2013

Clustering Methods and Their Applications to Adolescent Healthcare Data

Morgan Mayer-Jochimsen
Scripps College

Clustering Methods and Their Applications to Adolescent Healthcare Data

Morgan Mayer-Jochimsen

Deanna Needell, Advisor
Winston Ou, Reader

Submitted to Scripps College in Partial Fulfillment
of the Degree of Bachelor of Arts

March 15, 2013

Department of Mathematics

Abstract

Clustering is a mathematical method of data analysis which identifies trends in data by efficiently separating data into a specified number of clusters so is incredibly useful and widely applicable for questions of interrelatedness of data. Two methods of clustering are considered here. K-means clustering defines clusters in relation to the centroid, or center, of a cluster. Spectral clustering establishes connections between all of the data points to be clustered, then eliminates those connections that link dissimilar points. This is represented as an eigenvector problem where the solution is given by the eigenvectors of the Normalized Graph Laplacian. Spectral clustering establishes groups so that the similarity between points of the same cluster is stronger than similarity between different clusters. K-means and spectral clustering are used to analyze adolescent data from the 2009 California Health Interview Survey. Differences were observed between the results of the clustering methods on 3294 individuals and 22 health-related attributes. K-means clustered the adolescents by exercise, poverty, and variables related to psychological health while spectral clustering groups were informed by smoking, alcohol use, low exercise, psychological distress, low parental involvement, and poverty. We posit some guesses as to this difference, observe characteristics of the clustering methods, and comment on the viability of spectral clustering on healthcare data.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to extend thanks to all of the incredible professors at the Claremont Colleges who have supported me in my education of math and social justice and who encouraged me in finding my place between the two. My classes have inspired me in learning about the various approaches and knowledge systems to understand the multitude of factors interacting to shape the world we live in.

I am tremendously thankful for my wise, patient, and generous advisor, Deanna Needell, who helped to inspire this research and assisted me throughout the process. I was fortified by her support in general and her advice from personal experiences combining skill in mathematics with the "real world."

Thank you to Winston Ou for his role as my second reader, but perhaps more importantly for his support as a friend and mentor since my first days at Scripps. I would also like to thank Jerry Grenard for his incredible willingness to help with access to the California Health Interview Survey adolescent data set, finding applicable, interesting sources, and providing inspiration in the direction of my data analysis. Finally, I am thankful for the inspiring Scripps students that I am fortunate enough to learn from every day and for the love and support of my family and friends.

Personal Motivation

This thesis, like many student works at Scripps, is interdisciplinary in motivation. In fact, the topic of this section was informed by a reading from a Philosophy course I took at Scripps in the Fall of 2012. In Feminist Ethics, we read a compelling piece by Joyce Trebilcot about the importance of an author sharing motivations with her audience so that the readership can more expertly understand the author's opinion and the conclusion reached [12]. Trebilcot argued that understanding an author's history and viewpoint is vital to the most comprehensive reading of their work. As such, I begin by briefly situating my reader in the background to and inspiration for this work.

Math has always been a comfortable realm for me, however, I have been a "well-rounded" (now interdisciplinary) woman for all of my years as a student and came to college with the intention to attend medical school after graduation. Beginning in math and the sciences felt natural, but I was soon pulled more strongly toward classes that investigated problems with humans, rather than with cells. I changed my plan from practicing medicine to working in the field of public health so that I could dedicate myself to worldwide human issues of health inequality. As my Scripps education exposed me to the many issues people face in addition to securing their health, such as sexism and racism in everything from immigration policies to the prison industrial complex, the media, and every day life, my scope of future work was opened further still. My plan is now the least specific that it has ever been, but I feel confident that this thesis in mathematics as applied to health data is a great first step into the future.

Math is captivating and powerful, but is most interesting and useful to me when applied to something tangible. Due to this fact and my interest in contributing meaningful work that is helpful to people, this thesis investigates K-means and spectral clustering, methods that I believe can

be applied to healthcare data. To my advisor, Deanna Needell's, and my knowledge, spectral clustering has yet to be extensively applied to issues of health. With the availability of the California Health Institute Survey data of teen respondents, I am excited to put this kind of application forth as an option.

Chapter 1

Introduction

1.1 Purpose and Motivation

Clustering is an interesting method of determining trends in data. This thesis serves first to illuminate the mathematics and methodology of K-means and spectral clustering, then applies these clustering techniques to the California Health Interview Survey (CHIS) data from adolescent respondents. The CHIS data set is specifically useful to research on healthful interventions to combat unhealthy behaviors and future health risks in teens. In a society where chronic diseases are becoming increasingly prevalent, this work is motivated by the project of meaningfully clustering the CHIS adolescent data to provide insight as to how to specifically tailor health interventions for adolescents based on the health characteristics that clustering uses to identify them as unhealthy or at risk of leading unhealthy adult lives.

1.2 Introduction to Clustering

We begin with a verbal overview intended to situate the reader in the methods and goals of clustering before providing rigorous definitions in Chapter 2. Clustering is a mathematical problem concerned with separating objects, points, or other data into meaningful groups. When separated through clustering, nodes assigned to the same cluster are more similar to each other than they are to nodes assigned to other clusters, for some definition of similarity. Clustering is often used on data for which there is little prior information because implementation does not require many assumptions to be made on the data. Clustering is extremely useful to applications

across the spectrum of data analysis, where researchers are often concerned with finding trends, locating patterns, uncovering similarities, and making predictions. A few examples of clustering applications include identifying the various components of an image as investigated in [10], which could be relevant to medical imaging, and determining which employees at a company are satisfied and which are at risk of leaving, which could help companies determine strategies towards achieving low employee turnover [2].

In determining which data points best fit in which cluster, we do not wish to cluster through enumeration, or "by trial and error," because this is time-consuming and combinatorially difficult. This enumeration problem is well represented by the Stirling number of the second kind, denoted $S(n, k)$, which gives the number of ways to partition n objects into k distinguishable and nonempty groups, here, clusters. The Stirling number of the second kind is given by [9]

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n. \tag{1.1}$$

In the later parts of this thesis, 3294 people are clustered into 2 groups. By the enumeration represented in (1.1) this would be $S(3294, 2) \approx \infty$ ways the people could be partitioned into 2 clusters, which is unrealistic analysis to perform and evaluate. So, we turn to a discussion of the various established clustering methods used here. The primary interest of this thesis is spectral clustering, but K-means clustering is included for comparison, because Mistry et al. used K-means on the 2003 CHIS data [7], and due to the fact that spectral clustering utilizes K-means in the latter-most steps of the algorithm if the data are being clustered into more than 2 groups.

1.3 Overview of Clustering Methods

Clustering is characterized by the ability to handle a wide range of complex data. Due to this diversity, there are a multitude of ways to portray the data or objects, to define similarity, and to create the groups mentioned in the previous section. Clustering is complicated by this subjectivity, since there is no single algorithm or approach that is adequate to solve each of the multitude of potential clustering problems [6]. As a result, there are many different clustering approaches and algorithms. Two methods are utilized here, K-means and spectral clustering.

1.3.1 Overview of K-means

Where data are represented in an $m \times n$ matrix and are graphed in n-dimensional space, K-means creates K clusters from the data where clusters are defined by their center, called the centroid. The centroid of a cluster is a representation of the means of the variables that make up the points it is nearest to. The distance from the centroid to vertices in n-dimensional space is measured by the Euclidean distance formula. K-means works to minimize this distance, which is effectively the distance between the vectors representing the vertices of a group and the vector of the centroid of that group.

In an iteration of minimizing distances, if a node in cluster 1 is closer to the centroid of cluster 2, the vertex will instead be assigned to cluster 2. The centroid of the first cluster will move away from this vertex as the means change due to that point's variables being removed from the calculation of the centroid. When this process is complete, we have K clusters defined by their centroids where maximized similarity between the vertices of a cluster and its centroid is indicated by a minimized square error between the two [6]. K-means is easily carried out by standard software due to its foundation in linear algebra.

