Differences in Personality Structure by Age: Analyzing Clusters with Persistent Homology

Jake Howell, BS Mathematics (May 2018), Texas Christian University

Faculty Advisor: Dr. Eric Hanson

Simplicial Homology

(This introduction to simplicial homology is based on [2]).

In order to study the shape of our data, we need to have some idea of connected components. In topology, we can determine the number of components of a simplicial complex. Informally, a simplicial complex is a collection of simplices with a couple qualifying conditions. To fully understand the definition of a simplicial complex, we first need to understand simplices. A 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, etc. Let's look at some examples:

Figure 1

Figure 2

In figure 1, there exist three 0-simplices and one 1-simplex. In figure 2, there exist five 0-simplices, six 1-simplices, and zero 2-simplices. The "rabbit ears" formed in figure 2 are not 2-simplices as the triangles are not filled. Now that we have an understanding of simplices, we can define a simplicial complex:

Def. A **simplicial complex** K is subspace of $\mathbf{R}^n$ together with a finite list of simplices such that:

1. The union of the simplices is the set K and each point in K lies in the interior of only one complex;
2. Every face of every simplex in the list is also a list.

Essentially, a *k-simplex* is described by a list of k+1 vertices $v_0,\ldots,v_k$ (points in some $\mathbf{R}^n$). To ensure that our k-simplex is actually k-dimensional, we assume each of these vertices are in general position. Thus we define a **k-simplex** as the smallest convex subspace of $\mathbf{R}^n$ containing a given list of k+1 vertices in general position, written as $[v_0,\ldots,v_k]$. We require that each vertex has dimension no greater than k-1. If a vertex has dimension v-1, call it a **face**. Next, we want to further describe a simplicial complex by defining its boundary:

Def. Given a simplicial complex K, define $\mathbf{S_n(K)}$ as the set of all n-simplices of K. Define the **boundary** of each element in $S_n(K)$ as the list of elements in $S_{n-1}(K)$.

To illustrate this definition, consider our 2-simplex, call it J:

We can write J = [a,b,c]. Not that $S_2(J) = \{[a,b,c]\}$. Then, the boundary of J is $S_1(J) = \{[a,b], [b,c], [a,c]\}$ = the line segments (1-simplices) forming the 2-simplex. We want to form a boundary operation $S_n(K) \longrightarrow S_{n-1}(K)$. However, the boundary of $S_n(K)$ is a list of simplices, not a single simplex. So, we need some new terminology.

Def. Let $\mathbf{C_n(K)}$ be the collection of all subsets of $S_n(K)$. Note that the boundary of an n-simplex is in $C_n(K)$.

Another way to interpret $C_n(K)$ is that it is the $\mathbf{Z}/2$ vector space spanned by $S_n(K)$. Thus, we can now define a boundary operation:

Def. Call the linear transformation δ_n: $C_n(K) \longrightarrow C_{n-1}(K)$ the **boundary operator**, given by the formula

$$\delta_n[v_0,\ldots,v_n] = \sum_{i=0}^{n}[v_0,\ldots,\widehat{v_i},\ldots,vn]$$

For example, let's apply δ_n to our 2-simplex, J:

$$\delta_n[a,b,c] = [b,c] + [a,c] + [a,b]$$

We can compose these boundary operations to form the **chain complex**:

$$\ldots \longrightarrow C_n(K) \xrightarrow{\delta_n} C_{n-1}(K) \xrightarrow{\delta_{n-1}} C_{n-2}(K) \longrightarrow \ldots \longrightarrow C_1(K) \xrightarrow{\delta_1} C_0(K)$$

Now we have a way to detect simplices that *could* be boundaries. This is the basis of homology, which we shall define:

Def. The *n*th **homology group** of a simplicial complex K is the quotient

$$H_n(K) = Z_n(K) \,/\, B_n(K)$$

with $\mathbf{Z_n(K)} = \mathrm{Ker}\delta_n$, which we will call **cycles** (all possible candidates for the boundary). Define $\mathbf{B_n(K)} = \mathrm{Im}\delta_{n+1}$, which is the set of all **boundaries** of an n+1simplex.

The **homology** of K is the collection $H_*(K) = \{H_0(K), H_1(K), H_2(K),\ldots\}$

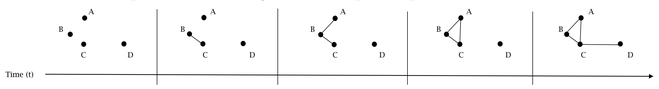
For convenience, define $Z_0(K) = C_0(K)$.

