converting a point cloud dataset into a global complex is knowing when a specific value of ϵ captures the topology of the dataset. As ϵ increases, the complex will transition from a discrete set (the total number of nodes in the data set) to a single high-dimensional complex. To illustrate the importance of the value of ϵ, consider a case when a sample of points is taken from a genus-2 surface (i.e., a double torus). For what value(s) (if any) of ϵ will the Rips Complex indicate that there are precisely two holes? Furthermore, how can you know if the holes indicated are not just noise? Ghrist [8, p. 65] offers a helpful diagram of a sequence of Rips complexes for a point cloud data produced from an annulus to show the different complexes produced for different values of ϵ (Figure 3.3).

In order to provide insight into this issue, topological data analysis techniques invoke the notion of persistence. The idea is to create a family of topological spaces by parameterizing with respect to the size of ϵ. A specific topological feature (e.g., a hole) is interpreted as signal over noise if the hole persists for a "significant" param-

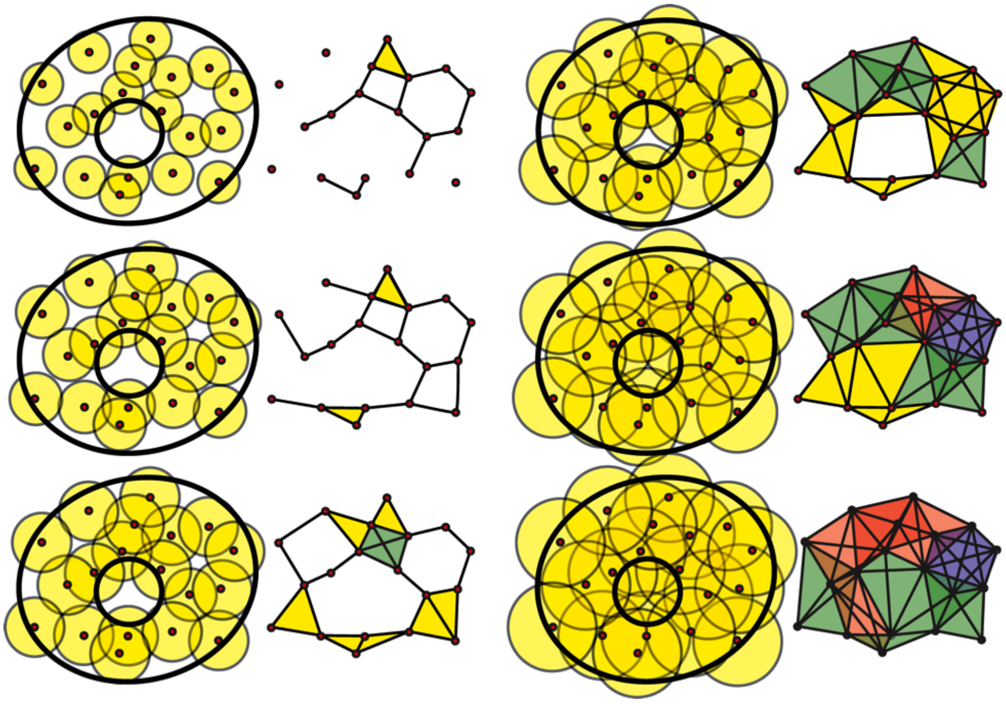


Figure 3.3: Ghrist's [8, p. 65] sequence of Rips complexes for point cloud data constructed from an annulus.

eter range. What is deemed as significant depends on the researcher's interpretation of the data set and is an active area of research, but see Lemma 1 for the minimum needed to guarantee a good approximation of the Čech complex.

Ghrist [8, p. 65] offers a more formal way of thinking about the idea of persistence.

Definition 5. *Assume that $R = (\mathcal{R}_i)_1^N$ is a sequence of Rips complexes associated to a fixed point cloud for an increasing sequence of parameter values $(\epsilon_i)_1^N$. There are natural inclusion maps*

$$\mathcal{R}_1 \overset{\iota}{\hookrightarrow} \mathcal{R}_2 \overset{\iota}{\hookrightarrow} \ldots \overset{\iota}{\hookrightarrow} \mathcal{R}_{N-1} \overset{\iota}{\hookrightarrow} \mathcal{R}_N.$$

Instead of examining homology of the individual terms $\mathcal{R}_i$, one examines the homology of the iterated inclusions $\iota : H_*\mathcal{R}_i \to H_*\mathcal{R}_j$ for all $i < j$. These maps reveal which features ***persist****.*

One crucial component to consider is whether or not using this idea of persistence actually gives a good approximation to a single Čech complex. This result would be ideal since by the Čech Theorem, the Čech complex can give an accurate representation of the homotopy of the topological space. The Rips complex is the same as that of the of the Čech complex for a particular value of epsilon. Thus, by finding a good approximation of the Čech complex using the less computationally expensive Rips complex method, it is possible to conclude that the homotopy type determined by the Rips complex matches the topological space created by the union of the ϵ-balls created from the metric space. Thus, homology will not distinguish the spaces. de Silva and Ghrist [6, p. 66]provide a proof for the lemma which provides us with the aforementioned property of the Rips complexes:

Lemma 1. *For any ϵ>0, there is a chain of inclusion maps*

$$\mathcal{R}_\epsilon \hookrightarrow \mathcal{C}_{\epsilon\sqrt{2}} \hookrightarrow \mathcal{R}_{\epsilon\sqrt{2}}.$$

Thus, topological features that persist between two Rips complexes formed from $\mathcal{R}_\epsilon \hookrightarrow \mathcal{R}_{\epsilon'}$ is a topological feature of the Čech complex $\mathcal{C}_{\epsilon'}$ when $\epsilon'/\epsilon \geq \sqrt{2}$ [8, p. 66].

3.2 Representations and Interpretations of Persistent Homologies

The next step is to represent these persistent homologies. There are two main methods of accomplishing this goal: barcodes and persistence diagrams. The results of this study contain persistence diagrams, but some of studies using topological data analysis reported here use barcodes. In order to clarify some terminology, H_i will represent a vector space, where i represents the dimension of the vector space. The dimension of H_*(C) corresponds to the persistent homology of a sequence of chain complexes, C. Edelsbrunner and Harer [7] describe homology as "the mathematical formalism for talking in a quantitative and unambiguous manner about how a space is connected" (p. 79). More specifically, homology groups "provide a mathematical language for the holes in a topological space" (p. 79). The idea of homology is to relate topologically meaningful subset of a space in order to capture holes in the space. The dimension of H_*(C) corresponds to its Betti number (i.e., B_0:= dimH_i). In other words, the p^{th} Betti number refers to the rank of the p^{th} homology group, where p defines the dimension of the space [7, p. 81]. Thus, the Betti-0 number of a specific simplicial complex refers to the number of connected components, Betti-1 is the number of one-dimensional "circular" holes, and Betti-2 is the number of two-dimensional "cavities."

3.2.1 Bargraphs

In order to represent the persistent homology, there needs to be a way to represent the data as ϵ increases. Barcodes represent the data in a way that simplices are added but never removed as ϵ increases. To do so, a barcode shows a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter ϵ and whose vertical axis represents an (arbitrary) ordering homology generators [8, p. 67]. Consider Figure 3.4, which shows a barcode for the persistent homology of the data points from the annulus from Figure 3.3.

The homologies up to H_2 are visible for a range of ϵ. In order to interpret the barcodes, consider the starting point (ϵ =0) and the seven constructed simplicial complexes from various sizes of ϵ. At the left of the barcode, when ϵ =0, the number of bars corresponds to the number of vertices in the dataset (i.e., the Betti-0). As ϵ increases, vertices are connected and so the number of connected components decreases. For instance, the second simplicial complex has a total of six connected

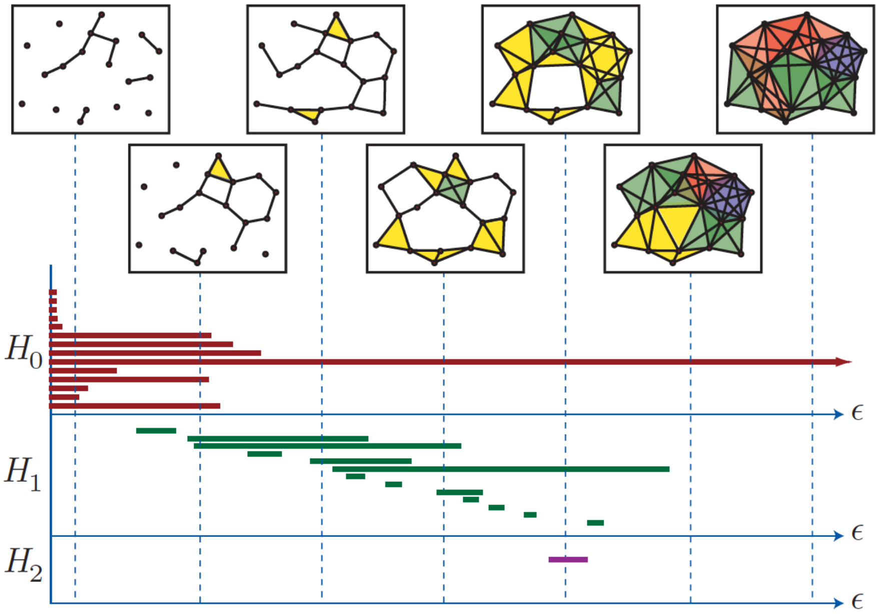


Figure 3.4: Ghrist [8, p. 67] example of barcodes using the data cloud annulus from Figure 3.3.

components, which is indicated by the six red bars intersecting the blue dotted line. When ϵ reaches the size indicated by the third blue dotted line from the left, all of the vertices have been connected and so only one red bar indicated by the 0^{th} homology, H_0, remains. Thus, when two vertices are connected at the appropriate ϵ size, the barcodes initially displays two red bars and then at the appropriate ϵ size, an arbitrary one of the bars will end and the other will continue. Because there is no way of distinguishing which bar should be the one to end, the barcode will accurately represent the situation regardless of the bar it chooses to end and which it chooses to continue. Looking at the H_1 section of the diagram, we can determine the number of holes in the data set for given sizes of ϵ. For instance, at the point where ϵ is the value indicated by the second dotted blue line, there are two holes in the dataset. These holes are the non-filled holes visible in the example simplicial complexes.