1.3.2 Overview of Spectral Clustering

The spectral clustering process similarly begins by considering data as vertices in an n-dimensional space. For example, if we consider the data set consisting of people and their age, weight, and height, we would have points that exist in 3-dimensional space. Then we wish to create a graph, the diagram that results from representing data as nodes joined or unjoined by lines, which are called edges. In this implementation of spectral clustering, we choose to begin by connecting all points and then proceed by adding weight to the edges (an action which can be conceived of as making the connecting lines thicker) based on how similar the points are. If a photograph is being examined, a good measure for similarity may be the color of the pixels. The closer in color two points are, the more weighted the edge between them will be, and the more likely they are to be grouped together. Then the data are grouped by maintaining the connections that have the most weight, but eliminating edges of minimal weight to separate the points into disjoint clusters. This "cut" is achieved through eigenvector analysis, one of the defining characteristics of spectral clustering. This allows spectral clustering problems to be efficiently computed by current

linear algebra software [13].

Organization
Chapter 2 presents the theorems and mathematical concepts of K-means and spectral clustering necessary to an understanding of their application and concludes with the algorithm for implementation on the CHIS data. The chapter ends with a comparison of the characteristics of the two clustering methods. Then, Chapter 3 introduces the CHIS data, its uses, and the results of the clustering algorithms. Finally, Chapter 4 contemplates the practical issues encountered during this work and suggests potential future analyses in the intersection of spectral clustering and health. The MATLAB code written by the author to carry out analyses on the data is included in the appendices.

Chapter 2

K-means and Spectral Clustering

2.1 Graph Theory Definitions

Both K-means and spectral clustering are informed by a conception of data in a graph. First, data are organized in an $m \times n$ matrix, A, featuring m objects and their n attributes. Each vector of this matrix $(x_{m1}, x_{m2}, ..., x_{mn})$ defines an individual by their characteristics. Then, a graph G is a pair (V, E) where V is the set of vertices and E is the set of unordered pairs denoting edges between the vertices of V. Each vertex is a row of A, so v_m = $(x_{m1}, x_{m2}, ..., x_{mn})$ and E is the pairwise set of vertices (v_i, v_j). The graphing of points as vertices and establishing edges that connect them occurs in n-dimensional space.

We note two commonly used approaches to connect the points of G with edges. The graph informed by the k-nearest neighbors method requires the choosing of a parameter k, where a vertex v_i is connected to the k vertices that are nearest to it. Here, we prefer the fully connected graph where edges are established between all vertices. The fully connected graph produces an undirected graph and allows for all points to be evaluated in the formation of the clusters.

2.2 K-means Clustering

K-means is defined by clustering data into K groups and by the concept of the centroid μ_k, a point in the n-dimensional space of the graph that demarcates the center of a cluster. Specifically, the centroid is given by the

mean of each of the n measurements, so is a vector of length n. For μ_1, the centroid for cluster 1,

$$\mu_1 = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_n).$$

We will have K centroids for each of the K groups to be identified. The centroid is a useful tool in organizing clusters since it gives a defining set of characteristics for each group, which also makes the centroid useful in data analysis.

Initially, the centroids are randomly placed by the K-means algorithm, then enhanced by iterations to create clusters in such a way that the Euclidean distances between the vertices of a group and the centroid of that group are minimized and are smaller than the distances of these vertices to the centroids of other clusters [11]. In terms of our m vectors in n-dimensional space, $v_i = (x_{i1}, x_{i2}, ..., x_{in})$, K-means works to minimize a sum of squares cost function, alternatively referred to as the sum of the squared intra-cluster distances [14][2]:

$$\sum \sum ||v_i - \mu_k||_2^2.$$

When the sum of squares function is minimized after an established number of iterations, the clusters are established. The algorithm used in K-means implementation on data is given by the MATLAB command `kmeans`.

2.3 Spectral Clustering

We derive spectral clustering through graph theory organized in matrix notation. We begin with data represented as points in the graph G and fully connected by edges from E. Then, we wish to weight the edges of the graph so vertices representing similar data are connected with an edge that is weighted more heavily than the edge between dissimilar vertices.

We call this representation a Similarity Graph and denote the similarity between vertices v_i and v_j, s_{ij}. W, the Weighted Adjacency Matrix, contains the pairwise weights from the similarity function, so $W(i, j) = s_{ij}$. Similarity can be defined in a multitude of ways, but a good, commonly used function to determine weightedness is the Gaussian similarity function,

$$s_{ij} = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right). \tag{2.1}$$

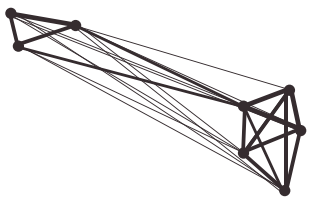

Figure 2.1: Similarity Graph Example

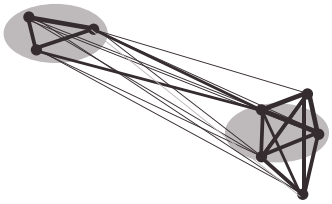

Figure 2.2: Similarity Graph with Neighborhoods

2.3.1 Considerations on σ

The Gaussian similarity function is normalized by the tuning parameter σ, which works in a similar way to the parameter k in the k-nearest neighbor graph. σ identifies a neighborhood which provides information to the Gaussian function in its establishment of appropriate weights, s_{ij}. Within a neighborhood, the edges between points are given substantial weight, while edges to vertices outside the neighborhood are assigned negligible (yet still positive) weight [13]. σ is thus an important parameter with substantial influence on the determination of clusters. Unfortunately, there are few guidelines to choosing σ other than trial and error. σ is also problematic in that the use of a constant σ stipulates that the data must be on the same scale, or clustering will not provide good groups since the same sized neighborhood will not be appropriate everywhere on the graph [8]. In 2.2, we see that the neighborhood used to identify the cluster on the left is insufficient to capture the characteristics and points included in the cluster on the right.

Perona and Zelnik-Manor write on the effects of σ on clustering in [8]. They conclude that rather than establishing a constant σ, it is better to set a σ for each point. It makes sense intuitively to define each neighborhood by the points it includes since data may be more dense in some areas in the graph than in others. A local scaling σ also augments the ability of spectral clustering algorithms to handle complex data through appropriate analysis of every node.

Through empirical and theoretical analysis, Perona and Zelnik-Manor give a successful alternative to the constant σ, instead defining the neighborhood by a local scaling parameter, $\sigma_i\sigma_j$. Generally, σ_i is defined as the Euclidean distance between vertex v_i and v_k where v_k is the kth nearest node to v_i. In a multitude of different experimental settings, the authors found that k $= 7$ resulted in good clustering. So, we borrow this parameter and change the Gaussian similarity function to

$$s(x_i, s_j) = exp\left(\frac{-||x_i - x_j||^2}{(\sigma_i * \sigma_j)}\right),$$

where $\sigma_i = d(x_i, x_{7^{th}})$, $\sigma_j = d(x_j, x_{7^{th}})$, and $x_{7^{th}}$ is the 7th nearest neighbor to point x_i or x_j.

Once we have established W and the Similarity Graph, we create the Degree Matrix, D, which contains on its diagonal the total weight of the connections of each vertex,

$$D = diag(d_i), \tag{2.2}$$

where $d_i = \sum_{k=1}^{m} w_{i,k}$. D represents the overall connectedness of each vertex to all others.

We are now ready to eliminate some edges from the fully connected graph to create distinct clusters of connected points. Following [10], we can conceive of making two clusters as splitting our graph $G = (V, E)$ into two disjoint sets A and B where $A \cup B = V$ and $A \cap B = \emptyset$. To achieve this separation, we "cut" the edges that maintain connections between the two sets. We measure the weight of our cuts by summing the similarity of the eliminated edges in what is called the cut.

$$cut(A, B) = \sum_{\substack{v_a \in A \\ v_b \in B}} W(v_a, v_b) \tag{2.3}$$

The optimal clusters will be derived by minimizing cut so that the formation of clusters occurs by eliminating minimally weighted edges rather than edges between nodes that have high similarity. Shi and Malik enhance cut within their proposed calculation, $Ncut$, an unbiased measure which considers the value of a cut between A and B as a fraction of the total connections between all nodes in the graph. The use of $Ncut$ eliminates the risk that cut will make a trivial cut and produce a cluster consisting of a single vertex in A. $Ncut$ is given by

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}, \tag{2.4}$$

where $assoc(A,V)$ is the total connection between nodes in A to all other nodes in the graph,

$$assoc(A,V) = \sum_{\substack{v_a \in A \\ v_i \in V}} W(v_a, v_i),$$

and cut as defined as in (2.3).

