Alexandria University
Faculty of Engineering

Bioinformatics
Spring 2024
Lab 5
Assigned: 21/3/2024
Due: 26/3/2024

Lab 5: PCA and Clustering (80 Points)

Objective

Students should be able to perform Principal Component Analysis (PCA) and clustering on genetic data using PLINK and R.

Requirements

- You are obligated to attempt all tasks
- You **MUST** use R Markdown to submit your report
- You **MUST** submit the screenshots of PLINK logs
- State your name and ID at the first cell in your markdown report
- For all the tasks, you should use the Dataset of 156 Qataris.

Part 1: Principal Component Analysis Using PLINK (35 Points)

Task 1.1: QC and PCA (15 Points)

- Run Minor Allele Frequency count on your dataset using PLINK and the flag --freq. Provide a screenshot of the head of the file. Explain the output
- Run QC on your dataset using PLINK
    - Try the following flags separately --maf --geno and --hwe filters. Try and report different levels and thresholds focusing on the number of variants removed. Add screenshots of the log files output at the end of each trial (3 max with meaningful values for each flag).
    - Run the final version of your QC using all the flags combined and report the final number of variants. Use the following thresholds (hwe: 0.01, maf: 0.1, geno: 0.001)
- Run PCA on your dataset using the PLINK and --pca flag
- You should recode the data to be in the format ped/map

- Use meaningful names for your output files

Task 1.2: PCA Visualization (20 Points)

- Explore Egienvectors and Eigenvalues. For Ubuntu/MacOS/Windows WSL users, use the awk, vi, and head commands to view the eigenvalues and eigenvectors.For Windows users, you can use PowerShell commands like Select-String, ForEach-Object, regular expressions, Get-Content and Select-Object.
- Load the PCA results into R.
- Create 2D scatter plots comparing PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3 using ggplot2.
- Create a scree plot for the first 20 components with the explained variance.
- Create a 3D plot of the first three principal components.
- Install and use the scatterplot3d package for 3D plots.

Part 2: Clustering in R (40 Points)

Task 2.1: Perform Clustering (10 Points)

- Reduce the dimensionality of the dataset by only choosing the first three pricipal components PC1, PC2, PC3
- We will perform clustering techniques to find the clusters that correspond to different subpopulations in this population
    - Use k-means clustering with kmeans function
    - Try different number of clusters
    - Determine the optimality of the number of clusters using Dunn's index or Xie Beni's index

HINT: for Dunn's index use the dunn function. For Xie Beni's index use the fclustIndex function

Task 2.2: Perform Hierarchical Clustering (10 Points)

- Try hierarchical clustering methods, once with single linkage and another with average linkage clustering.
    - Use the hclust function. Note that you need to specify the method as either "single" for single linkage or "average" for average linkage.
    - Try different number of clusters
    - Determine the optimality of the number of clusters using Dunn's index or Xie Beni's index for the average linkage clustering
    - Determine the optimality of the number of clusters using common sense (expert eye) for single linkage clustering.
    - Plot the dendrograms using the plot function

Task 2.3: Visualize Clusters (20 Points)

- Visualize the clusters correposponding to the subpopulations that were produced from each clustering on the pca plots (PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3) using ggplot2 and do not forget to color them.
- Create a side-by-side comparison of the clusters formed by k-means, single linkage, and average linkage methods. Use gridExtra's grid.arrange function to plot multiple clustering results side by side for easy comparison. Make sure each plot uses different colors for each cluster to aid in comparison.

Reporting (5 points)

- You are required to write a report discussing the commands, and file formats and include any screenshots of the output of such commands
- Write your comments on how the data quality control affected the dataset

References

- Dataset of 156 Qataris
- PLINK file formats
- ggplot2 Documentation
- grid.arrange
- scatterplot3d
- kmeans
- Hierarchical Clustering
- Dunn's Index
- Xie Beni's Index
- Markdown Cheatsheet