The barcode is simply an ϵ-parameterized rank, which provides a way to filter out topological noise and capture significant features qualitatively. How to determine what is considered significant has varied over the years. In the early years of topological analysis, it seemed to be the case that the *significant* findings in the data were those holes that persistent over a large range of ϵ values. For instance, Ghrist [8] remarked that in his results of the data cloud taken from an annulus (Figure 3.4) "that the point cloud likely represents a connected object with one or two significant 'holes' as measured by H_1 and no significant higher homology" (p. 68). The one or two significant holes results from the two relatively longer green bars visible in Figure 3.4, and although there did appear to be a higher dimensional hole captured as H_2, this hole only occupied a small range of ϵ, and so Ghrist deemed it insignificant. It should be noted that the goal of this analysis was to see if the persistent homologies captured in the analysis indicated that the point cloud's shape appeared to have come from an annulus. Thus, in this case, it was appropriate for Ghrist to use the length of the bar codes to determine significance in the findings.

Some research done after Ghrist's [8] piece has found that the length of the bars in a bar graph is not always an appropriate determining factor for the significance of the hole in terms of the interpretation of the dataset. That is, the persistent generators may not correspond to relevant meaningful structures in the data. Rather, it is the birth and death rates that become significant. For instance, consider the work of Lee, Chung, Kang, Kim, and Soo [11], who used the process of persistent homology through the Rips filtration "to construct brain networks consists out of fluorodeoxyglucose-position emission tomography (FDG-PET) data for 24 attention deficit hyperactivity disorder (ADHD) children, 26 autism spectrum disorder (ASD) children and 11 pediatric control subjects (PedCon)" (p. 1). Their goal was to study modeling brain networks using persistent homology. To do this, the researchers chose

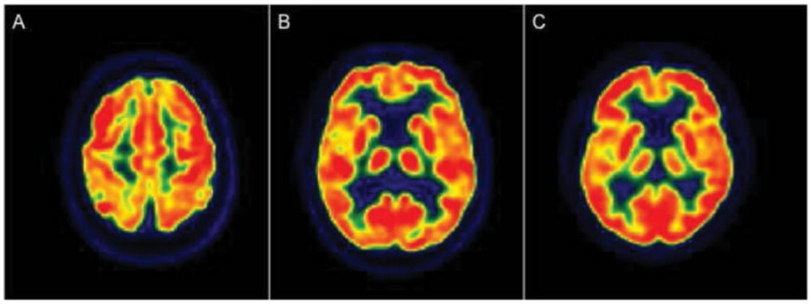


Figure 3.5: 18F-FDG PET of 3 cognitively normal subjects between 76-82 years old from Berti, Mosconi, and Pupi [2].

97 regions of interest (ROIs) from the aforementioned dataset. The researchers were only interested in looking at the number of connected components (i.e., the Betti-0 number).

Lee et al. [11] used FDG-PET data, which is a form of PET scan that uses FDG as its tracer to create three-dimensional images of the tracer concentration (see Figure 3.5 for an example image of a FDG-PET). In this case, the 3-dimensional images construct images of the brain network. They also used the Rips-Complex on their dataset. The researchers prepared the dataset by taking the mean FDG uptake within the 97 ROIs and globally normalizing the value to the individual's total gray matter mean count. The result is shown in Figure 3.6.

What Lee et al. [11, p.3] noticed is that the filtration values at which all connected components are created are identical because of the normalization of their dataset. They noted "common underconnectivity and local over-connectivity in ASD and ADHD group compared to control groups" and that "the barcode changes faster in PedCon than other groups". From these observations, Lee et al. concluded "the brain networks of ASD and ADHD groups might be more difficult to be merged into a giant connected component which connects values at which connects all ROIs when

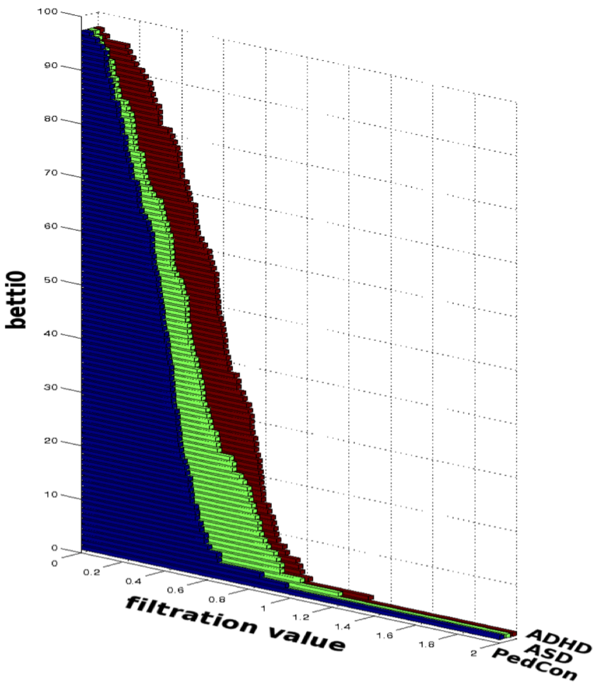


Figure 3.6: Barcode of the Betti-0 number for ADHD (red), ASD (green), and PedCon (blue) networks.

the filtration increases". Lee et al.'s [11] study is an example of the importance of considering the context in which the dataset is placed to interpret the results. In this case, the death rate of connected components between ADHD, ASK, and PedCon networks is of more significance than the number of connected components.

3.2.2 Persistence Diagrams

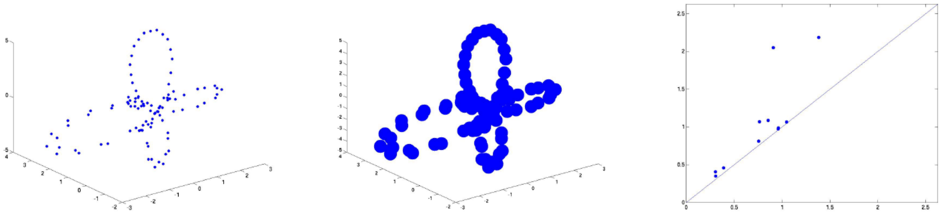
Although the previous studies illustrate the method for creating complexes (specifically Rips complexes) and the method of topological data analysis through persistent homologies, the studies use bar graphs instead of persistence diagrams to visualize the results. The following discussion describes what a persistence diagram is and how to interpret one.

In general, a persistence diagram is a multiset of dots in $\mathbb{R}^2$. The planar point (a,b) indicates the value of ϵ-ball is at when the connected component is formed (i.e., its birth), a, and the value of the ϵ-ball when the connected component dies, b. Note that this construction implies that each dot will be above the major diagonal because a component always dies after it is born. Moreover, the persistence of a dot can be thought of as b-a, which corresponds to a dot's vertical distance to the major diagonal. [1, p. 10]

For each component, C, C is not present before the ϵ value of a, C exists at every value of ϵ between a and b, and C joins with another component at the ϵ value of b. Thus, similar to the length of the bars in the graphs, the higher up the dot in the persistence diagram, the larger the ϵ-range in which that component persists. Alternatively, the smaller the persistence the dot has (i.e., the closer to the diagonal a dot is), the more likely that the feature is likely to be noise.

Consider the new point cloud below, $\mathbb{Y}$, from Bendich, Marron, Miller, Pieloch, and Skwerer [1, p. 12] and associated persistence diagram (Figure 3.7). In this example, (a) is a point cloud in $\mathbb{R}^3$ that appears to contains four loops (i.e., a compact subset of some Euclidean space $\mathbb{R}^D$, where D=3 in this case), (b) is an image of the ϵ balls created around each point at a specific value of ϵ, and (c) is the resulting one-dimensional persistence diagram (i.e., H_1 persistence diagram) over a range of ϵ values from 0 to 2.5.

The four loops can be identified in this persistence diagram by looking at the four dots that are furthest from the diagonal. According to Bendich et al. [1, p. 12], the two dots with the highest persistence correspond to the two larger loops. The dot with the later birth time (i.e., the right-most dot) corresponds to the leftmost loop in the image; the sparseness of data points in this loop requires a larger value of ϵ to fill to create the loop. The smaller loops correspond to the two dots with the



(a) Point cloud $\mathbb{Y}$ (b) Thickening of $\mathbb{Y}$ (c) Persistence diagram $\mathrm{Dgm}_1(\mathbb{Y})$

Figure 3.7: Point cloud $\mathbb{Y}$ to persistence diagram from Bendich et al. [1].

next highest persistence. The smaller persistence of these loops occurs because the value of ϵ needed to close the loops is much less than the value needed to close the aforementioned loops. The remaining dots near the diagonal indicate small loops that appear and disappear within a small range of ϵ values as a result of holes being created between overlapping sets. Thus, the persistence diagram accurately depicts these holes as noise. Figure 3.7 only shows the persistence diagram for the Betti-1 case, but similar diagrams can be created for other dimensions (see persistence diagrams in Chapter 5).

In summary, this overview contains a description of how a point cloud is converted into complexes. Complexes are then formed for a range of sizes of ϵ-balls around each point. Bar graphs and persistence diagrams provide a way to visualize the homologies that persist for different ranges of the parameter ϵ. Lastly, when interpreting either the bar graphs and persistence diagrams, researchers should always consider the context of the dataset used to create the representations.