Unfortunately, minimizing $Ncut$ is NP-hard [10], so instead, Shi and Malik represent $Ncut$ with an eigenvector problem where solving $Ncut$ is equivalent to solving

$$(D-W)y = \lambda D y,$$

for y, the matrix of eigenvectors. D is diagonal Degree Matrix with $d_i = \sum_{j=1}^{n} W(i,j)$ on the diagonal (2.2) and W is the symmetrical Weighted Adjacency Matrix where $W(i,j) = s_{ij}$ with s_{ij} defined in (2.1). This eigenvector problem can be reformulated as

$$D^{-1/2}(D-W)D^{-1/2}z = \lambda z,$$

where the Graph Laplacian Matrix, $(D-W)$, is a representation of the similarity graph. We see that the solution results from normalizing the Graph Laplacian, giving the Normalized Graph Laplacian, L,

$$L = D^{-1/2}(D-W)D^{-1/2}.$$

We refer the reader to [10] for details on the derivation of these formulas and the mathematical explanations for the approximation of $Ncut$ by this

eigenvector problem. For our purposes, it is important to note that z_0 is the smallest eigenvector of L with an eigenvalue of 0 and z_1 is the second smallest eigenvector of L which gives the solution to the $Ncut$ problem [10]. Nodes are partitioned into groups by identifying a dividing threshold in the second smallest eigenvector, z_1, and assigning the points above the threshold to one cluster and the points below the threshold to the other, effectively cutting the connections between individuals on opposite sides of the threshold. Examples are given in Figure 2.4.

2.3.2 The Clustering Properties of the Eigenvector

The reason why the second eigenvector of the Normalized Graph Laplacian gives the threshold for separating data into two clusters is not completely obvious. It is similarly unclear why other eigenvectors are useful in further separating the data. Deep mathematical theory (involving the Rayleigh quotient) that is beyond the scope of this work can be utilized to indicate the role of the second eigenvector as seen in [10]. Superficially, eigenvectors point in the "right" directions and provide intuition about the actions or properties of a function.

For example, consider a function F that changes a circle into an oval, vectors v_1 and v_2 define how F works on the circle.

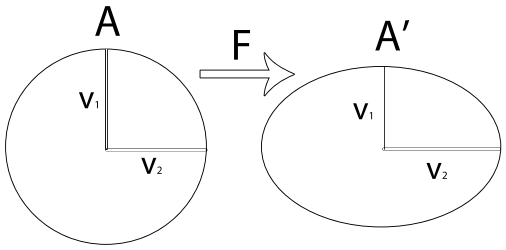

Figure 2.3: Eigenvector Conceptualization

In spectral clustering, the eigenvectors to the Normalized Graph Laplacian give direction to the data in the Weighted Adjacency Matrix (which informs the Laplacian and is used in creating the similarity graph), so if the similarity graph indicates two directions or two groups in the data, the second eigenvector will represent them.

In the ideal case, there is an obvious break in the second eigenvector that clearly defines the two clusters in the data. Sometimes, however, the

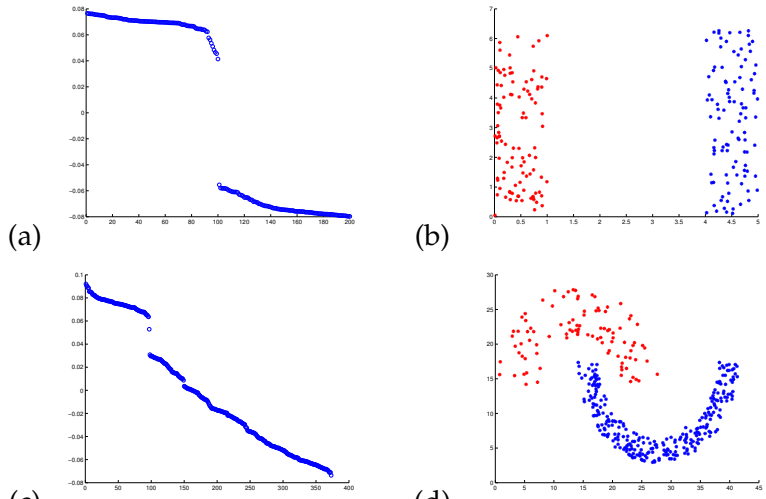

Figure 2.4: Local Scaling σ producing good clustering results on Toy Examples using (a) Local Scaling σ Eigenvector for the Gaussian Distribution, (b) Local Scaling σ on the Gaussian Distribution, (c) Local Scaling σ Eigenvector for Interlocking U's, (d) Local Scaling σ on Interlocking U's

second eigenvector takes on continuous values and to create two clusters, we must identify a splitting point. One option for the threshold is the median of the eigenvector, but a better option is to search for the splitting that gives the best (most minimal) $Ncut(A, B)$ value for the resulting partition by testing potential thresholds and choosing the value with the best $Ncut$ as the splitting point [10]. Sometimes a continuous eigenvector indicates that there is no clear divide to the data we aim to separate, in some instances it implies the scaling parameter σ is poorly chosen, and still other times it may simply represent a complicated data set, where we must work harder to conceive of the separation. In establishing clusters we can check their validity in part by the divide in the eigenvector, looking for a clear, visible break or a very small $Ncut$ value for the chosen threshold. We can also check algorithms generally by applying them to empirical examples where the distributions are known.

While this thesis is primarily concerned with clustering into two groups, we note that the third eigenvector can be used to further partition the first two clusters. After the threshold in the second eigenvector is identified or created and two clusters are established, then the threshold in the third

eigenvector is found and used to subdivide the first two clusters.

Alternatively, the first K eigenvectors can be found and used as representatives of the characteristics of the individuals. Then K-means is run on this $m \times K$ matrix to find K clusters in the data. If 4 groups are desired, K-means should be run on the first 4 eigenvectors produced by the spectral clustering algorithm as the 4 new "attributes" of the people to be clustered.

Algorithm 1 Spectral Clustering for Two or K Clusters, Input data matrix A, Output clusters

1: Define A as an $m \times n$ matrix of data with m objects and n attributes.

2: Find B, an $m \times m$ matrix of pairwise distances characterized by 0's along the diagonal. Arrange the distances by row in ascending order in matrix bx.

3: Create W, the $m \times m$ Weighted Adjacency Matrix of pairwise similarities where

$$W(i,j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right)$$

or

$$W(i,j) = \exp\left(\frac{-||x_i - x_j||^2}{bx(i,7) * bx(j,7)}\right)$$

4: Create D, the $m \times m$ Degree Matrix of the similarity graph which has elements d_i along its diagonal where d_i is equal to the total similarity of point i to all others, $d_i = \sum_{k=1}^{m} w_{i,k}$.

5: Solve

$$D^{-1/2}(D - W)D^{-1/2}z = \lambda z$$

for z, the matrix of eigenvectors by column of the Normalized Graph Laplacian matrix.
Identify z_1, the second eigenvector and order its values

or

identify the first K eigenvectors, order their values, and place them in the mxK matrix Z.

6: Use the break in the second eigenvector to separate the objects into two groups

or

run `kmeans(Z,K)`.

2.4 Comparison of Clustering Methods

Parameters
Both K-means and spectral clustering are vulnerable to choices in parameter. Any time a clustering problem is considered, the number of clusters into which the data is partitioned. This can cause problems for researchers, since clustering is often used when there is little prior information on the distribution of data. In lower dimensions it is possible to plot clusters and visually determine the effectiveness of the clustering, but when clustering with large data sets, accuracy is generally difficult to define since there is most often no notion of what constitutes a "correct" cluster.

Ae also have the parameter σ in spectral clustering that can drastically change clustering results. If σ is too large, the algorithm will maintain too many connections between the data creating large groups of slightly similar nodes. If σ is too small, the clustering algorithm may miss large, general trends in the data. The measure of the appropriateness of σ can sometimes be understood by looking at the innate thresholding characteristics of the second smallest eigenvector, but, again, is generally only clear through data analysis.

Scales of Data
Spectral clustering most often features either the k-nearest neighbor similarity graph or the fully connected graph, which are both characterized by handling data on different scales very well [13]. K-means clustering, on the other hand, most often does poorly with data on different scales, since if one of the initial centroids does not land near a set of points on a scale different from the majority of the data, it may never "make it over" to those points. This problem with K-means is compounded by its tendency to create clusters of similar size.

