MSCR 509: High Dimensional Analysis

Homework 11

Anish Shah

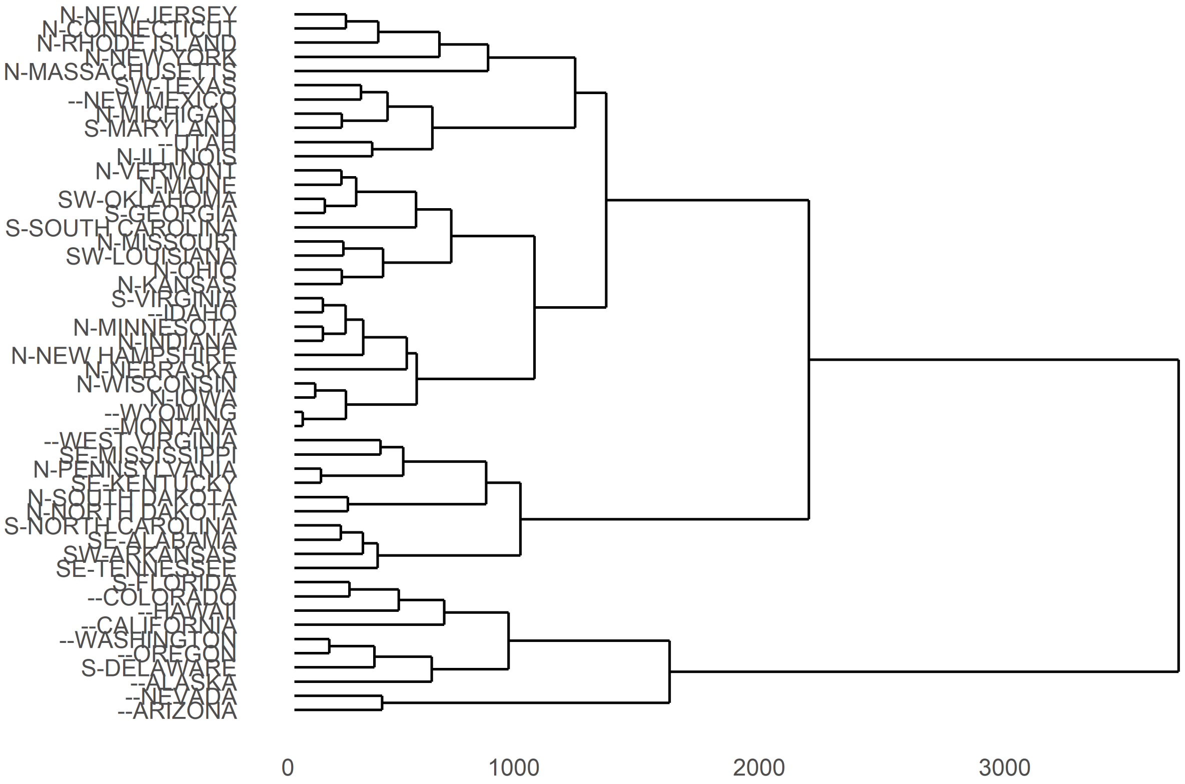
April 20, 2020

Question 1

The data set in the attached SAS program (crime.sas) contains variables describing the types of crime rates in each U.S. state. A researcher is interested in summarizing these crime rates and determining whether the Southern States have different patterns of crimes than the Northern States. Use a cluster analysis to determine which states have similar crime rates.

Part A

Perform hierarchical clustering on the data, and provide the resulting dendrogram.



Part B

Report the distance (similarity) measure and linkage function used.

The distance method used was **euclidean**. The linkage function was **complete** for the clustering.

Part C

How many clusters are there in the data? Provide justification using the criteria discussed in class.

There are in the original data labels for many states: *N*, *S*, *SE*, *SW*. With this clustering approach, we have to select how many iterations we would like for the solution. In this case, we should select somewhere between 3-6 clusters to approach the pre-existing or known clusters of geographic regions.

Three clusters:

10, 10, 30

Four clusters:

10, 8, 2, 30

Five clusters:

10, 8, 2, 11, 19

If we use 5 clusters, then there are two states that have their own cluster. That state is 3, 3. These are geographically and culturally very similar, thus, would not divid this further.

Part D

Provide an interpretation of your final clustering solution, in the context of the original research question.

When reviewing all the classification groups, we can see that although there are some patterns, where certain states tend to cluster together, it is not a clear division. We cannot easily argue that there is a robust difference between northern and southern states.

Question 2

Read the article titled "Identifying Heterogeneity Among Injection Drug Users: A Cluster Analysis Approach."

Part A

Summarize the findings of this article.

This article used cluster analysis to characterize the population of injection drug users with the intent to identify novel behavioral patterns. They used information on syringe sharing, ethnicity, and drug types. They found seven clusters. Some of these cluster relationships were known from prior knowledge (e.g. certain drugs that go together, like petnazocine and methylphenidate). They confirmed that injection use and HCV were related. The novel relationships they found included female crystal methamphetamine users who had high-risk behaviors but low prevalence of blood-stream infections. This novel-group finding was a major take away from the article.

Part B

Provide a brief discussion, including strengths and weaknesses of the cluster analysis approach.

The most important part of any analyses is the data input stage. This data set uses community volunteers, which serves as a starting point for recruitment bias. They also used an age cut off of >15 years, which adds a bias to the data. Also, they had self-initiated recruitment, which likely biases away from high risk groups that are unable to participate.

The cluster analysis used Ward's linkage in a agglomerative hierarchical approach. A strenght includes using Ward's linkage that is better for binary data and creates equal group sizes. Another strength is using several clustering approaches, and then deciding to report only one due to similarity between approaches. They used evidence based stopping rule's (Duda's pseudo T and Calinski pseudo F). They were also thoughtful in variable selection, using evidence for reasons.

Additional weaknesses include the lack of a dengrogram that shows the clustering analysis. They also do not show how many iterations it took to achieve the clustering. They also separated men and women, which could be a strength and a weakness in that there may be patterns that occur regardless of sex, but also patterns that would only occur within genders (which they document in table 2). An additional weakness is the alpha chosen, as they are performing multiple hypothesis testing with their regression models.