Chapter 4

Methods

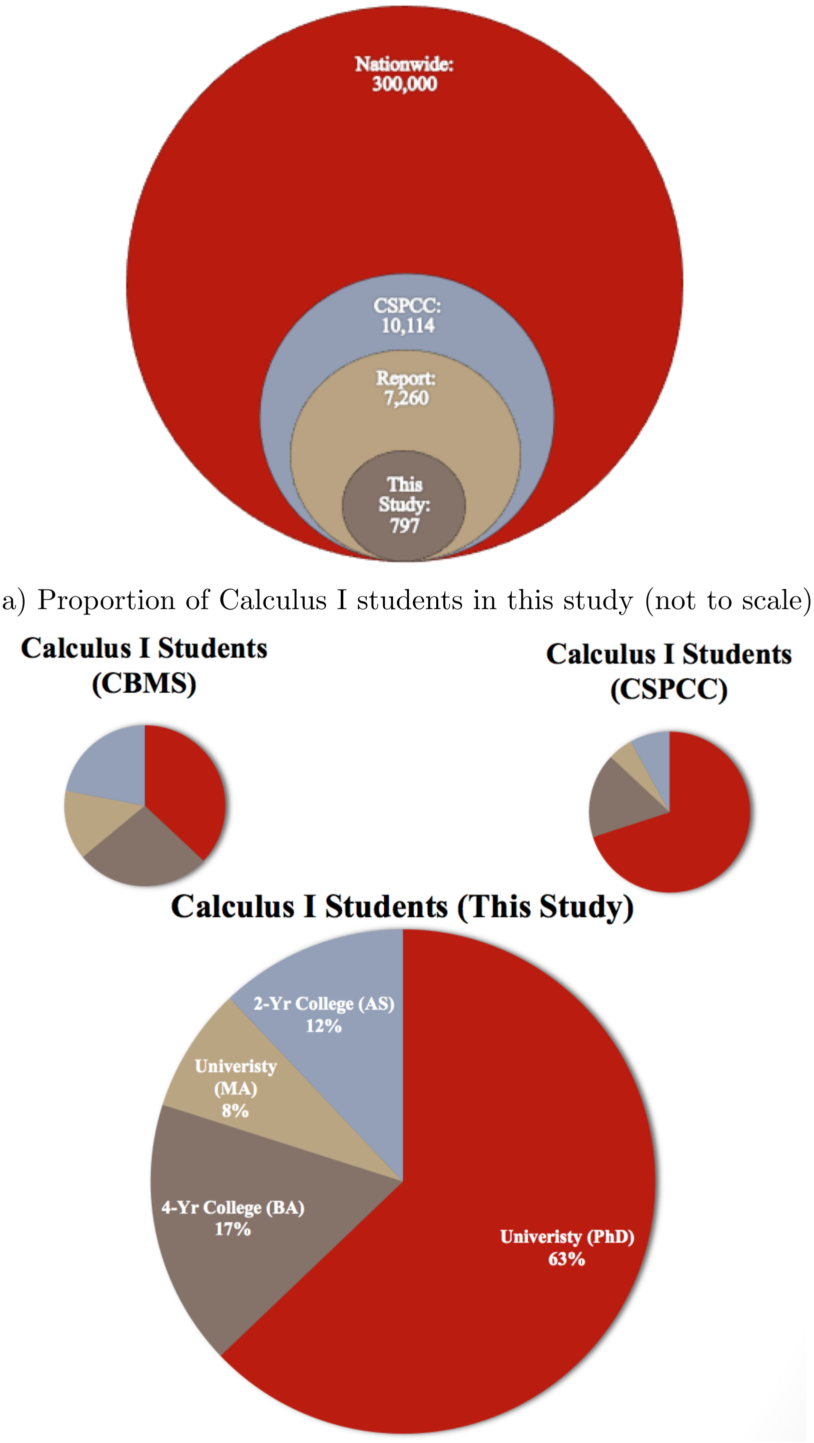
The goal of this study is to apply topological data analysis techniques to the dataset described in Chapter 2. The dataset, maalongdatafile_ANON.csv, was obtained with the permission of the MAA. The data was imported into the statistical software, R. After preparing the data for analysis, the dataset was analyzed using R's persistent homology code. The resulting persistent diagrams, and in some cases, three-dimensional plots of the dataset are in Chapter 5.

4.1 Preparing the Dataset

The datafile, maalongdatafile_ANON.csv, contains the survey responses from all of the participants in the study. In order to limit the course of the study, only the Calculus students who responded to both the pre- and post-course surveys and had their grades recorded were considered. This restriction reduced the number of participants down to 797 Calculus students (Figure 4.1a). The distribution of Calculus I students by institution type is displayed in Figure 4.1b. Since much of the survey data was recorded using a Likert scale and topological data analysis requires numerical values to run, the Likert scale responses were converted into numerical values that also included 1 to indicate students who did not respond to that particular survey question (e.g., 1-No response, 2-Strongly disagree, 3-Disagree, 4-Slightly Disagree, 5-Slightly Agree, 6-Agree, 7-Strongly agree). Similarly, student grades A-F were converted to the numbers 1-13 (1-F, 2-D-, 3-D, 4-D+, 5-C-, 6-C, 7-C+, 8-B-, 9-B, 10-B+, 11-A-, 12-A, 13-A+). This conversion method was applied to all of the variables of interest for this study (see Chapter 5).

R has a function that computes the persistent homology of a given dataset [12]. As described in Chapter 3, the steps for computing persistent homology include creating a filtered simplicial complex and then computing the persistent homology of the filtered simplicial complex. R has the ability to use the Rips complex to construct the simplicial complexes, and this method was chosen due to its computational advantages (see Chapter 3). In this study, the maximum dimension considered was 2 (i.e, H_0, H_1, and H_2). This range was considered in order to explore possible interpretations at various levels. The maximum filtration was set to 4. I chose this number by experimenting with various values; the number appeared to be small enough not to obtain a memory error and large enough to reveal persistent information. In some cases, the maximum filtration needed to be larger to display all of the persistent information, and I made the appropriate adjustments. The output of the function is a matrix with three columns. Each row corresponds to a persistence interval. The first column contains the dimension of the interval, the second column contains the starting point, and the third column stores the ending point (i.e., the values of ϵ that corresponded to the birth and death).

In order to produce a persistence diagram from the output of the function that computes the persistent homology of the dataset, I used another function in R which plots the starting and ending points as the coordinates $(x_i,\ y_i)$, respectively, in the plane. Markers identify the dimensions: red circles indicate H_0, green triangles indicate H_1, and blue plus signs indicate H_2. One issue to note is that there is a possibility that two dots have the same starting and ending points, and the multiplicity of an interval is not visible on these persistence diagrams beyond a bolder marker. In this case, using barcodes to visualize the persistent homologies may be more ideal.



(a) Proportion of Calculus I students in this study (not to scale).

(b) Calculus I students in this study, nationwide (CBMS) (top left), and the CSPCC (top right) by institution type.

Figure 4.1: Distribution of Calculus I students in this study.

4.2 Topological Data Analysis

I used three different methods of data input in this study, and I interpreted the resulting persistence diagrams. First, I included the numbers of responses that corresponded to matching responses for two of the survey questions and normalized the result. I plotted the persistence diagrams for these cases and compared them to three-dimensional plots of the data for further analysis and contextual interpretation. Second, I analyzed the persistence diagrams of the survey questions asked both at the beginning and end of the students' semesters in Calculus I. Third, I considered students' responses from multiple survey questions. Each of these methods are discussed in more detail in the following sections.

4.2.1 Normalizing the Data

The main method of analysis involved considering the responses to two survey questions and the frequency of each of the responses. The reason for considering the frequency of the responses was due to the large sample size combined with the low number of variables considered and the low number of potential responses. The result of this combination of factors meant that most of the possible data points were filled when only considering students' responses to two variables (i.e, survey questions); that is, all possible combinations of responses were present in the dataset, even when a lower sample was taken. See Figure 4.2 for an example of this result created from two variables and 100 randomly sampled responses. Thus, resulting persistence diagrams of any two variables looked very similar and there was no way to get a sense of the structure of the data.

In order to resolve this issue, I created a third variable that corresponded to the frequency of each type of response. For instance, say 30 people responded 3 to

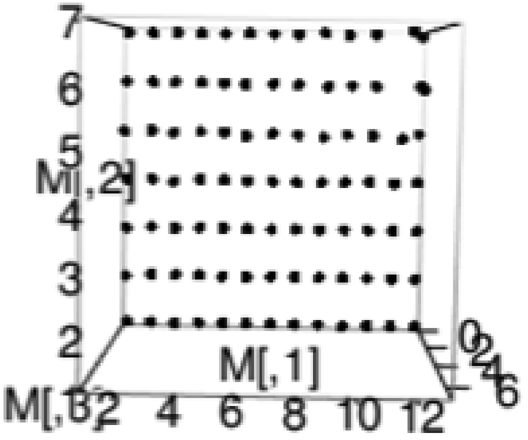


Figure 4.2: Point cloud covers the space.

the first survey question considered and 5 to the second survey question considered. Then the frequency of the coordinate pair (3, 5) is 30. However, because of the high value of the frequency of pairs relative to the values of the responses (mostly 1-7), the data was normalized by setting the value of the highest frequency equal to the maximum value of the survey responses considered. This strategy avoids dealing with the scaling issues that could be involved in determining the persistent homologies for a range of ϵ values by making what is considered to be a "significant" range of ϵ closer in value across each of the variables.

After normalizing the data, the resulting matrix was run through the function for topological data analysis and I analyzed the resulting persistence diagrams. To draw conclusions from the diagrams, I also considered the 3-D plots of the data in order to get a sense of the location of any identified loops and to determine what the loops could indicate about the data.