We see that whenever Euclidean distance is used, which is often throughout both the K-means and spectral clustering methods of analysis previously discussed, the formation of clusters is dominated by those attributes that are largest in scale [6]. To achieve a data set where all variables are on the same scale, researchers may choose to manipulate the data.

Assumptions on and Complexity of Data
Spectral clustering evaluates each point so that the entire data set informs the formation of groups [13], which allows spectral clustering to group complex data where a model of the distribution of data is unknown [8]. This property of spectral clustering is encompassed in references to spec-

tral clustering as divisive, meaning all nodes are first connected in one large cluster and then edges are cut to arrive at the desired number of disjoint groups [6]. Alternatively, due to K-means performing clustering by establishing the centers of clusters first, it does not function as well on non-linear, complex models since it first establishes connections, then changes them to minimize Euclidean based distance measures. It is important to note that the added considerations of spectral clustering lead to a longer run time than K-means, which is executed very quickly and easily, especially given the `kmeans` command in MATLAB. This has led to a popularity of K-means due to its easy implementation (which is especially useful on large data sets) and general success when clusters are of comparable size [6][2].

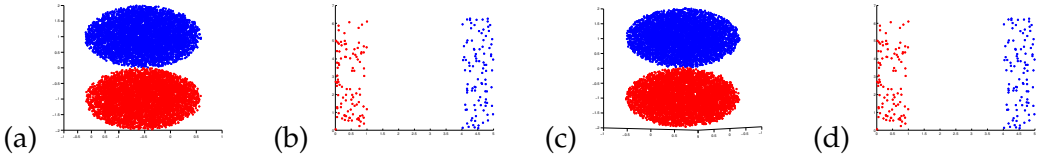

Figure 2.5: Comparison of K-means and Spectral Clustering on Empirical Examples with (a) and (b) by K-means and (c) and (d) by spectral clustering producing the same results

Empirical Examples

When examining the effectiveness and behavior of K-means on toy examples in 2.5 and 2.6, we note that K-means is successful when it can linearly divide the data. So while K-means works identically to spectral clustering for Spheres and Gaussian, images in which a line can clearly separate the clusters, spectral clustering outperforms it on Eye and Interlocking U's.

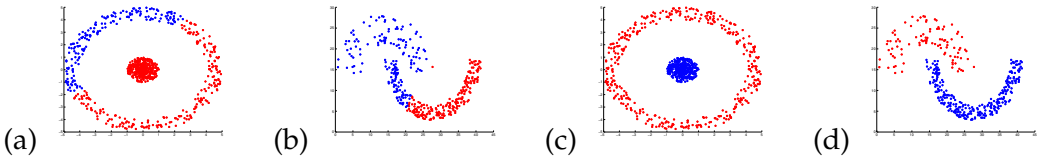

Figure 2.6: Comparison of K-means and Spectral Clustering on 2D and 3D Empirical Examples with (a) and (b) failing by K-means and (c) and (d) succeeding by spectral clustering

Chapter 3

Application to the California Health Interview Survey

3.1 Description of Data

The California Health Interview Survey (CHIS) is a telephone survey of California residents performed by the UCLA Center for Health Policy Research. It is the largest health survey in the nation and is performed through a random dial of landlines and cellular phones [4]. The survey provides critical data for wide usage by researchers, health professionals, and policymakers [5].

The survey includes 186 questions on qualitative and quantifiable health-related behavior and indicators. The answers to qualitative questions such as "Do you currently smoke cigarettes?" are represented numerically for data analysis purposes. Often "yes" is coded as a 1 and "no" is denoted with a 2, however there is no steadfast rule to the numerical assignments. Other questions are represented differently, for example insurance type is entered as 1: uninsured, 5: Medicaid, 6: healthy families, 7: employment-based, 8: privately purchased, 9: other public. Quantitative data are recorded as given.

The CHIS 2009 data dictionary notes that interviewers achieved a high level of responses on the majority of questions and that most questions have missing responses for less than 2% of the sample [4]. There are of course exceptions, such as household income, where there were initially valid missing responses for more than 20% of the people interviewed. The data dictionary describes that where missing responses were identified, values were imputed using two methods of impution. The first technique

was a random selection from the observed distribution and was used when the percentage of items missing from a question was very small. Hot deck imputation was also used, a process by which information from a similar person is imputed to the respondent with missing data where similarity was defined by household and individual characteristics. Once a "similar" individual had been used, they were eliminated from further use in the hot deck imputation process.

The survey includes responses from adults, adolescents, and children (given by their guardians), but only the data from adolescent respondents is utilized here. Among the questions asked to adolescents, some were proposed to all respondents and others specifically geared to the adolescent population. Further, there were many questions specific to adolescent subpopulations, such as those diagnosed with asthma. As given, before any manipulation of data by the author, there were 3,382 adolescents interviewed on 186 health related questions. A few overall characteristics of the data are that males are more often smokers, females are more likely to exercise infrequently and experience psychological stress, and all adolescents are likely to consume less than the recommended servings of fruits and vegetables per day.

3.2 Informing Work

The Center for Disease Control (CDC) studies on adolescents indicate that adolescent health is largely informed by obesity, reckless or violent behavior, poverty, substance use, and healthcare access. The limited research on the clustering behavior of adolescent health data indicates clusters are largely defined by gender and psychological distress [7]. CHIS data on adolescents has previously been clustered in [7] where Mistry et al. apply K-means clustering to the 2003 CHIS adolescent data set using smoking, alcohol use, low fruit/vegetable consumption, physical activity, gender, parental involvement, parental supervision, adult role models, age, race, income, and parental education as variables informing the analysis. In group formation, unhealthy behavior is identified as smoking, using alcohol, infrequent exercise, and low fruit and vegetable consumption. The authors first separate data by gender, then perform K-means to create 4 clusters for each gender which they classified (in order of healthfulness) as Salutary Adherents, Active Snackers, Sedentary Snackers, and Risk Takers. None of the individuals in the Salutary Adherents group displayed any of the unhealthy behaviors identified by the authors. Active Snackers are

distinguished by adolescents who consume few servings of fruits and vegetables, but exercise frequently, while Sedentary Snackers are characterized as exercising infrequently with a smattering of low fruit and vegetable consumers, cigarette users, and alcohol users. Finally, the Risk Takers cluster includes many individuals currently using alcohol and rarely consuming fruits and vegetables. Some Risk Takers use alcohol and exercise rarely.

The authors are specifically concerned with the gendered characteristics of the clusters, which they characterize as gender specific in that males and females engage in risk-taking behavior, perform exercise, experience depression, and respond to parental involvement in different ways and at different rates. On average, females have poorer psychological health, greater parental supervision, and lower physical activity levels than males. Generally adult involvement decreases the risk of being in an unhealthful cluster for both genders [7].

This analysis shares with Mistry et al. a motivation to identify adolescent populations which engage in unhealthy lifestyles so that health outreach programs can be tailored to the specific behaviors that are indicative of low health and to the groups of adolescents which could most benefit from healthful interventions. It is also valuable for researchers and health professionals to be aware of the correlations between multiple behaviors and risk of an unhealthy lifestyle. On this front, Mistry et al. conclude that all interventions planned for adolescents should address a multitude of overlapping and interacting health behaviors. In the two unhealthy clusters, for example, they identify low fruit and vegetable consumption is correlated with low physical activity or alcohol use, so interventions should address these overlapping issues simultaneously.

3.3 Implementation of Clustering

This analysis takes its lead from Mistry et al. in using the variables utilized in their successful clustering analysis: gender, cigarette use and alcohol use, physical activity, consumption of fruits and vegetables, exercise, parental involvement, adult role models, age, race, income, and parental education. The variables and their values that determine a healthy or unhealthy classification are given in Table 3.1. These variables also align with CDC reports of accurate indicators of adolescent health, substance use, diet, exercise, and psychological health [3].