To better understand this, let's look back at Figure 1, our first example. Let's call this complex M.

Since the highest degree simplex is 1, we can just examine $C_1 \longrightarrow C_0$. Observe that dim C_0 = 3 (there exist three 0-simplices) and dim C_1 = 1 (there exists only one 1-simplex). When we apply δ_1 to C_1, we see that it is mapped to the addition of the two 0-simplices creating the 1-simplex. Thus, Im δ_1 = 1. Therefore, dim H_0 = = dim [C_0 / Im δ_1] = 3-1 = 2. This is the number of connected components. From a similar computation, we find that H_1 = 0.

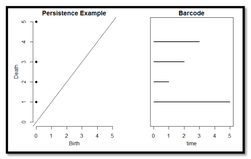
Persistent Homology

One way to perform cluster analysis is to define a simplicial complex on a data set and determine its connected components. A simplicial complex can be defined by connecting two points in the data with an edge if there pairwise distance is less than some value ε and then completing the complex by filling in any higher dimensional faces. By varying ε the data can then be studied at various scales. For example, consider the following set of nested simplicial complexes:



What persistence homology analyzes is how long certain simplices *persist*, i.e. the amount of time for which they exist. For this example, we could represent our initial dataset with a pairwise distance matrix. We can then create a persistence and barcode diagram, which we will learn to interpret in the analysis section.

	A	B	C	D
A	0	1	2.5	4
B	1	0	2	4.5
C	2.5	2	0	3
D	4	4.5	3	0



Data

To study personality structure, psychologists have commonly applied clustering techniques on questionnaire data. In [2], Costa et. al performed cluster analysis with the Cattell's Sixteen Personality Factor Questionnaire. The subjects were 969 adult male volunteers divided into three age groups: 25 to 34, 35 to 54, and 55 to 82 (from now on we shall call these the young, middle, and old age groups respectively. Below is a table outlining the personality traits measured by the 16PF Questionnaire (adapted from [3]):

Primary Factor	Descriptors of Low Range	Descriptors of High Range	Primary Factor	Descriptors of Low Range	Descriptors of High Range
Warmth (A)	Reserved, detached, formal	Warm, outgoing, easy-going	Vigilance (L)	Trusting, accepting	Suspicious, skeptical
Reasoning (B)	Concrete thinking, unable to handle abstract problems	Fast learner, abstract-thinking	Abstractedness (M)	Solution-oriented, steady, conventional	Imaginative, absent minded, impractical
Emotional Stability (C)	Reactive emotionally, easily upset, changeable	Emotionally stable, adaptive, mature	Privateness (N)	Open, guiltless, naive	Discreet, non-disclosing, shrewd, astute
Dominance (E)	Cooperative, humble, obedient, accommodating	Dominant, forceful, assertive, aggressive	Apprehension (O)	Self-assured, unworried, confident	Self doubting, worried, insecure
Liveliness (F)	Serious, restrained, introspective	Animated, spontaneous, enthusiastic, impulsive	Openness to Change (Q1)	Traditional, conservative	Experimental, liberal, free thinking
Rule-consciousness (G)	Expedient, self-indulgent, nonconforming	Dutiful, conscientious, conforming, moralistic	Self-Reliance (Q2)	Group-oriented, affiliative	Solitary, resourceful, individualistic
Social Boldness (H)	Shy, timid, hesitant	Venturesome, thick skinned, uninhibited	Perfectionism (Q3)	Tolerates disorder, flexible, impulsive	Organized, compulsive, self-disciplined
Sensitivity (I)	Objective, self-reliant	Sentimental, tender	Tension (Q4)	Relaxed, tranquil, patient	High energy, driven, frustrated

Analysis

The authors of [2] provided us with a pairwise correlation matrix for the questions [A,B,C,…,Q3,Q4]. We converted these to a pairwise distance matrix by taking the absolute value of the correlations * 100 and subtracting them from 100.

For example, a correlation coefficient of .45 would represent a distance of 55 in our matrix: (100 – (|.45|*100)).

A correlation of -.90 would have a distance value of 10 in our matrix: (100 – (|-.90|*100)).

We did this to say that two variables are *far away* from each other if their correlation is weak (i.e. near 0). Once this matrix was created, we utilized the *TDA* package [4] in R to create the persistence diagrams and barcodes to the right.