4.2.2 Pre-/Post- Persistence Diagrams

As described earlier, some of the survey questions were asked both before and after the Calculus I course. Although analyzing pre- and post-course survey data needs to consider the context of the course (i.e., most students who took the post-survey successfully completed the course), there is still information to be gleaned from the pre- and post-course survey responses from these students. That being said, I used two methods of inputting the data for these types of questions.

In the first method, I considered the persistent homologies of several of the high dimensional data cloud of pre-course survey questions identified to be about students' attitudes towards mathematics. Then I constructed the persistence diagram for the same survey questions using the post-course data. Lastly, I compared the two persistence diagrams to one another to see if there were any significant differences between them.

It should be noted that a sample of 100 random students were taken from the total sample in this pre- and post-course analysis and the high dimensional input form of analysis described in Section 4.2.3. This decision was made because of the likelihood that most of the possible data points would be represented in the dataset with a large enough sample size as discussed in Section 4.2.1 and because persistence diagrams do not account for multiplicity or frequency of results in these cases. Thus, only a proportion of the results were considered in an effort to identify some underlying structure in the dataset. Each random sample of 100 students was used to create pre- and post-course variable analysis. Several random samples of 100 students were used to create multiple persistence diagrams in an effort to choose a persistence diagram that seemed to be representative of total data cloud. Chapter 5 depicts the persistence diagrams that I deemed to be a typical resulting persistence diagram for the different random samples of 100 students.

For the second method, I plotted the pre-course survey question responses with the post-course survey question responses for the same survey questions along with the frequency of the responses. The survey questions chosen were identified as the two variables of attitude that had the largest effect size on attitude by the CSPCC. This configuration enabled me to consider loops in the data that could potentially correspond with shifts in the student responses because responses along the diagonal of the survey responses in the 3-D plot would indicate that the students did not vary their responses from the pre- and post-course survey question.

4.2.3 High Dimensional Input

One last configuration of the data involved producing the persistence diagrams from several responses to survey questions. This form of data input was mostly exploratory in measure because the interpretation of the results becomes difficult at high dimensions. However, the variables considered for this method were based off of the results from the MAA Calculus Report. For instance, the section on factor analysis identified three clusters of variables that all seemed to indicate the same information; that is, the covariance between the variables was high. I created persistence diagrams of some of the variables placed in each of these categories in order to observe the homologies produced by these variables.

Chapter 5

Results

This section begins by outlining some of the observations made from analyzing the persistence diagrams that were particular to the discrete nature of the dataset. Then, it includes the results of the study using each of the three methods of data input described in Chapter 4.

5.1 Discrete Data Points

In this section, I describe how the discrete nature of the dataset affects the resulting persistence diagrams. Recall that most of the survey questions considered were based on Likert scale responses. Thus, most of the responses were based on answers from the integer values 1-7. Similarly, yes/no responses were given the discrete values 0 and 1, and the grading scale was translated into a discrete scale as well. The discrete and integer-valued nature of the data set implies some commonalities across the persistence diagrams.

For example, consider the case when two survey questions are the variables used to create the point cloud. The resulting point cloud is a unit grid of the possible responses. Figure 5.1 shows a simple four-point version of this case. Since the distance between each of these points is 1, two ϵ-balls initially connect when ϵ=1 (Figure 5.1 (left)). Moreover, if using the Rips complex, the components connect and a loop will form. The loop dies once $\epsilon=\sqrt{2}$ because that is the value of ϵ which results in the ϵ-balls meeting in the center and, thus, closes the loop (Figure

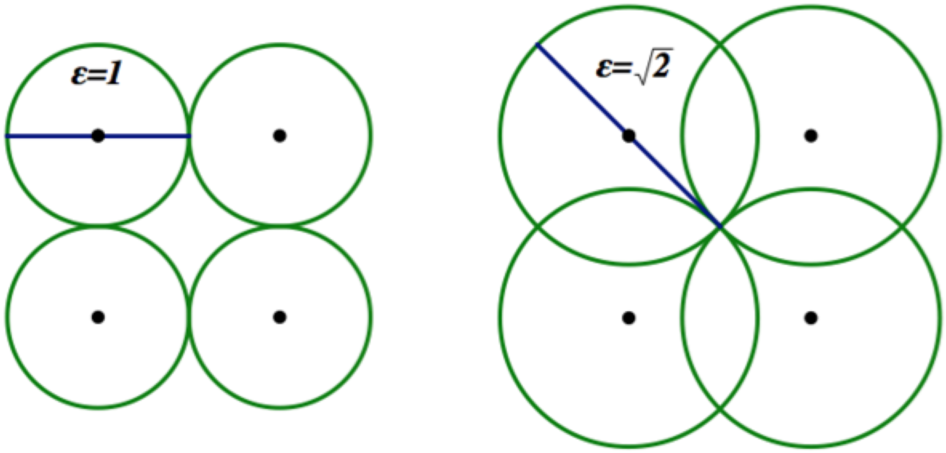


Figure 5.1: Point cloud with ϵ=1 (left) and ϵ=$\sqrt{2}$ (right) to show the birth and death rate of a loop in a unit square point cloud.

5.1(right)). Hence, the point (1, $\sqrt{2}$) is plotted on the persistence diagrams to indicate the birth and death of this loop in the data (Figure 5.2).

In the case when the frequency of each of the combinations of the responses of the two variables is also considered, this same result may not occur because the normalized frequency value is not necessarily zero or the same value as surrounding combinations of responses. Building on the example in Figure 5.1, each of the points in the point cloud would be raised depending on the frequency of times that response was found in the dataset. Thus, the minimum starting ϵ value of a loop in the data is when ϵ=1 and the minimum ending ϵ value of a loop is when ϵ=$\sqrt{2}$.

In terms of the number of connected components, no matter the situation presented in this study, each point in the point cloud is a connected component that starts at ϵ=0 and ends at various values of ϵ. Thus, the coordinate pairs, (x_i, y_i), for each point in the point cloud, p_i, on the persistence diagram for these connected components will fall on $(0, y_i)$, where $1 \leq y_i \leq m$, where m is the minimum value of ϵ at which only one connected component exists. Furthermore, it should be noted

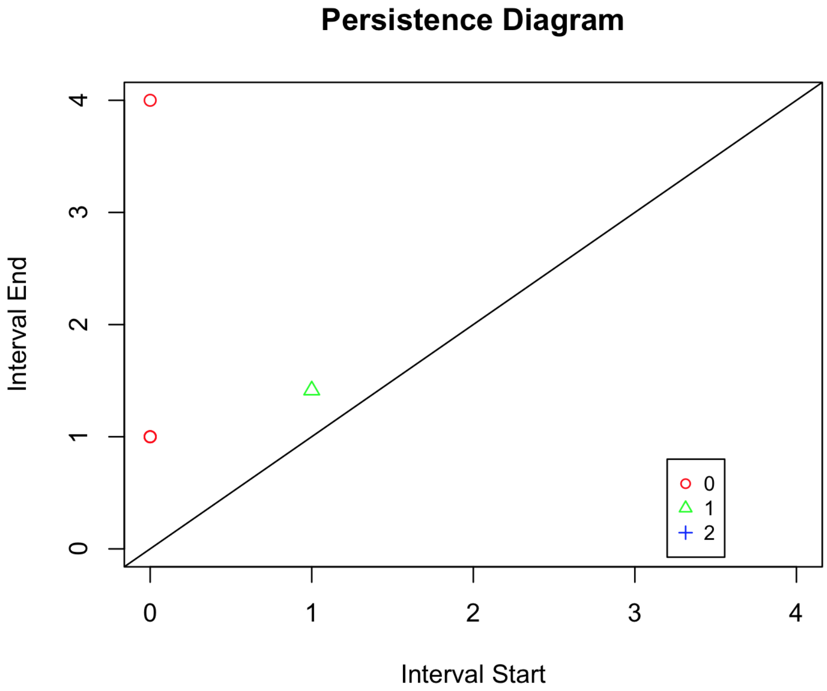


Figure 5.2: Persistence diagram of a unit square point cloud.