Table 3.1: Variables and Their Meaning for Clustering

Variable	Healthy	Unhealthy
Smoking	Doesn't Smoke	Current Smoker
Alcohol	Has never had more than a few sips of alcohol	Has ever had more than a few sips of alcohol
Exercise	Exercises for an hour 4, 5, 6, or 7 days of week	Exercises for an hour 0, 1, 2, or 3 days of week
Fruits and Vegetables	5 or more servings yesterday	Less than 5 servings yesterday
Psychological Distress	Feeling nervous, hopeless, restless, depressed, apathetic some, few, or none of the last 30 days	Feeling nervous, hopeless, restless, depressed, apathetic most or all of the last 30 days
Hero	Admires and wants to be like someone	Does not admire and want to be like someone
Adult Involvement	Parent is present during after school hours and knows where adolescent goes at night	Parent is not present after school and does not know where adolescent goes at night

3.3.1 Data Characteristics and Manipulations

While the 2009 CHIS Data Dictionary describes each health attribute as having very few missing values, there are a fair number of individuals in the adolescent data set where responses are labeled unknown for some of questions. Due to the relatively small number of these individuals compared to the total (88 of 3382), they were eliminated to arrive at the 3,294 individuals used in the following analysis. Additionally, 55 questions were asked only to subpopulations and were coded as inapplicable (denoted with a -1) for the remaining set of adolescents. An example of such a question is, "Do you have a written copy of your asthma plan?" Only 295 adolescents answered this question, while the remaining 3,084 individuals, a substantial 91% of respondents were assigned -1 for inapplicable. In some cases, the percentage of inapplicable answers was as high as 97%.

Since spectral clustering relies heavily on the attributes of individuals, it is concerning to utilize data where some variables have entries that are, in a way, meaningless. Further, we do not necessarily want two individuals to be seen as more similar by the algorithm because they did not have specific experience with the same subset specific questions. Since the analysis here did not utilize all of the 300 measurements, most of these problematic variables were eliminated in the paring down of the data. For the few others, the approach used here was to change inapplicables to the answer deemed appropriate (if applicable), which in most cases was ”no”. This approach was used on questions in the same vein as, ”Have you used an inhaler in the last 12 months?” No other manipulations were required since the variables used in analysis were all fairly similar in scale.

3.4 Analysis and Results

3.4.1 How Many Clusters?

Other than examining subtle indicators or completing a full analysis of clustering results, there is no set method of analysis to determine if the clusters produced are ”correct,” because each instance of clustering is unique and we wish for clustering to give information on unknown distributions. However, this can make the selection of the number of groups to cluster a difficult decision. It is an important one though, because the number of groups will change how the patterns in the data are recognized and interpreted.

In order to gather information on the clustering tendencies of the 3294 individuals and 22 variables, clustering was performed in a variety of combinations of 2, 4, and 20 groups based on strictly K-means, K-means on 4 spectral clustering eigenvectors, and K-means on 20 spectral clustering eigenvectors. The results provide some subtle indications for the CHIS adolescent data as seen in Table 3.2. By ”stable” and ”volatile” we refer to the ability of K-means to produce the same results each time it is run on a data matrix, A. Strictly K-means on 2 groups is very stable, while strictly K-means clustering to create 4 groups is very volatile, leading to the conjecture that 2 groups are more clearly identifiable in the data than 4 groups. K-means on 4 eigenvectors for both 2 and 4 groups is stable, but places a trivial number of people in some of the groups; with 2 clusters, there are only 7 people assigned to group 1 with the remaining 3287 individuals in group 2. Similarly for K-means on 4 eigenvectors for 4 clusters, only 0.24%

of the data are clustered into group 3 and group 4 with 99.76% of the data in clusters 1 and 2, which does not provide much useful information about overall trends in the data, but may indicate a propensity of the data to exist in two clusters. Both K-means as applied to create 2 and 4 clusters produced volatile results when used on 20 spectral clustering eigenvectors, which may indicate that using many more eigenvectors than needed (here, 18 and 16 respectively) introduces noise that prevents the clustering algorithms from working effectively. Out of curiosity, the K-means algorithm was run on different numbers of spectral clustering eigenvectors to create 20 clusters. The resulting groupings are volatile and often create groups of less than 4 individuals, which is not entirely helpful when our aim is to identify large scale patterns in the data.

Table 3.2: Stability or Volatility of Clustering Attempts

Method	Two Clusters	Four Clusters
Strictly K-means	Stable	Volatile
K-means on 4 eigenvectors	Stable, but trivial groups	Stable, but trivial groups
K-means on 20 eigenvectors	Volatile	Volatile

Other than these inferences, we must run analysis on the clusters produced by the clustering algorithms in order to determine if they provide useful information about the data. Considering the information we gathered from the stability and volatility of clustering attempts along with the idea that the simplest groupings may provide the clearest distinctions between the resulting groups, we move forward in analyzing the composition of the data in two clusters provided by strictly K-means and strictly spectral clustering. Two clusters may also allow us to see large patterns of adolescents who are healthy or not.

3.4.2 Two Groups by K-means

The K-means function in MATLAB was used to cluster the 3294 x 22 data matrix A into 2 groups. Group 1 has 1710 individuals (888 males and 822 females) and group 2 has 1584 individuals (935 males and 749 females). The tendency of K-means to create clusters of the same size is true here. It is interesting to note that the clusters also have similar numbers of males and females. The composition of the groups was analyzed by running χ^2

tests on the data to examine if the proportion of adolescents in each cluster reporting a certain variable significantly varies from the expected frequencies given the sizes of the two groups.

Group 1 is characterized by adolescents who have no hero ($p < 0.0001$), exercise infrequently ($p < 0.0001$), and come from impoverished families ($p < 0.0001$) as well as females who experience psychological stress ($p = 0.0036$). Group 2 has no specific characteristics other than the inclusion of adolescents exhibiting the opposite behaviors of cluster 1 (in more often having a hero, exercising, being financial security, and having low levels of psychological stress). Individuals who smoke, consume alcohol, consume low levels of fruits and vegetables, and have low parental involvement or supervision are distributed insignificantly between the clusters.

Table 3.3: K-means: Percentage of adolescents by cluster that exhibit unhealthy behaviors

All	Low Exercise	Psychological Distress	No hero	Impoverished
Cluster 1	17.91	3.73	23.16	31.44
Cluster 2	12.57	2.13	15.73	5.41
Females	Low Exercise	Psychological Distress	No hero	Impoverished
Cluster 1	21.45	5.16	23.36	31.44
Cluster 2	14.32	2.67	14.51	5.41
Males	Low Exercise	Psychological Distress	No hero	Impoverished
Cluster 1	14.68	2.44	22.98	30.06
Cluster 2	10.97	1.63	16.83	5.40

As to the meaning of the clustering, we can conjecture that adolescents in impoverished homes may experience more stress than other adolescents due to their being held to standards in school for which they lack the resources to attain and the potential extra burdens of childcare for younger siblings, working outside of the home to support the family, or providing emotional support. These factors would also decrease the time available to the individual for physical activity. These guesses combined with the fact that individuals who cannot identify a hero are more often in cluster 1 may indicate that the adolescents are separated mainly due to their psychological health. Cluster 2 would then represent adolescents with low levels of psychological stress.

This clustering result may be useful to health professionals in the field of psychology or to middle and high school counselors in helping to

determine adolescents who could benefit from counseling services. Since adolescents may not be forthcoming in issues of mental health, this clustering by K-means could indicate the propensity of an individual to be assigned with the group associated with psychological stress and low exercise. The result may also be helpful in studies of the effects on poverty on adolescents in supporting the notion that financial insecurity is a causal factor in stress.

3.4.3 Two Groups by Spectral Clustering

The spectral clustering algorithm was run to create two groups using the local scaling 7th nearest neighbor parameter for σ in the Gaussian. However, this clustering attempt failed. All eigenvectors of the Normalized Graph Laplacian were equal to zero, giving no information on the data. This is potentially a result of clustering integer data where repeats in data points cause the 7th nearest neighbor to be the same node as the one being analyzed. This problem is fully discussed in Chapter 4.

Instead, a constant parameter was established, $\sigma = 5$, a choice made according to the parameters found appropriate for the toy examples and by a good shape of the resulting second eigenvector. Unfortunately, the second eigenvector of the Normalized Graph Laplacian for our 3294 individuals and 22 characteristics does not represent the ideal case in that it is a continuous eignevector, as seen in Figure 3.1. Thus, a threshold at which to split the individuals into groups was found by the $Ncut$ method detailed above, (2.4), where different splitting points are tested and a threshold is chosen by the splitting point that minimizes $Ncut$ [10]. 40 potential thresholds were tested and the point that minimized Ncut was found to be individual 2399 of 3294 with an Ncut value of 0.9973. So, the 2 clusters are defined by grouping the individuals above and below this point on the eigenvector, putting 2399 individuals in group 1 (1267 males and 1132 females) and 895 individuals in cluster 2 (456 males and 439 females). Interestingly, while the clusters are of different sizes, the number of males and females in each is very similar.