The persistence diagram visualizes the "birth" and "death" of each variable in our matrix. Since our questionnaire has 16 questions, we have 16 variables. Since all of our variables exist from the beginning, they all have a birth value of 0. The death value of a variable is the "time" a component "dies". In the context of persistent homology, this represents the distance parameter ε such that a component connects to another component. Note that all points in the plot must lie above the birth-death line; a variable cannot die before it is born. The red triangles in the persistence diagram represent topological holes in the dataset, which may or may not have psychological significance.

The barcode diagram visualizes how many connected components exist at a specific time t. Each black bar represents a connected component, and the total number of black bars represent the number of connected components. For example, at t=0, each barcode has 16 bars, which makes sense since we started with 16 variables. The length of the black bar represents how long a component *persists*. When some component connects to another, we lose one black bar. At t=100, we expect to have only one black bar remaining, since every variable should be connected. The same can be said for any red bars; these represent the existence and persistence of any topological holes.

To gain a better understanding of these diagrams, let's analyze the barcode for the young age group. At t = 0, we have 16 bars, as to be expected. As t passes through the high 20's, we see that we lose two black bars. This corresponds with the two points in the persistence diagram; the components "die" at this time. We interpret this as two different components connecting with another component at this distance parameter. Thus, we go from 16 to 14 connected components. Once t is approximately in the mid-70s, we see that we have only three black bars left. This means that there exist only three connected components in our simplicial complex.

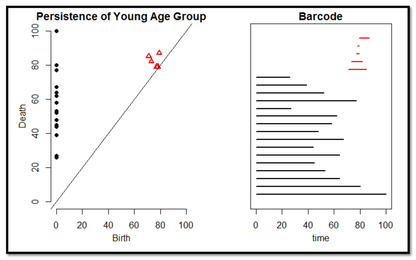
After creating our persistence diagrams, we wanted to be able to analyze which components were connected at a specified distance parameter. To do this, we created a function that transformed our pairwise distance matrix into an adjacency matrix. If two components are further than a specified distance *d*, we say that they are not connected and enter a 0 into that entry of the matrix. If two components are within distance *d* of each other, we say that they are connected and enter a 1 into the matrix. This creates an adjacency matrix for a specified distance.

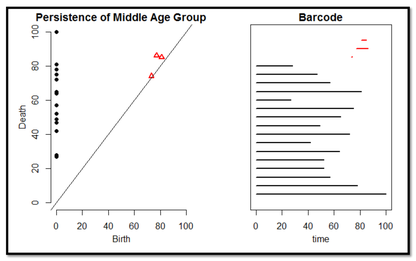
Utilizing the *igraph* library [5] in R, we converted the adjacency matrices into graphs. This allowed us to determine which components were connected at the specified distance. This is how we formed our clusters; we can even specify the correlation value at which each cluster is created. "Important" results are listed in the three tables below.

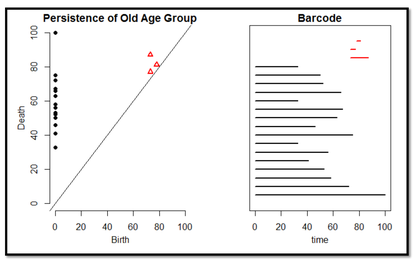
Let's analyze the "important" distances for the young age group. At d = 0 (correlation coefficient of +/- 1.00), all variables are singletons. This means that there are 16 "clusters", but each contain only one point (hardly a cluster). At d = 40, we see that there is a significant cluster forming. The variables C, O, Q3, and Q4 form a cluster for correlation coefficients of +/- .60. All other "clusters" are singletons. At d = 59 (correlation of +/- .41), we see that there are now three significant clusters. These persist until d = 65, at which the first and second cluster connect. This indicates that, for the young age group, personality structure appears to follow a three-cluster (or three-factor) model. At d = 81 we have one connected component.

While the tables are useful, we would like to be able to visualize the clusters at certain distance parameters. To do so, we simply plotted the graphs formed from the respective adjacency matrices. Observe the following graphs:

*All computations were done in the statistical computing language R.