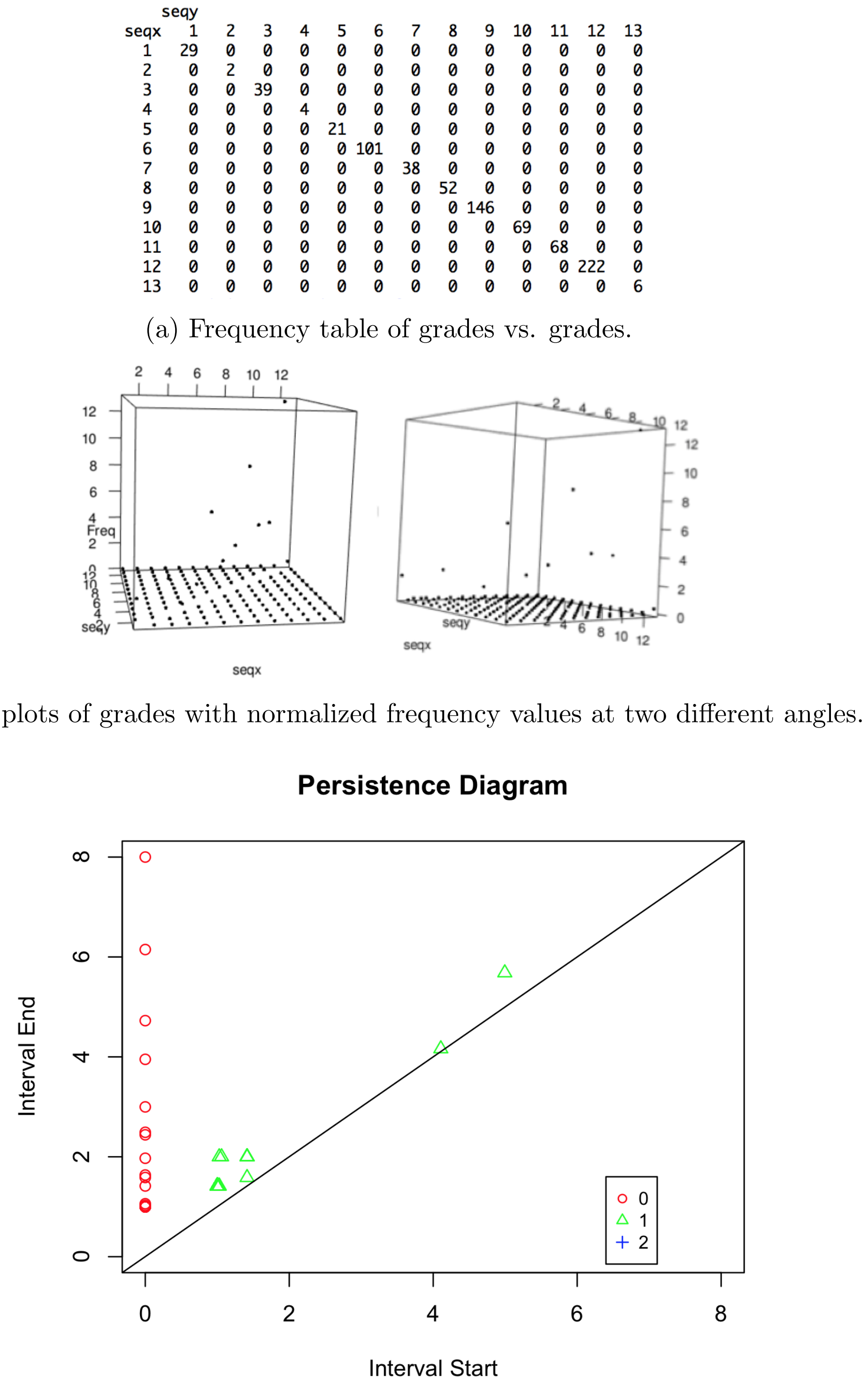
that the persistence diagrams created in this study always plot a connected component in the upper left corner of the diagram regardless of the maximum filtration size because there is always one remaining connected component that never has an ϵ value for which the connected component ends.

5.2 Normalized Data

This section contains another base case for comparison before there is a discussion of specific examples from the dataset. For instance, consider the case in which students respond with exactly the same numerical value response to both variables. In this case, the 3D-plot of the data would simply be along the diagonal of the two axes representing the two variables. In this method of analysis, the frequency of the responses is considered, and so the third variable would correspond to the frequency of each of the responses. As a simple example, Figure 5.3a shows an example of the frequency table produced when students' final grades were compared against students' final grades. The students' actual grades in their respective Calculus I courses were provided by the departments based on student ID matches. The counting table shows how many students scored each grade (where the values for *seqx* and *seqy* are 1=F through 13=A+) along the diagonal (e.g., of the 797 total students, 222 students received an A). These frequency values were normalized before analyzing the persistent homologies, so that the maximum value (i.e., 222) became the maximum value of the other variables (i.e., 13), and every other number was scaled using that same scaling factor (see Figure 5.3b to see the result of this scaling). The resulting persistence diagram is in Figure 5.3c.

The diagram shows that most of the loops occur around the $(1, \sqrt{2})$ point mentioned earlier. The distribution of the grades in this example lends itself to producing loops with larger values of ϵ as well. That is, more students were likely to receive letter grades that did not have a + or - attached to it (e.g., a B is more common than a B+ or B-), which ends up producing a point cloud that is susceptible to loop detection. Thus, two loops were detected between the $4 < \epsilon < 6$.

With these base cases established, consider the following result in which one loop has a significantly larger lifespan than any of the other identified loops in the data (Figure 5.4a). The variables considered in this case are the students' final grades and their responses to the statement, "My Calculus Instructor discouraged me from wanting to continue taking Calculus." The third variable is the normalized frequency of the combinations of responses. Notice the one H_1 loop with a much higher persistence than the rest of the identified loops. Its birth is just above $\epsilon = 1$ and it persists until its death at just over $\epsilon = 3$, giving it a persistence value of



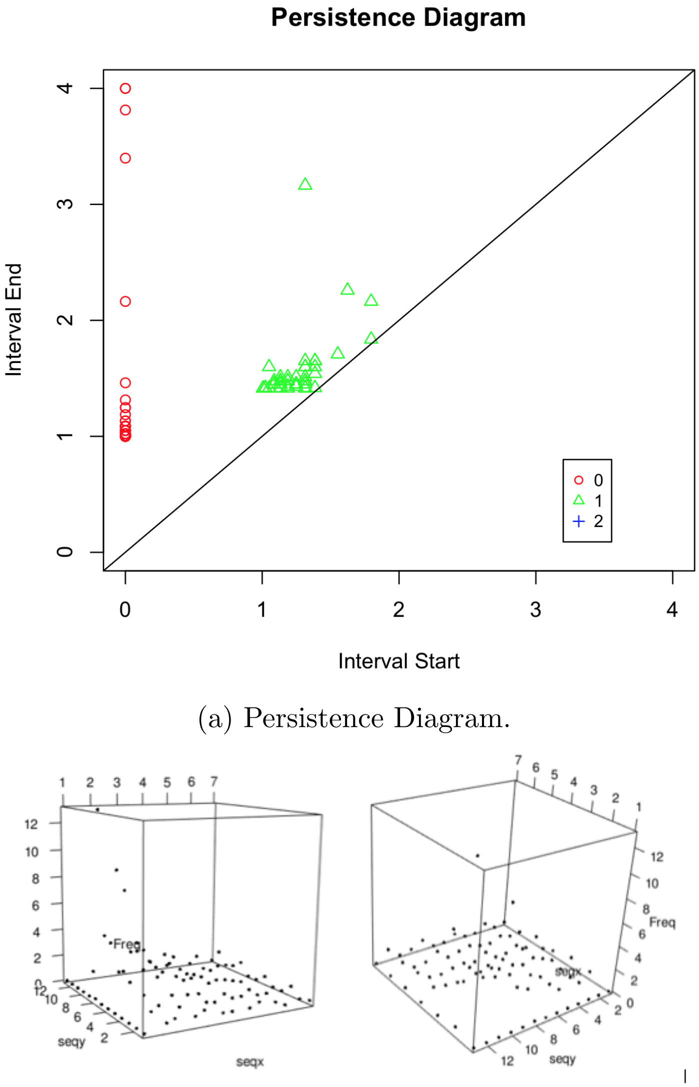
```
      seqy
seqx   1   2   3   4   5   6   7   8   9  10  11  12  13
  1   29   0   0   0   0   0   0   0   0   0   0   0   0
  2    0   2   0   0   0   0   0   0   0   0   0   0   0
  3    0   0  39   0   0   0   0   0   0   0   0   0   0
  4    0   0   0   4   0   0   0   0   0   0   0   0   0
  5    0   0   0   0  21   0   0   0   0   0   0   0   0
  6    0   0   0   0   0 101   0   0   0   0   0   0   0
  7    0   0   0   0   0   0  38   0   0   0   0   0   0
  8    0   0   0   0   0   0   0  52   0   0   0   0   0
  9    0   0   0   0   0   0   0   0 146   0   0   0   0
  10   0   0   0   0   0   0   0   0   0  69   0   0   0
  11   0   0   0   0   0   0   0   0   0   0  68   0   0
  12   0   0   0   0   0   0   0   0   0   0   0 222   0
  13   0   0   0   0   0   0   0   0   0   0   0   0   6
```

(a) Frequency table of grades vs. grades.

(b) 3D-plots of grades with normalized frequency values at two different angles.

(c) Persistence diagram of grades with frequency.

Figure 5.3: Topological data analysis of final grades vs. final grades vs. normalized frequency.



(a) Persistence Diagram.

(b) 3-D plots at different angles.

Figure 5.4: Topological data analysis of grades, discouragement from continuing Calculus, and frequency.

approximately 2. The low starting value of the loop indicates that the points in the point cloud that resulted in the loop are closely spaced together. The large persistence value indicates a much larger loop than the one given in the unit square example earlier. In order for a loop to persist for a large range of ϵ, the variable in the third dimension must be spaced much further away from the points included in the loop. There does appear to be a large loop around the *seqx*-*seqy* plane in the 3-D plot of this situation (Figure 5.4b), which corresponds to the survey responses and their grades. Prior analysis of the grades shows that the students' grades were unequally distributed; in fact, much of the grades are distributed between the values 6-12 (corresponding to C-A grade-wise). Thus, a loop with high persistence might indicate that a majority of students leaned towards a narrow range of responding to the question. This leaning would cause higher frequency values and result in moving the points away from the *seqx*-*seqy* plane while the remaining combinations of responses would have lower frequencies and thus a much shorter persistence. In the diagram, this conclusion is noted by the large cluster of values around $\epsilon = 1$. The 3-D plot structure seems to indicate that many of the responses of the students were 2-3 (i.e., "Strongly disagree" and "disagree" because a value of 1 was reserved for missing responses). Thus, we can conclude that most students did not feel as though their Calculus I instructors discouraged them from wanting to continue to take Calculus.