The statistical analyses used on the K-means clustering results were also applied to the spectral clustering groups to determine if specific variables caused adolescents to be assigned to the clusters in significantly different frequencies than expected, based on the size of the clusters. The results of the χ^2 tests indicate that spectral clustering on the adolescent data holistically grouped the individuals into a comprehensively unhealthy cluster, cluster 1, and a more healthful group, cluster 2. For both gen-

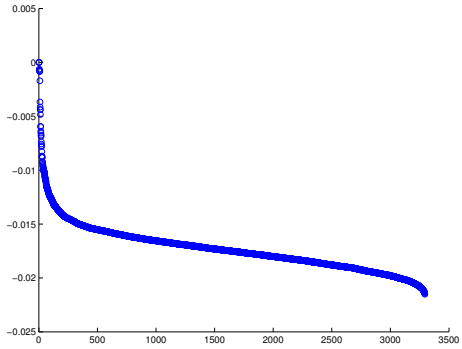

Figure 3.1: Spectral Clustering Second Eigenvector

ders, group 1 is characterized by smoking ($p < 0.0001$), low exercise ($p < 0.0001$), psychological distress ($p < 0.0001$), and poverty ($p < 0.0001$). Within the variables that define overall adolescent group assignment, smoking is a significantly more important factor for males ($p = 0.0032$), while exercise ($p < 0.0001$) and mental health ($p < 0.0001$) are more determining for females. Beyond these four decisive variables, the clusters are determined in part according to gender-exclusive variables (with higher, yet still significant, p-values than the variables most determining in cluster formation). While low adult supervision has no significant effect on female placement in either cluster, when looking specifically at males, low adult supervision increases the risk of being placed in the unhealthy cluster ($p = 0.0227$). Females who have ever consumed more than a sip of alcohol are also more likely to be placed in cluster 1 ($p = 0.0390$) while alcohol has no significant effect for males. Factors that are non-determining in the creation of clusters are adolescents not having a hero and a low consumption of fruits and vegetables, which is likely due to the fact that the majority of individuals in the data set are classified as low consumers.

These spectral clustering results may be useful for organizations seeking to determine general causal factors of unhealthy adolescents. Campaigns for healthy lifestyles could tailor their efforts to the variables indicative of unhealthy behaviors as given by spectral clustering, such as exercise and substance use. The results could also be used indicate to physicians the status of overall health of their teen patients, which may close the gap in communication of information about adolescents' habits.

Table 3.4: Spectral clustering: Percentage of adolescents by cluster that exhibit unhealthy behaviors

All	Smoking	Alcohol Use	Low Exercise	Psych. Distress	Low Parental Involvement	Poverty
Cluster 1	4.12	25.05	26.87	5.74	15.15	29.87
Cluster 2	0.12	7.65	3.61	0.12	4.40	6.25
Females	Smoking	Alcohol Use	Low Exercise	Psych. Distress	Low Parental Involvement	Poverty
Cluster 1	3.06	24.19	31.06	7.64	14.83	29.98
Cluster 2	0.06	7.51	4.71	0.19	4.77	6.87
Males	Smoking	Alcohol Use	Low Exercise	Psych. Distress	Low Parental Involvement	Poverty
Cluster 1	5.11	25.83	23.04	4.00	15.44	29.77
Cluster 2	0.17	7.78	2.61	0.06	4.06	5.69

3.4.4 Comparison of Clustering Analyses

Clustering by both K-means and spectral clustering appear successful in that the variables most affecting group formation are among those presented by the CDC as high indicators of adolescent health [3]. Further, the differences between healthy and unhealthy individuals in the clusters were significant for at least 4 variables in each clustering method. The clusters identified by spectral clustering depended on 6 variables: smoking, alcohol use, exercise, psychological distress, low parental involvement, and poverty. Other than the overlap of substance abuse in smoking and alcohol use, these variables seem to present a inclusive view of the health of the adolescent. On the other hand, the unhealthier cluster identified by K-means is characterized by 4 variables: exercise, psychological distress, no hero, and poverty. Two of these variables overlap in the category of mental health, psychological distress and hero, and the other two, low exercise and poverty, provide information about the lifestyle of the adolescent. In general, the spectral clustering clusters were defined by being informed by more variables and having differences between the clusters that were more significant than those of the K-means algorithm. It is possible that the tendency of K-means to create closely sized groups based on linear data caused these variables to be more important to clustering than others. Then, we may infer that the patterns of relationship between the variables determining the spectral clustering groups may be more complex.

Three of the variables overlap between the clustering approaches: exercise, psychological distress, and poverty, which may indicate that these variables are the most deterministic of adolescent health attributes, which may be concerning to mathematicians and social scientists alike as the variables affecting overall teen health. Due to the different characteristics of the clustering methods, the overlap between them may be the best starting point for future work, since K-means has been proven to cluster healthcare data before and spectral clustering has been argued as a better approach to clustering problems. Both the results from the groups evaluated individually and the conclusions from the combined results indicate a successful application of clustering to health care data.

In addition to differences between the results of K-means and spectral clustering, we can compare the analysis here to the K-means clustering of the 2003 CHIS data in [7]. A first obvious difference is that K-means clustering on the 2003 data produced clusters very well characterized by the factors noted by the authors. Very high percentages of adolescents reported the behaviors of the group they were clustered into (even reaching 100% in some cases). While the differences noted in the clustering results on the 2009 data are significant, the variables do not divide the groups as starkly, seen in Tables 3.3 and 3.4. Also, while this analysis shares with [7] cluster formation informed by smoking, alcohol use, and low exercise, a final determining variable in Mistry et al. is low consumption of fruits and vegetables, which is not seen in the analysis of the 2009 data. Instead, poverty and psychological health play a more important role. This difference in results when the current K-means clusters are compared to the 2003 K-means groups may be a product of the 6 years between data collection changing health concerns for adolescents, or the non-inclusion of low fruit and vegetable consumption here may be a result of most adolescents in the data set reporting low consumption. Additionally, the process of analysis here differs in utilizing the spectral clustering algorithm to group data and not separating the data before it is clustered, so as not to assume gendered differences in the data and to be able to utilize gender as another informing variable in clusters creation.

Chapter 4

Looking Forward

4.1 Practical Problems

Using Data Collected by Others
Problems may occur in the analysis of data collected by other researchers. In general, there may be missing data or the data may be coded inappropriately for a specific kind of analysis. While the subset of data used here alleviated many of the problems in the larger set, a researcher aiming to cluster the entire data set would have to work through the missing values and the many instances of ”inapplicable” as an answer. Depending on the kind of analysis performed, the inapplicable answers may not matter, but if they were problematic, imputation could potentially be used. Another characteristic of the 2009 CHIS adolescent data set is a high variability in the number of questions asked on each health topic. While there are four questions to determine cigarette use, there are 15 related to asthma. With respect to these variables, clustering the data will result in groups that are defined more by characteristics related to asthma than to smoking. This problem of over-representation of certain variables could potentially be remedied through careful scaling.

The general differences in scale in the data could be an issue since the spectral clustering algorithm works by grouping people with similar numerical profiles and involves the Euclidean distance formula which weights larger data more heavily in grouping considerations. It is consequently important to have data within a similar range. For example, while the answer to the question ”Are you currently taking physical education in school?” is coded with a 1 for ”no” and a 2 for ”yes,” the question ”How much do you weigh?” yields values as large as 200 lbs. These values will be weighted

differently by the Gaussian similarity function, so not only will some questions be given larger importance than others, but these values may appear dissimilar to others. Scaling is an easily fixed issue since the researcher can manipulate the data by a constant, but it will not necessarily by clear in which range all variables should exist.