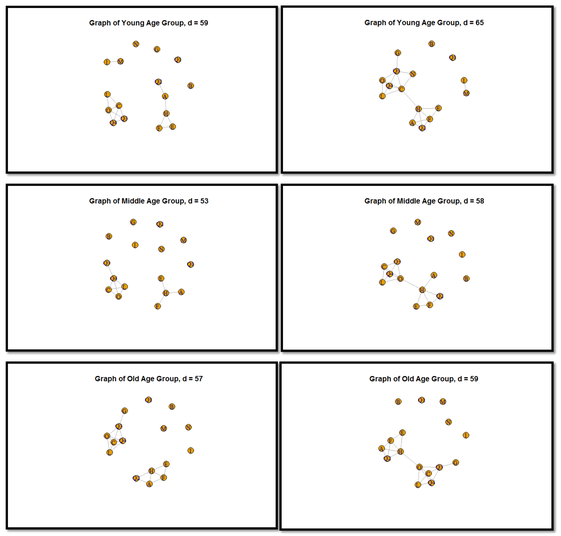


Distance Value	Young Age Group Clusters
d = 0	Only singletons, i.e. 16 components
d = 40	[C,O,Q3,Q4], all else singletons
d = 59	[C,L,O,Q3,Q4], [A,E,F,H,Q2], [I,M], all else singletons
d = 65	[C,L,N,O,Q3,Q4,A,E,F,H,Q2,G], [I,M] , all else singletons
d = 81	One connected component, i.e. everything is connected

Distance Value	Middle Age Group Clusters
d = 0	Only singletons, i.e. 16 components
d = 53	[C,O,L,Q3,Q4], [A,E,F,H], all else singletons
d = 58	[C,O,L,Q2,Q3,Q4,A,E,F,H], all else singletons
d = 66	[C,G,O,L,Q2,Q3,Q4,A,E,F,H], [I,M], all else singletons
d = 82	One connected component, i.e. everything is connected

Distance Value	Old Age Group Clusters
d = 0	Only singletons, i.e. 16 components
d = 34	[C,O,Q4] , [F,H], all else singletons
d = 57	[C,G,L,O,Q3,Q4] , [A,E,F,H,Q2], all else singletons
d = 69	[C,G,L,O,Q3,Q4,A,E,F,H,Q2], all else singletons
d = 76	One connected component, i.e. everything is connected

Analysis cont.



For the young age group, we visualize the graphs for d = 59 and d = 65.

For the middle age group, we visualize the graphs for d = 53 and d = 58.

For the middle age group, we visualize the graphs for d = 57 and d = 59.

Note: These graphs **DO NOT** accurately represent distance between components nor their "location". This visualization simply shows which components are connected. While this may not accurately portray distance, it does give us some information though. For example, in the old age group graph for d = 57, since E is connected to F but not connected I, we can say that the distance from E to I is greater than the distance from E to F. However, these graphs do not accurately portray *how much* father E is from I.

Comparisons and Conclusion

In [2], the authors claim that each age group can be defined by three clusters. They claim that the first two clusters remain relatively the same regardless of age, and that the third cluster changes significantly. We compare these to what we find to be the most significant clusters:

	Their Clusters	Our Clusters (with d)
Young Age Group	[C,O,L,Q3,Q4], [A,E,F,H,Q2], [I,M]	[C,O,L,Q3,Q4], [A,E,F,H,Q2], [I,M] (d = 59)
Middle Age Group	[C,L,O,Q3,Q4], [A,E,F,H,Q2], [M]	[C,L,O,Q3,Q4], [A,E,F,H] (d = 53)
Old Age Group	[C,G,L,O,Q3,Q4], [A,E,F,H,Q2], [B,I,M,Q2]	[C,L,O,Q3,Q4], [A,E,F,H,Q2] (d = 57)

Based on our analysis, we agree with the statement that the first two clusters change very little with respect to age. However, we claim that, for the middle and old age groups, the dataset is best described by two clusters rather than three. As we can see in the graphs, for the middle and old age groups, the first two clusters connect before a third cluster is even formed. This seems to support our claim that the last two age groups fit a two-cluster system better.

There are still questions of how to interpret these clusters and the psychological significance of this cluster-based approach. Similarly, we can ask questions about the psychological significance of the topological holes found in the dataset. However, what we can say is that there appears to be a difference in personality structure based on age. In future analyses we plan to analyze further the relationship between personality and age.

References

[1] Conn & Rieke, The 16PF Fifth Edition Technical Manual, Institute for Personality and Ability Testing, (1994).

[2] Costa, P. T., & Mccrae, R. R. (1976). Age Differences in Personality Structure: a Cluster Analytic Approach. *Journal of Gerontology, 31*(5), 564-570.

[3] M.D Crossley, Essential Topology, Springer, (2005).

[4] Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, Vincent Rouvreau. The included GUDHI is authored by Clement Maria, Dionysus by Dmitriy Morozov, PHAT by Ulrich Bauer, Michael Kerber and Jan Reininghaus. (2016). TDA: Statistical Tools for Topological Data Analysis. R package version 1.5. https://CRAN.R-project.org/package=TDA.

[5] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

****Special thanks to Dr. Cathy Cox of the TCU Psychology Department for her contributions to this project.**