There are limitations to this analysis. First, it is difficult to note how students who scored lower grades (which occurred much less frequently) responded to the questions because the frequencies of these combinations of responses are already likely to be smaller and closer together in value than the others. Thus, future research could consider normalizing the variables in a way that accounts for an unequal distribution of the results of a variable. Second, when a student did not respond to a question,

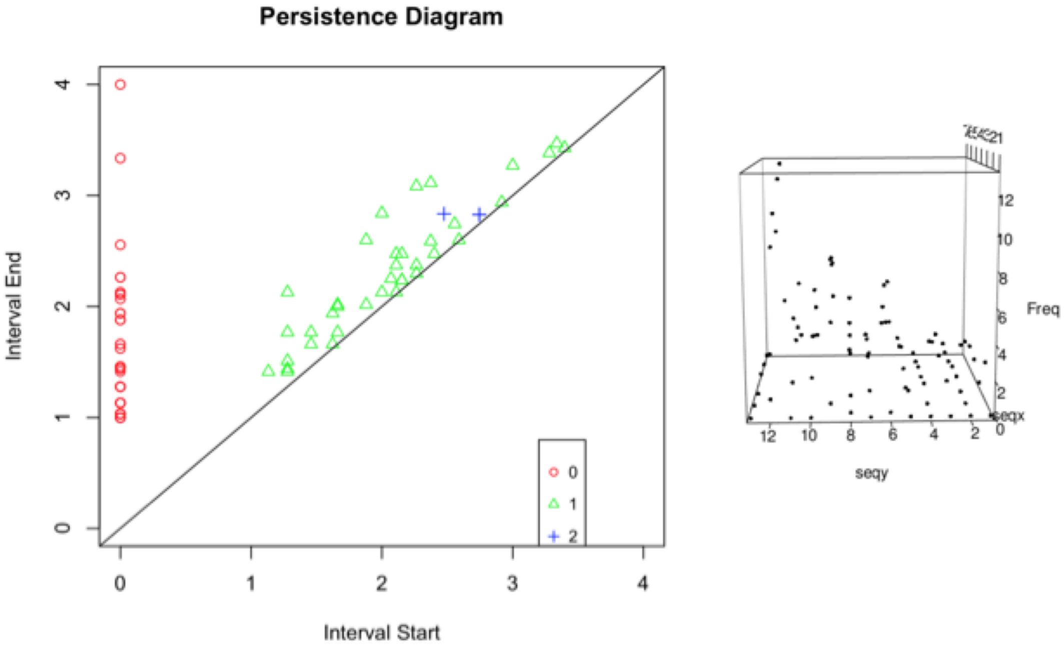


Figure 5.5: Persistence diagram and 3-D plot of grades, prior experience with group work, and frequency.

they were given a score of a 1. Only a handful of students did not respond to any particular question in this study, and so this portion of the 3-D graph usually results in a line close to the *seqx*-*seqy* plane. Consequently, this information appears to form part of the loop that appears in the 3-D graph. Thus, the loop would remain undetected if missing responses were not included in the data cloud.

Nevertheless, we can compare this result to Figure 5.5, which shows the results of the persistence diagram and 3-D plot of three variables. The first variable is the students' grades and the second variable is the students' responses to the following statement: "The teacher of my last mathematics course in high school frequently had us work in groups." The third variable is the normalized frequency of the pairs of responses. The persistence diagrams have several loops away from the $(1, \sqrt{2})$ point, most of which do not persist for a large range of ϵ. This result seems to indicate that there several combinations of responses that have varying levels of frequencies where some points forms loops close together and others are farther apart, unlike the previous example. Considering the context, particularly keeping in mind the unequal distribution of the variables, this result seems to indicate that students at all different grade levels have all had varying degrees of experience with group work in their previous mathematics course in high school. Thus, it seems likely that past experiences with group work is not a good predictor of student success in Calculus I (in terms of final grades).

5.3 Pre- and Post- Data

The purpose of testing pre- and post-course data analysis in this study is to get a sense of (1) whether or not structural changes exist in the persistence diagrams, and (2) whether or not these results relate to the results of the CSPCC.

Figure 5.6 shows the persistence diagrams of the students' responses to questions that were asked in both the pre- and post-course surveys. In this example, only three variables were considered. These variables include the students' responses to the following statements: "I am confident in my mathematics abilities," "I enjoy doing mathematics," and "Do you intend to take Calculus II?". The last variable has the possible responses "No," "I don't know yet," and "Yes," and were given the values 1, 2, and 3, respectively. These three variables were used as an "Attitude" composite score based on their high correlations with one another ($r = 0.52$ to $r = .70$) [4, p. 17]. The variable "confidence in math" had the largest effect size (Table 5.1) [4, p. 18], but even that size is not typically considered to be a large effect size. However, the persistence diagrams between the pre-course and post-course attitude variables are different, so some structural change was observed. In particular, the ϵ values for

Table 5.1: CSPCC report of pre- and post-course survey responses to attitude statements.

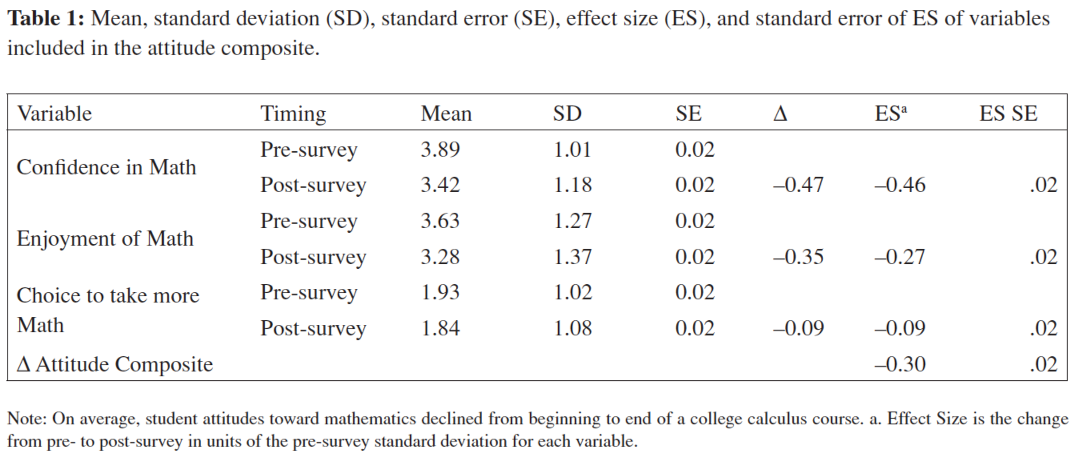


Table 1: Mean, standard deviation (SD), standard error (SE), effect size (ES), and standard error of ES of variables included in the attitude composite.

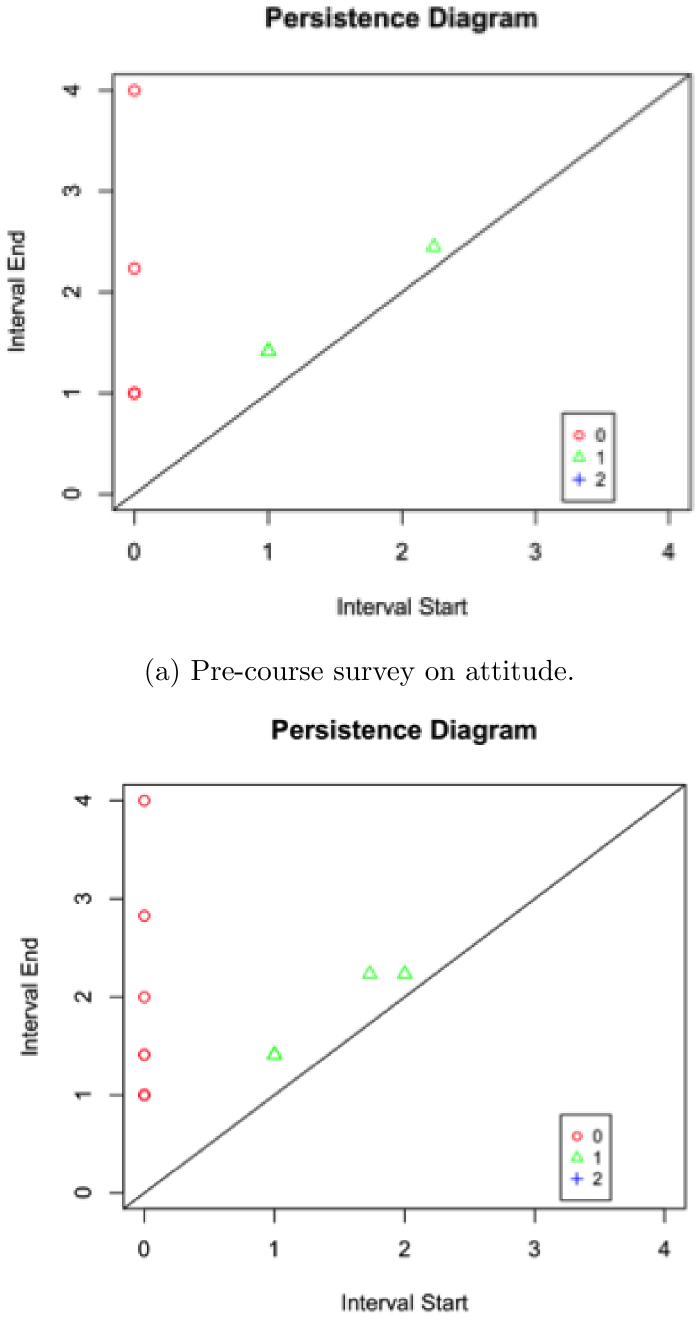
Variable	Timing	Mean	SD	SE	Δ	ES[a]	ES SE
Confidence in Math	Pre-survey	3.89	1.01	0.02			
	Post-survey	3.42	1.18	0.02	–0.47	–0.46	.02
Enjoyment of Math	Pre-survey	3.63	1.27	0.02			
	Post-survey	3.28	1.37	0.02	–0.35	–0.27	.02
Choice to take more Math	Pre-survey	1.93	1.02	0.02			
	Post-survey	1.84	1.08	0.02	–0.09	–0.09	.02
Δ Attitude Composite						–0.30	.02

Note: On average, student attitudes toward mathematics declined from beginning to end of a college calculus course. a. Effect Size is the change from pre- to post-survey in units of the pre-survey standard deviation for each variable.

the end of the connected components are more scattered and there are two plotted locations where loops occurred that differ from the pre-course persistence diagram.

