QBS 177 - Assignment 3 - Bertsch

1. Run PCA using spectral decomposition on the Lung cancer data PCA.example1.txt

Download PCA.example1.txt
(you can find the program for PCA already run on Melanoma data, just adapt the program to be run on the lung data). The melanoma data and lung data are under principle component folder. The Lung data you are to analyze is named PCA.example1.txt. Use date() to report the run time. (2pt)

Run time was 4.649158 minutes.

2. Run PCA using singular value decomposition on the Lung cancer data and use date() to report the run time. Compare it with spectral decomposition.(2pt)

Run time was 3.989584 minutes. A little faster! I also discuss this in the R-Markdown notebook.

3. There are three populations (European -CEU, African-YRI and Asian-CHB). The last 505 samples can be used as three types of index samples CEU(19662:19826)-European;CHB(19827:19963)-Asian;YRI(19964:20166)-African Calculate the centroids for these populations using the first 2 principle components.

This question is completed and explained in greater detail in the attached PDF of my R-Markdown notebook.

Europe centroid: (-1.736351, 0.390416)

Asia centroid: (17.177547, 6.374995)

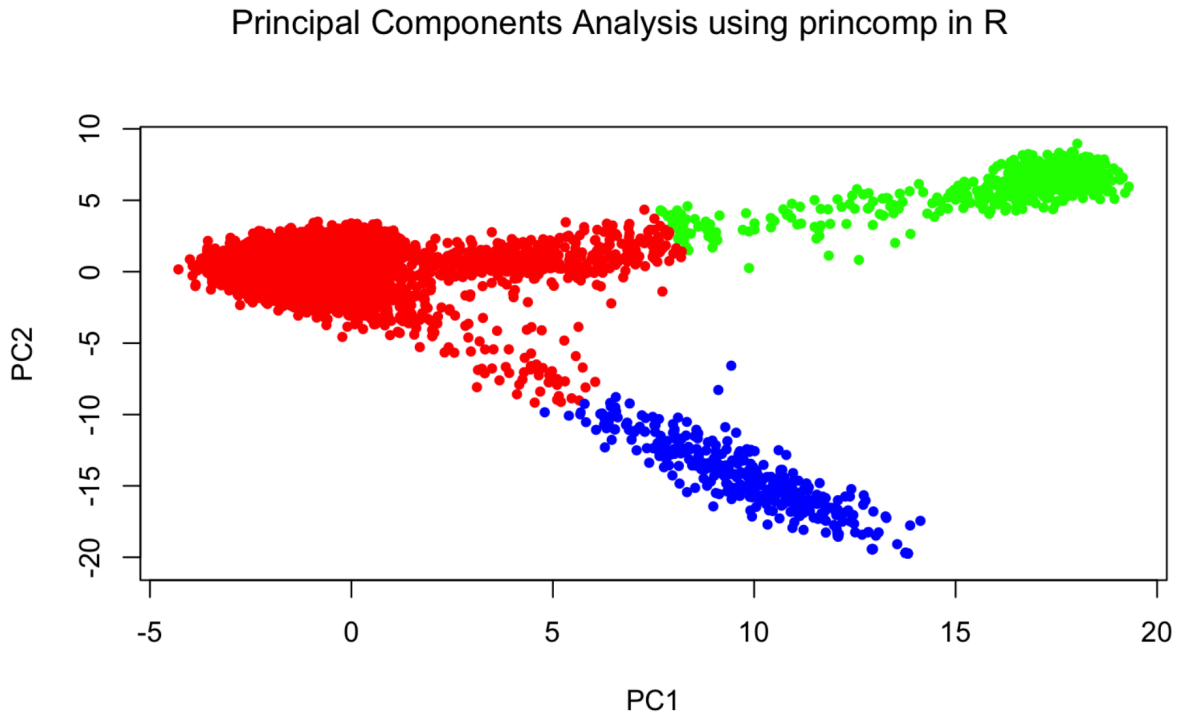
Africa centroid: (12.80415, -18.83316)

Then cluster the first 19661 samples into the closest centroid.

You can see the results of this in the attached R-Markdown notebook. I did this by creating lists and iteratively measuring the Euclidean distance between the point created by the first two principal components and the centroid of each of the three clusters. I also counted the number of samples that were closest to each centroid and found that the majority were closest to the European centroid.

4. plot all samples points based on the first two principle components and use different color to denote their population group. (3 pt)

QBS 177 - Assignment 3 - Bertsch



The above plot and the code that generated it is discussed in detail in the attached R-Markdown report.