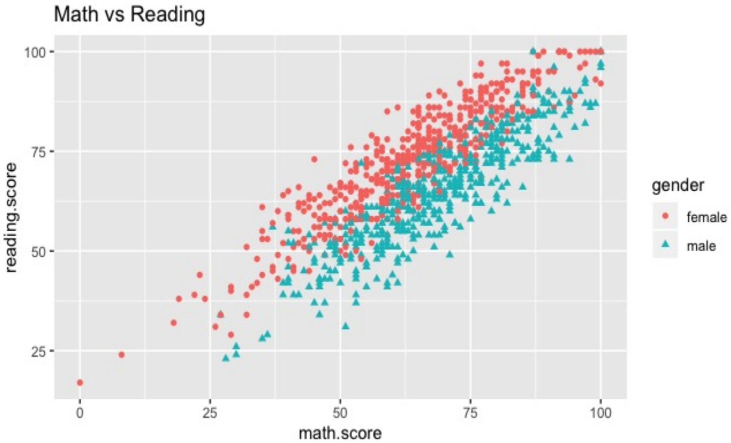
Impact of linkage on hierarchical clustering and covariance type on GMM clustering

by Calvin Smith

April 2020

Dataset

- Dataset contains math and reading scores for students and their gender.
- 1000 observations (518 female, 482 male)



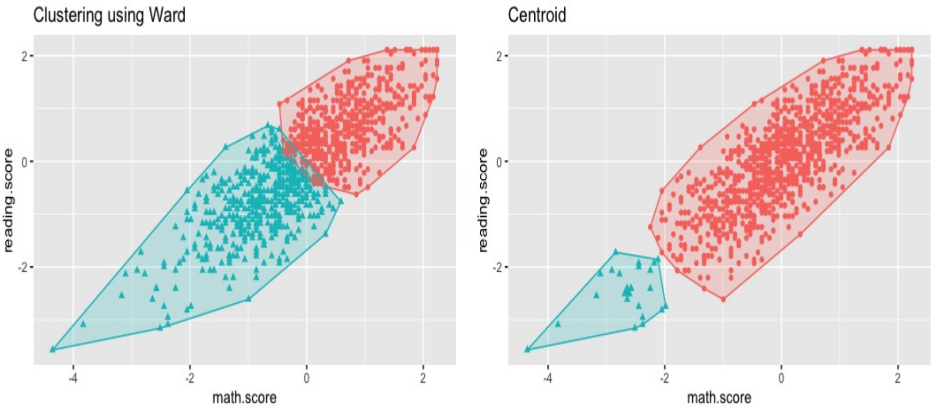
Dataset

- Looks like females perform better in reading and males perform better at math.
- However, there is quite a bit of overlap, and not a huge difference between groups.
- There is no clear boundary separating the groups.

Linkage

- Comparing two different linkage methods: Ward and Centroid.
- The Ward method tries to minimize the total within-cluster variance. At each step, it finds the pair of clusters that leads to a minimum increase in total within cluster variance after merging. The initial distance between clusters (points) is the squared euclidian distance.
- The Centroid method uses the distance between the centroids of two clusters. At each step, the Centroid of a cluster will be moved according to which clusters have been merged.
- Since we know that there are two classes (male and female) in our dataset, this is how we will evaluate the different linkage methods. How well can they capture the structure in our data when forcing them to create two clusters?

Linkage


Clustering using Ward
Centroid
reading.score
math.score
2
0
-2
-4
-2
0
2

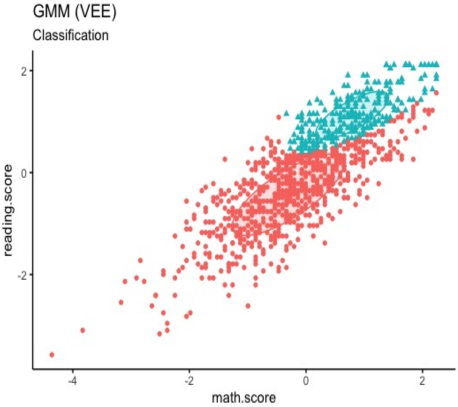
Linkage

- Neither methods are very good at classifying according to the original data.
- Both of the methods uses only the geometry of the feature space to determine clusters. That is, we do not make any assumptions about the distribution of the data.
- Since the original clusters (male and female) are not clearly separated, using only the geometry of the feature space for clustering is difficult.
- Using only the geometry of the feature space, capturing the shape of this dataset proves to be difficult. Both of the methods create clusters where the cut is diagonally "straight through" the dataset, but what we really want is to cut diagonally down the middle of the dataset (think straight line $y = x$). Based on the geometry of the feature space, the methods are unable to do so, because it would not minimize the within cluster variance (ward method) and the centroid method would need a larger separation between the clusters to be effective, and perhaps need the clusters to be more "circular".

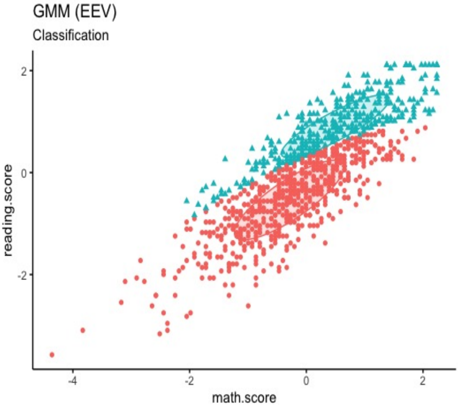
GMM

- Comparing two different Covariance structures.
- Based on the BIC for two clusters, the "Optimal" strucure is VEE (Variable volume, Equal shape and Equal orientation).
- Comparing with the second "best" structure EEV (Equal volume, Equal shape, Variable orientation).
- Based on the original clusters, it seems reasonable to assume at least equal shape and equal orientation.

GMM


GMM (VEE)

Classification


GMM (EEV)

Classification

GMM

- Surprisingly, the EEV Covariance structure seems to capture the shape of the data better than the VEE. This is surprising since it looks like the two clusters from the original data have the same orientation.
- However, I would say that both structures are better compared to linkage. At least they are "closer" to the true boundary compared to the linkage methods.

Summary

- Since the two clusters are overlapped and not clearly separated, the methods presented are unable to accurately cluster the data in two groups.
- With this particular shape of data, using only the geometry of the feature space seems like a very uneffective way of clustering.
- GMM is closer to the "truth". But less overlap in the data would be needed for the method to be effective.
- The main conclusion is that if we have groups that are close to each other, using hierarchical and model-based clustering to identify the groups is difficult.