It is important to recall that one of the limitations of this study is that only a random subset of all of the students were analyzed, and so a random sample of 100 students is being used to represent the data cloud. Pre-course and post-course persistence diagrams were compared with five different random samples of 100 students, and the ones in Figure 5.6 are typical graphs. The number of loops varied slightly for different samples of 100 students, but it was always the case that the number of post-course loops was greater than the number of pre-course loops (i.e., the Betti-1 of pre-course survey responses was greater than the Betti-1 of the post-course survey responses). Future studies can be done to determine an optimal sample size to consider and ways to interpret the differences in the persistence diagrams (e.g., looking at the 3-D plot of the variables).

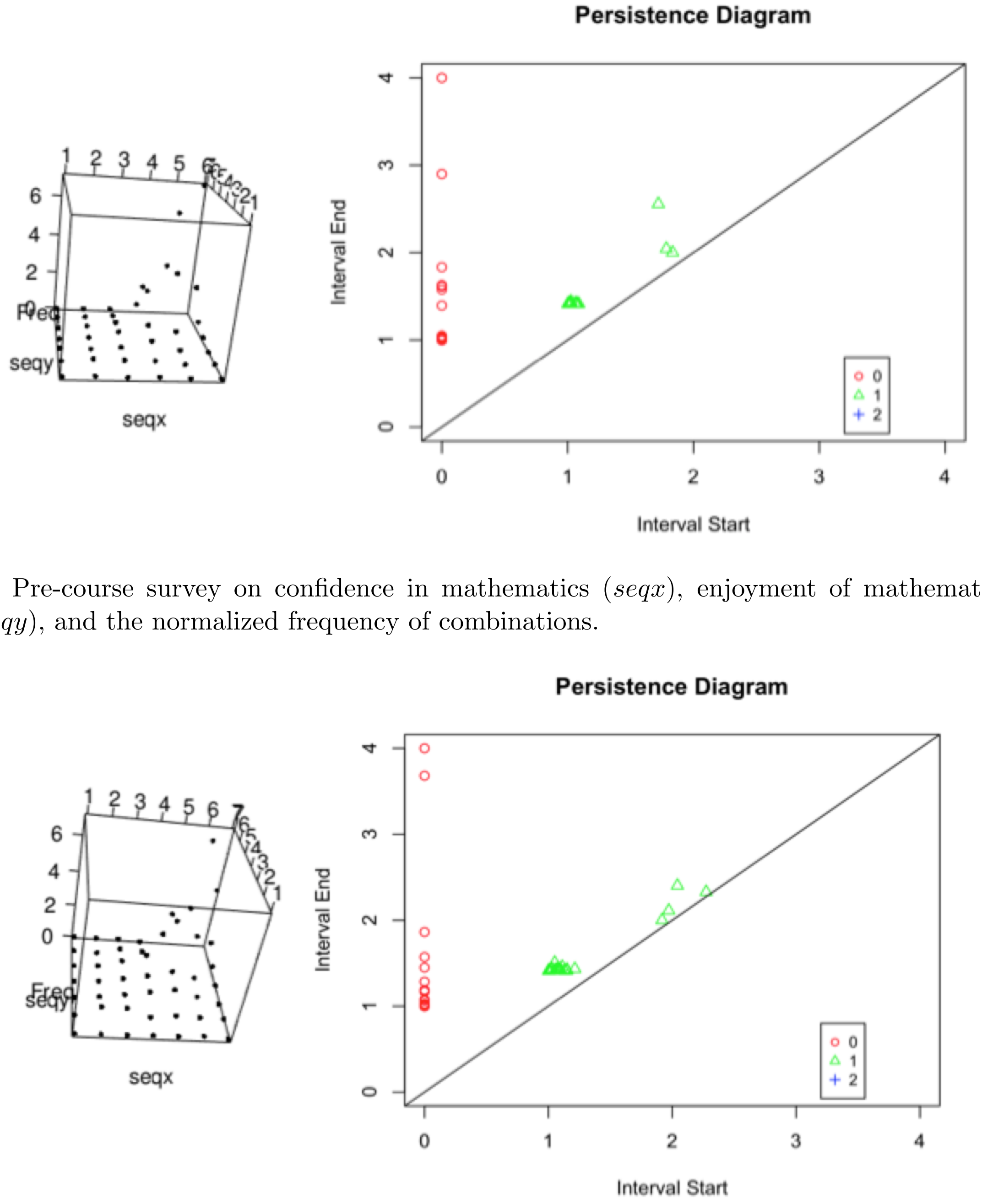
In order to compare the pre-course and post-course survey data to one another, the top two variables that contributed to the attitude composite score, confidence



(a) Pre-course survey on attitude.

(b) Post-course survey on attitude.

Figure 5.6: Pre-course vs. post-course persistence diagrams of attitude variables.



(a) Pre-course survey on confidence in mathematics (*seqx*), enjoyment of mathematics (*seqy*), and the normalized frequency of combinations.

(b) Post-course survey on confidence in mathematics (*seqx*), enjoyment of mathematics (*seqy*), and the normalized frequency of their combinations.

Figure 5.7: Pre-course vs. post-course persistence diagrams of confidence in mathematics, enjoyment of mathematics, and the normalized frequency of their combinations.

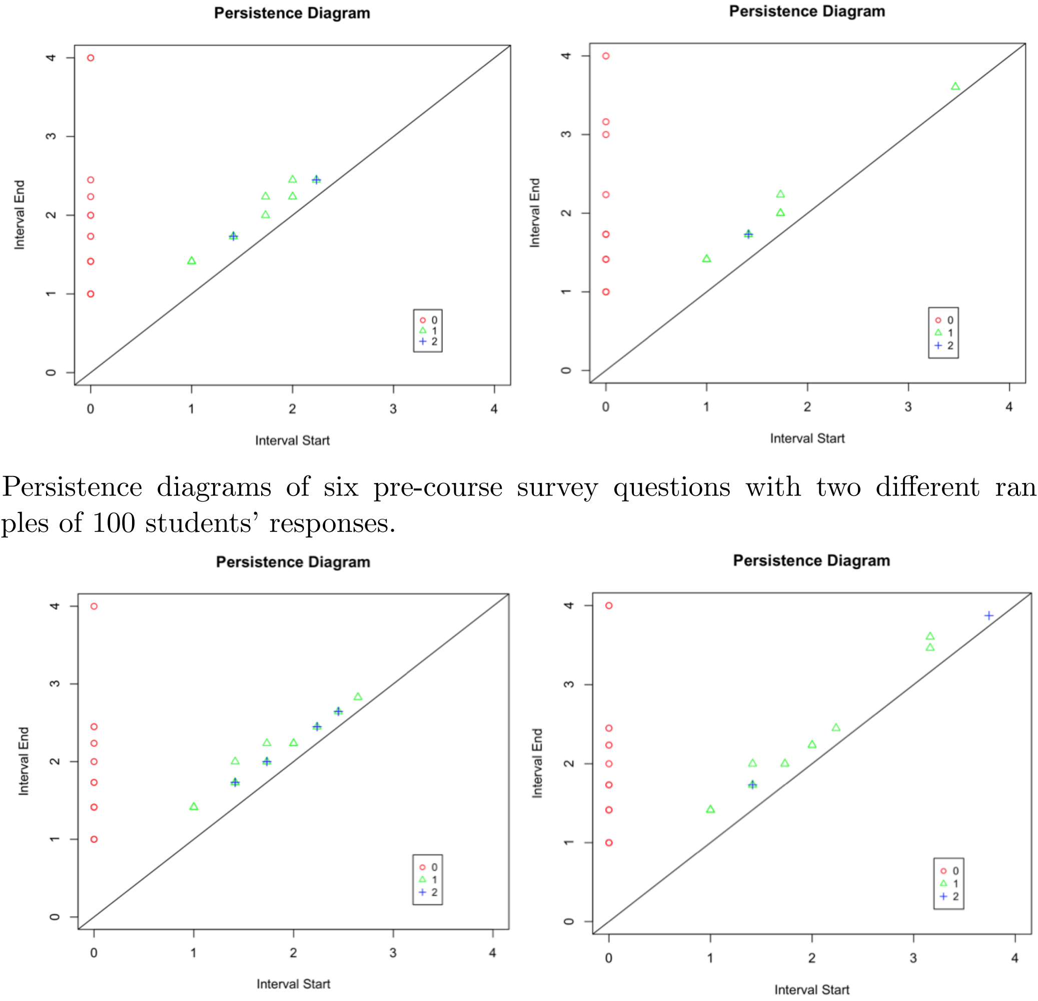
and enjoyment, were compared to one another along with the frequency of the combinations of the responses. Again, structural differences exist between the two resulting persistence diagrams (Figure 5.7).

5.4 High Dimensional Input

To see how topological data analysis fared with high dimensional input, persistence diagrams of six different variables were created. Two persistence diagrams were created using six variables that were asked pre-course and post-course. The input variables were responses to the following statements:“What grade do you expect in this calculus course?” (scaled 0=F to to 4=A), “Do you intend to take Calculus II?” (scaled 1=No, 2=Don’t know/unsure/N/A, 3=Yes), “How certain are you in what you intend to do after college?” (scaled 0=Not certain at all to 3=Very certain), “When experiencing a difficulty in my math class” (responses scaled 0= I try hard to figure it out on my own to 3= I quickly seek help or give up trying), “I am confident in my mathematics abilities” (scaled 1=missing, 1=Strongly disagree to 7=Strongly agree), and “I enjoy doing mathematics” (scaled same as previous).