7th Nearest Neighbor
The local scaling σ defined by the 7th nearest neighbor fails with the integer data used here. Because there are not the differences in data that result from non-contrived quantitative data, there is potential for repeated points, especially here when only 22 variables are used. Thus, in some cases, the 7th nearest point to a vertex was a point with the same coordinates, which inserted zeros into the Weighted Adjacency Matrix, a problem for further calculations of the Degree Matrix and the eigenvectors. This is a substantial issue, since the local scaling sigma given by the 7th nearest neighbor is one of few good options for choosing σ in the Gaussian. A possible solution may be to look at a different kth nearest neighbor for cases of integer data. An analysis could be devised in order to determine for a specific data set which nearest neighbor would be necessary to advance beyond the repetition of points. It is unclear, however, if this parameter would retain the good properties of the 7th nearest neighbor in characterizing useful neighborhoods and providing good clustering results. Alternatively, some randomness could be added to each integer, to make the data set look more like standard data, but this noise may interfere with clustering. A final potential solution is to add a dummy variable so that the points are recognized as different [2].

Number of Variables
The researcher must determine the number of variables to include in a clustering analysis. There are definite benefits to including many variables since more information is then given to the clustering algorithm for better, more useful clusters on the objects the variables describe. However, the run time of the algorithm will be greatly increased by the inclusion of many attributes, especially the time to create W, the Weighted Adjacency Matrix and to find the eigenvectors of the Normalized Graph Laplacian. Using few variables decreases computation time and allows for the visual representation of data, so provides visual confirmation that clustering has or has not worked (as in the toy examples). However, using few variables may result in a failure of the clustering algorithm if the variables chosen do not correlate in a significant way. Initially, for ease of computation, clus-

tering analyses were run on three variables from the CHIS adolescent data as informed by [3]. Self-health analysis, smoking in the household, nervousness, body mass index, and serious psychological stress were used in different combinations of three, all of which gave second eigenvectors equal to zero. This could indicate that the question that clustering was asking on these variables was poorly chosen or that three variables are simply insufficient with abstract, repeated, integer data.

4.2 Future Work

CHIS Adolescent Data
There are many potentials for the application of clustering to the 2009 CHIS data set other than the analysis performed here. A first, obvious, idea for future work is to use different variables in order to answer a more specific question. Since part of the purpose of this work was to see if spectral clustering produced meaningful results on healthcare data, no specific question informed the analysis on the 22 variables. This could provide interesting results to the healthcare community on specific adolescent health behaviors. For example, individuals with asthma could be assigned to one cluster and then other variables could be introduced into the algorithm to determine through clustering which health attributes are associated with asthma.

It could also be illustrative to intentionally weight the data according to the importance of certain variables to overall health, as informed by prior research on adolescent health behaviors. An objective measure could be obtained from the Department of Health and Human Services websites on causes of death or illness. For example, it is possible that by weighting "Do you walk or bike to school?" less than "How many cigarettes do you smoke per day?" could provide more informative indications on the overall health of adolescents.

Other Applications
The CHIS adolescent data set was not available at the onset of research for this project, so a different application was considered on Alzheimer's disease. Alzheimer's is important to study because it is a prevalent disease in the elderly population of the US (a segment of the population which is growing) and the disease remains much of a mystery to health professionals and the research community. It is possible that spectral clustering could provide information about the predictors of the disease to inform early detection of and preventive treatments for individuals likely to de-

velop Alzheimer's. A variety of research has already been performed on individual characteristics or experiences that increase the likelihood of developing Alzheimer's. Factors shown to increase the risk of Alzheimer's are having immediate family members with Alzheimer's and head injuries. Clustering may be able to help prove or disprove the influence of potential risk factors including stroke, high blood pressure, diet, social engagement, education, environmental toxins, and medications [1].

4.3 Conclusion

K-means and spectral clustering are informative methods of analysis that are easy to understand and to implement. Both are successful at clustering healthcare data, although the results given by each are unique. According to the overlapping results of K-means and spectral clustering, exercise, psychological distress, and poverty are strong indicators of health for adolescent respondents in the California Health Interview Survey. This kind of work and these results have applications for researchers in mathematics and the social sciences as well as health professionals in determining risk factors for low health and in identifying topics for successful health interventions.

Appendix A

Spectral Clustering for Two Groups

```
%A = mxn Data Matrix
m = length(A);

%Create B, an mxm matrix of pairwise distances to be used in the
%'7th nearest' calculation of W. B is characterized by 0's along
%the diagonal.
B = zeros(m,m);
for j=1:m
   for k=1:m
        B(j,k) = norm(A(j,:)-A(k,:));
    end
end

%Then, we wish to order the distances, so rearrange each row of
%B (dim = 2) so that the values are in ascending numerical order.
[bx, ix]=sort(B, 2, 'ascend');

%Create W, the Weighted Adjacency Matrix of the Similarity
%Graph. W is an mxm matrix with elements sjk pairwise similarities.
%W is characterized by 1's along the diagonal.
W = zeros(m,m);
for j=1:m
    for k=1:m
    W(j,k) = exp(-norm(A(j,:)-A(k,:))^2 / (bx(j,7)*bx(k,7)));
```

```
    %or
    W(j,k) = exp(-norm(A(j,:)-A(k,:))^2 / (2*(sigma)^2));
    end
end

%Create the Degree Matrix of the Similarity Graph. D is
%an mxm matrix with elements dj along its diagonal.
%This gives the total similarity of a point to the others.
%These are our eigenvalues.
D = zeros(m,m);
for n=1:m
    D(n,n)= sum(W(n,:));
end

%Solve D^(1/2)*(D-W)*D^(-1/2)*V = lambda*V for V, the
%matrix of eigenvectors by column, where D is a diagonal
%matrix of eigenvalues and V is a full matrix whose
%columns are the corresponding eigenvectors. X*V = V*D.
X = (D^(-1/2)) * (D-W) * (D^(-1/2));
[V,D] = eig(X);

%We want the eigenvector of the second smallest eigenvalue,
%so we pick the second column and sort it for ease.
v = V(:,2);
[vsorted, ix] = sort(v,'descend');

%Plotting the sorted, second smallest eigenvector allows
%us to see clearly the break in values, which indicates
%where the 'cut' should be made.
figure
hold on
plot(vsorted, 'O');
hold off

%index for t, the threshold at which we make our cut
ixx = 1:m;
clust1 = ixx(v>=t);
clust2 = ixx(v<t);

%If there are few enough dimensions that we plot (Toy
```

```
%Examples), we create the original plot with colors
%according to groups
figure
hold on
if size(A,2) == 2
    plot(A(clust1, 1), A(clust1, 2), '*r');
    plot(A(clust2, 1), A(clust2, 2), '*b');
elseif size(A,2) == 3
    plot3(A(clust1, 1), A(clust1, 2), A(clust1, 3), '*r');
    plot3(A(clust2, 1), A(clust2, 2), A(clust2, 3), '*b');
end
hold off

%Otherwise, we create indices
group = zeros(m,1);
for i=1:t\\
    group(i) = 1;
end
for i=(t+1):m
    group(i) = 2;
end

%Combine the index and the group assignments and sort them in a
%useful order.
index=(ix, group);\\
scsort = sortrows(index,1);
sc2 = scsort(2) ;
```

Appendix B

Spectral Clustering for Four or Twenty Groups

```
% Spectral Clustering and then Kmeans clustering on CHIS data

%A is the 3294 x 22 matrix of adolescents and 22 healthcare
%questions
m = length(A);

%Create W, the Weighted Adjacency Matrix of the Similarity Graph.
%W is an mxm matrix with elements sjk pairwise similarities. W
%is characterized by 1's along the diagonal.
W = zeros(m,m);
for j=1:m
    for k=1:m
        W(j,k) = exp(-norm(A(j,:)-A(k,:))^2 / (2*(sigma)^2));
    end
end

%Create the Degree Matrix of the Similarity Graph. D is an
%mxm matrix with elements dj along its diagonal. This gives
%the total similarity of a point to the others. These are our
%eigenvalues.
D = zeros(m,m);
for l=1:m
    D(l,l)= sum(W(l,:));
end
```

```
%Solve D^(1/2)*(D-W)*D^(-1/2)*V = lambda*V for V, matrix
%of eigenvectors by column. Where D is a diagonal matrix of
%eigenvalues and V is a full matrix whose columns are the
%corresponding eigenvectors. So, X*V = V*D
X = (D^(-1/2)) * (D-W) * (D^(-1/2));
[v,d] = eig(X);
[E order] = sort(diag(d),'ascend');
V = v(:,order);

%We want the first four (twenty) eigenvectors as the four
%(twenty) attributes of our people we will cluster by kmeans

v1 = V(:,1);
[vsorted1,ix] = sort(v1, 'descend');
v2 = V(:,2);
[vsorted2,ix] = sort(v2, 'descend');
v3 = V(:,3);
[vsorted3,ix] = sort(v3, 'descend');
v4 = V(:,4);
[vsorted4,ix] = sort(v4, 'descend');
...
etc.

%n = 4 or 20
for i=1:n
figure
hold on
plot (vsortedi, 'O');
hold off

vk4 = [v1,v2,v3,v4];
vk20 = [v1,v2,v3,v4,v5,v6,v7,v8,v9,v10,v11,v12,v13,v14,v15,
           v16,v17,v18,v19,v20];

%Run kmeans
%4 eigenvectors, 4 clusters:
z44 = kmeans(vk4,4);
%4 eigenvectors, 20 clusters:
z420 = kmeans(vk4,20);
```

```
%20 eigenvectors, 4 clusters:
z204 = kmeans(vk20,4);
%20 eigenvectors, 20 clusters:
z2020 = kmeans(vk20,20);
%20 eigenvectors, 2 clusters:
z202 = kmeans(vk20,2);
%4 eigenvectors, 2 clusters:
z42 = kmeans(vk4,2);
```

Appendix C

Identifying Characterics of Cluster Members

```
%Variables of A, the mxn data matrix

%gender =1 for male and =2 for female
gender = A(:,17)

% smoking =1 for current smoker, 2 for not
smoking = A(:,2)

%alcohol =1 for more than a few sips of alc, 2 for not
alcohol = A(:,4)

%exercise = # of days in past week active for > 60 mins
%count <= 3 as low physical activity
exercise = A(:,7)

%fruits and veggies. fg = combined servings of fruits and
%vegetables yesterday
% fg < 5 indicates low consumption
f = A(:,5)
g = A(:,6)
fg = f + g

%Psych sums nervousness, hopeless, restless, depressed, apathetic,
%worthless with responses 1:all the time, 2:most of the
```

```
%time, 3:some of the time, 4:a little of the time, 5:not
%at all psych < 20 indicates psychological distress
psych = A(:,8)+A(:,9)+A(:,10)+A(:,11)+A(:,12)+A(:,13)

%Hero =1 for admires and wants to be like some person,
%=2 for no heroes
hero = A(:,14)

%Adult involvement sums frequency of an adult around
%during after school hours measured as 1:always, 2:most
%time, 3:some time, 4:almost never, 5:never and how
%much guardian knows about whereabouts at night
%measured as 1:a lot, 2:most time, 3:nothing, 4:doesn't
%go out at night.
%if adult > 3.5 we say there is little parental involvement
adultaround = A(:,15)
adultknows = A(:,16)

know = zeros(3294,1)
for k=1:3294
    if adultknows(k) == 4
        know(k) = 0;
    else know(k) = adultknows(k);
    end
end

adult = adultaround + know

%Poverty where 1:0-99% FPL, 2:100-199%, 3:200-299%, 4:300%
%and above. We say an adolescent lives in an impoverished
%home if poverty <=2
poverty = A(:,18)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%a = number of eigenvectors
%b = number of groups
%index = vector of the cluster assignments where index(i) =
%cluster assignment for individual i
```

```
%Gender
gen = zeros(a,b);
for i = 1:a
    for j = 1:b
    gen(i,j)=sum((index == i).*(gender == j));
    end
end
gen

%Smoking
smok = zeros(a,b);
for i = 1:a
    for j = 1:b
    smok(i,j)=sum((index == i).*(gender == j).*(smoking == 1));
    end
end

%Alcohol
alc = zeros(a,b);
for i = 1:a
    for j = 1:b
    alc(i,j)=sum((index == i).*(gender == j).*(alcohol == 1));
    end
end

%Exercise
exer = zeros(a,b);
for i = 1:a
    for j = 1:b
    exer(i,j)=sum((index == i).*(gender == j).*(exercise <= 1));
    end
end

%Fruits and Veggies
fveg = zeros(a,b);
for i = 1:a
    for j = 1:b
    fveg(i,j)=sum((index == i).*(gender == j).*(fg < 5));
    end
end
```

```
%Psychological Distress
psy = zeros(a,b);
for i = 1:a
    for j = 1:b
    psy(i,j)=sum((index == i).*(gender == j).*(psych < 20));
    end
end

%Hero
her = zeros(a,b);
for i = 1:a
    for j = 1:b
    her(i,j)=sum((index == i).*(gender == j).*(hero == 2));
    end
end

%Adult Involvement
adu = zeros(a,b);
for i = 1:a
    for j = 1:b
    adu(i,j)=sum((index == i).*(gender == j).*(adult > 3.5));
    end
end

%Poverty
pov = zeros(a,b);
for i = 1:a
    for j = 1:b
    pov(i,j)=sum((index == i).*(gender == j).*(poverty <= 2));
    end
end
```

Appendix D

Ncut Implementation

```
%The best option to choose the threshold in a continuous
%eigenvector is to search for the splitting that gives the best
%Ncut(A;B) value for the resulting partition. This is achieved by
%testing potential thresholds and choosing the value with the
%best Ncut as the splitting point.

%l = individual at which to divide the eigenvector
la = A((1:l),:);
lb = A(((l+1):3294),:);

Wcut = zeros(l,l);

for j=1:l
    for k=1:(3294-l)
        Wcut(j,k) = exp(-norm(la(j,:)-lb(k,:))^2 / (2*(5)^2));
    end
end

Wcut1 = sum(Wcut,1);
cutAB = sum(Wcut1,2);

WassocAV = zeros(l,l);

for j=1:l
    for k=1:m
    WassocAV(j,k) = exp(-norm(la(j,:)-A(k,:))^2 / (2*(5)^2));
```

```
    end
end

WassocAV1 = sum(WassocAV,1);
assocAV = sum(WassocAV1,2);

WassocBV = zeros((3294-l),(3294-l));

for j=1:(3294-l)
    for k=1:m
    WassocBV(j,k) = exp(-norm(lb(j,:)-A(k,:))^2 / (2*(5)^2));
    end
end

WassocBV1 = sum(WassocBV,1);
assocBV = sum(WassocBV1,2);

NcutAB = (cutAB/assocAV) + (cutAB/assocBV)
```

Bibliography

[1] Alzheimer's Association. 2012 Alzheimer's Disease Facts and Figures. *Alzheimer's and Dementia*, 8.

[2] B Cung, T Jin, J Ramirez, A Thompson, C Boutsidis, and D Needell. Spectral Clustering: An empirical study of approximation algorithms and its application to the attrition problem. *arXiv preprint arXiv:1211.3444*, 2012.

[3] Centers for Disease Control and Prevention. Youth Risk Behavior Surveillance - United States, 2011. Morbidity and Mortality Weekly Report 61.4, Office of Surveillance, Epidemiology, and Laboratory Services, June 8 2012.

[4] UCLA Center for Health Policy Research. California Health Interview Survey 2009 Teen Survey Data Dictionary. Los Angeles, CA, November 2011.

[5] UCLA Center for Health Policy Research. California Health Interview Survey. `http://healthpolicy.ucla.edu/chis/Pages/default.aspx`, 2012.

[6] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data Clustering: A review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[7] Ritesh Mistry, William J McCarthy, Antronette K Yancey, Yao Lu, and Minal Patel. Resilience and Patterns of Health Risk Behaviors in California Adolescents. *Preventive medicine*, 48(3):291–297, 2009.

[8] P. Perona and L. Zelnik-Manor. Self-Tuning Spectral Clustering. *Advances in neural information processing systems*, 17:1601–1608, 2004.

[9] Fred S. Roberts and Barry Tesman. *Applied Combinatorics*. CRC Press, Boca Raton, FL, 2009.

[10] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[11] Douglas Steinley. K-means Clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.

[12] Joyce Trebilcot. Ethics of method: Greasing the machine and telling stories. *Feminist ethics*, pages 45–51, 1991.

[13] U. Von Luxburg. A Tutorial on Spectral Clustering. *Statistics and computing*, 17(4):395–416, 2007.

[14] Hongyuan Zha, Chris Ding, Ming Gu, Xiaofeng He, and Horst Simon. Spectral Relaxation for K-means Clustering. *Advances in neural information processing systems*, 14:1057–1064, 2001.