Figure 5.8 shows the persistence diagrams of all the aforementioned pre-course survey questions and all of the post-course survey questions using two different random samples of 100 students’ responses. The same subsets of students were considered in the pre- and post-course persistence diagrams on the left, and another subset of students were considered in the pre-and post- course persistence diagrams on the right. Notice the short-lived prevalence of more H_2 points visible in the post-course persistence diagram than in the pre-course persistence diagram with the same random sample. Furthermore, more H_2 points were plotted in the post-course persistence diagrams than their respective pre-course diagrams. However, it should be noted that when different random sample subsets of 100 students were compared, there was a lot of variability in the persistence diagrams. Thus, it is difficult to make any concluding remarks using a sample size of 100 points.

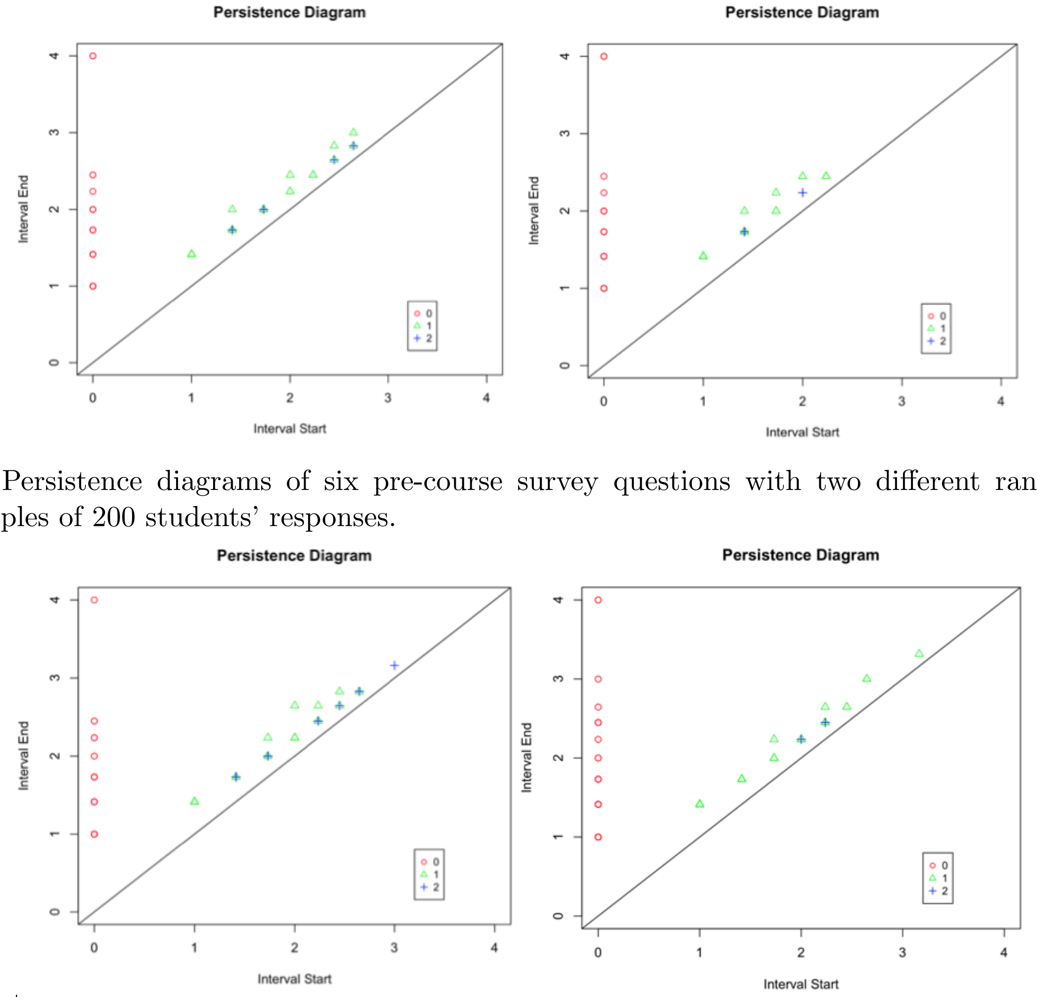
Therefore, a random sample size of 200 was used to compare persistence diagrams (Figure 5.9). The persistence diagrams are organized in the same way as Figure 5.8. Notice that although the post-course survey persistence diagrams with 200 responses seem to match closely, the pre-course survey persistence diagrams still vary. However, there does appear to be a consistent tower of two approximately equally spaced H_1 loops, possibly indicating some underlying structure in the data in both the pre- and post- course survey responses. However, comparisons between pre- and post-course survey responses using these subsets is still difficult because of the variations not only across different sample sizes but also between two different subsets of the same size. Thus, it appears that creating persistence diagrams without choosing subsets of data may not saturate all possible combinations of responses and a smaller sample is then not necessary. However, recall that creating the complexes still requires exponential time, and thus considering all the data points may not be a practical solution. Future researchers could look into calculating an optimal subset sample size to test based on the number of possible responses and the number of actual responses.



(a) Persistence diagrams of six pre-course survey questions with two different random samples of 100 students' responses.

(b) Persistence diagrams of six post-course survey questions with with two different random samples of 100 students' responses.

Figure 5.8: Persistence diagrams of pre-course and post-course survey questions with with two different random samples of 100 students' responses.



(a) Persistence diagrams of six pre-course survey questions with two different random samples of 200 students' responses.

(b) Persistence diagrams of six post-course survey questions with with two different random samples of 200 students' responses.

Figure 5.9: Persistence diagrams of pre-course and post-course survey questions with two different random samples of 200 students' responses.

Chapter 6

Conclusion

The goal of this study was to analyze a dataset using topological data analysis. The dataset came from a national study of Calculus I students in the United States. After describing the dataset and going into detail the ways in which the dataset has already been analyzed by the researchers involved in the study, there was a discussion of topological data analysis. The goal of topological data analysis is to identify the persistent homologies found in the dataset in order to get a sense of some of the underlying structure in the data. In order to do topological data analysis of this dataset, adaptations had to be made. For instance, all of the data needed to be given a numerical value and some of the data was normalized in order to enable consistent interpretations for different input variables. Moreover, two or three variables were often chosen in order to be able to use plotting software to get a sense of how to interpret the homologies at various dimensions that were visible in the persistence diagrams. Unlike the higher dimensional input, point clouds produced from two or three variables could be plotted in order to interpret the results in context. However, hopefully the results from analyzing the lower dimensional data will provide insights into how to interpret the persistence diagrams created from higher dimensional input. In this study, no final conclusions were made about high dimensional input because of the fragility of choosing random subsets of students' responses.

Nevertheless, the following conclusions can be made as a result of this study. First, the results indicate some of the structural patterns and common points exist when using discrete data; this study provided some insights into how to interpret these patterns. For instance, the prevalence of the point $(1, \sqrt{2})$ is due to the discrete nature of the points and the prevalence of clustered low frequencies of responses. Second, topological data analysis seems to be able to indicate shifts in data that occur with relatively small effect sizes. Third, if one loop persists while many others die quickly in a case where frequencies of combinations are also considered, the resulting persistence diagram may indicate favored responses. Fourth, if including frequencies of combinations, several loops with varying levels of persistence may indicate that the predictability power of one variable on the other is low. Lastly, when using large numbers of variables with topological data analysis, it is important to consider how different sample sizes will affect the resulting persistence diagrams.

Future studies could aim at exploring more of the results that come from using topological data analysis on discrete datasets. Some efforts to use statistical measures before or after topological data analysis is in progress, but further explorations into making this form of analysis rigorous is another direction for future research. Overall, this study is an exploratory first step to finding a way to bring topological data analysis into mathematics education research.

Bibliography

[1] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer. Persistent homology analysis of brain artery trees. *ArXiv e-prints*, November 2014.

[2] Valentina Berti, Lisa Mosconi, and Alberto Pupi. Brain: Normal variations and benign findings in fdg pet/ct imaging. *PET Clin.*, 9(2):129–140, 2014.

[3] College Board. Statistical abstract of undergraduate programs in the mathematical sciences in the united states: Fall 2010 cbms survey, 2010.

[4] David Bressoud, Vilma Mesa, and Chris Rasmussen. *Insights and recommendations from the MAA national study of college calculus.* The Mathematical Association of America, Washington, D.C., 2015.

[5] AB Costello and JW Osborne. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. pract assess res eval 2005; 10. *URL http://pareonline. net/getvn. asp*, 10:1–9, 2011.

[6] Vin De Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.

[7] Herbert Edelsbrunner and John L. Harer, editors. *Computational topology: An introduction.* American Mathematical Society, United States, 2009.

[8] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[9] Allen Hatcher. *Algebraic topology.* Cambridge University Press, 2002.

[10] Jeremy Kun. The čech complex and the vietoris-rips complex. http://jeremykun.com/2015/08/06/cech-vietoris-rips-complex/, 2015.

[11] Hyekyoung Lee, Moo K Chung, Hyejin Kang, Bung-Nyun Kim, and Dong Soo Lee. Discriminative persistent homology of brain networks. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 841–844. IEEE, 2011.

[12] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *pHom: Compute persistent homology*, 2013. R package version 1.14.4 — For new features, see the 'Changelog' file (in the package source).

[13] A. E. Maxwell. Factor analysis. In Samuel Kotz, Campbell B. Read, N. Balakrishnan, and Brani Vidakovic, editors, *Encylopedia of statistical sciences*, volume 3, pages 2–8. John Wiley and Sons, 2006.

[14] William Mendenhall and Terry Sinich, editors. *A second course in statistics: Regression Analysis*. Prentice Hall, United States, 7 edition, 2012.

[15] JR Munkres. Simplicial complexes and simplicial maps. *Elements of Algebraic Topology*, pages 7–14, 1993.

[16] Marc Offroy and Ludovic Duponchel. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta*, 910:1–11, 2016.

[17] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International Journal of Medical Education*, 2(7):53–55, 2011.