MvLA Report 2021

ANALYSIS OF PERIPHERAL BLOOD MONONUCLEAR CELLS

Ciara Lynch | STU33011 | 24/11/21

Introduction

The Healthy Flow data set (file hfd.csv) contains 5 variables, 4 of which represent the cell antibodies against the CD3, CD4, CD8 and CD19 protein markers. The other variable, variable number 5 is the common analysis of "gating" in flow cytometry. It aims to identify the populations of the above-mentioned cells within a data set. The first four variables mentioned are classed as continuous data, the gating variable is considered ordinal data.

The following report will be broken into two sections for further analysis;

1. Cluster Analysis:
   - To identify similar groups of subsets, present in the data set
   - To identify the subsets of data similar to the identified gated populations
2. Classification:
   - To predict which data will be assigned to the identified gates

CLUSTER ANALYSIS

Identifying the subsets present in the data

Section aim:

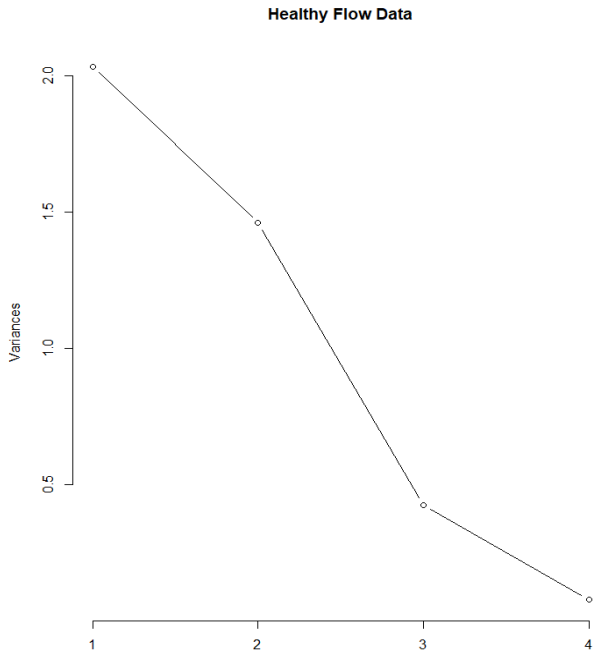
- As previously mentioned, this section is to identify subsets of cell data in the data set with similarities and determine if these newly identified subsets are similar to the identified gated populations

Principal Component Analysis (PCA)

Principal Component Analysis is a dimensionality-reduction method that's aim is to reduce the number of variables in a large data set (the dimensionality) to a smaller number while trying to preserve as much information as possible. By reducing the amount of variables in a data set we achieve simplicity, especially for visualization, at the cost of a little accuracy. This therefore makes it easier to identify the clusters and groups present in the data visualised. By applying this analysis method, it is important to note that the first component is the most important followed by component two and so on; As a result of the variance the components account for, is listed in decreasing order from Principal Component 1 (PC1).

For the data set I did conduct a standardised Principal Component Analysis. There are many reasons for conducting a standardisation on this data for this test the main being that the variables with the highest variance in the set do not dominate or skew the PCA. If for this analysis the data was not standardised the PCA would be dominated by the CD19 marker as it has the highest variance of 5.81783.

Standardised PCA of Healthy Flow Data Set



As seen above, the proportion of variance noted by the first principal component is quite large at just above 2.0, this is a good indication that the dataset is well suited to principal

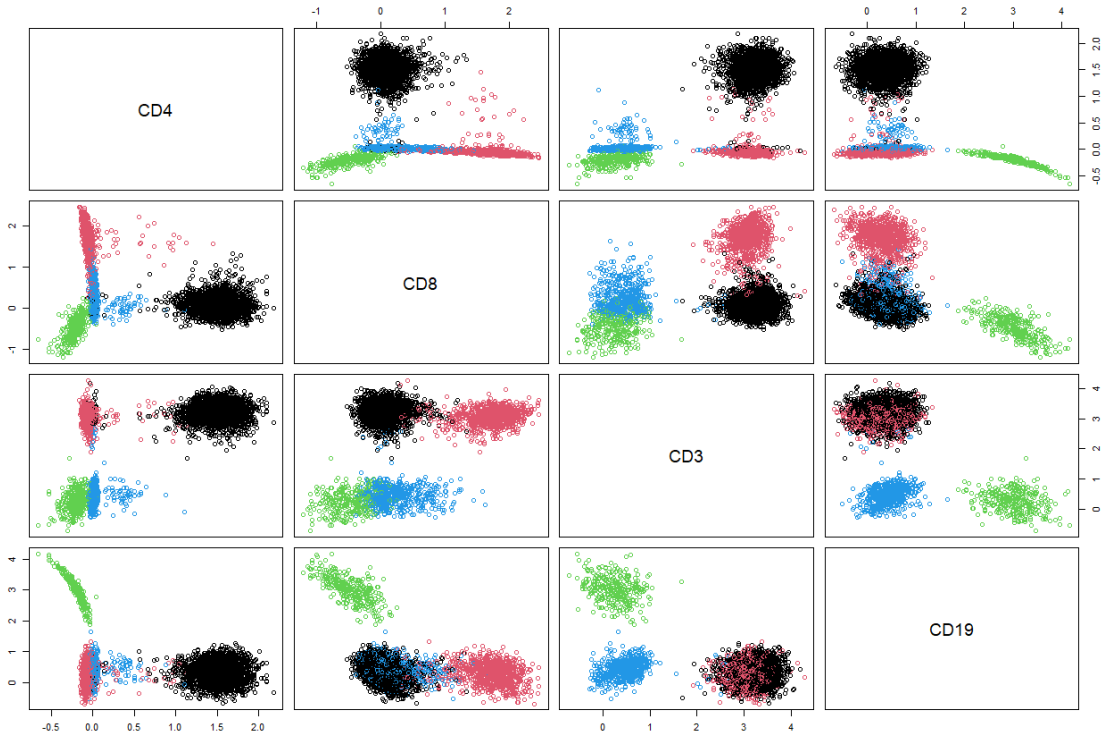
component analysis. Due to PC1 accounting for a large variation of the original data and more than any other principal component.

As expected, the most influential variables from the standardised PCA is within the first Principal Component which were CD8 and CD19. This can be seen in the data capture below where CD8 and CD19 in PC1 have the largest magnitude.

	PC1	PC2	PC3	PC4
[CD4]	-1.2039510	0.7013812	-0.01616878	-0.1346060
[CD8]	2.7849151	0.9054027	1.42907342	-0.3174678
[CD3]	2.2435587	-0.1322606	-1.33268766	-0.2437735
[CD19]	-0.8399212	1.1032951	0.44266840	-0.1463205

Visual Comparison of the Data

Below I have created a graph to compare the continuous variables of Healthy Flow dataset.



Plotted and pictured above, Healthy Flow Continuous Variables.

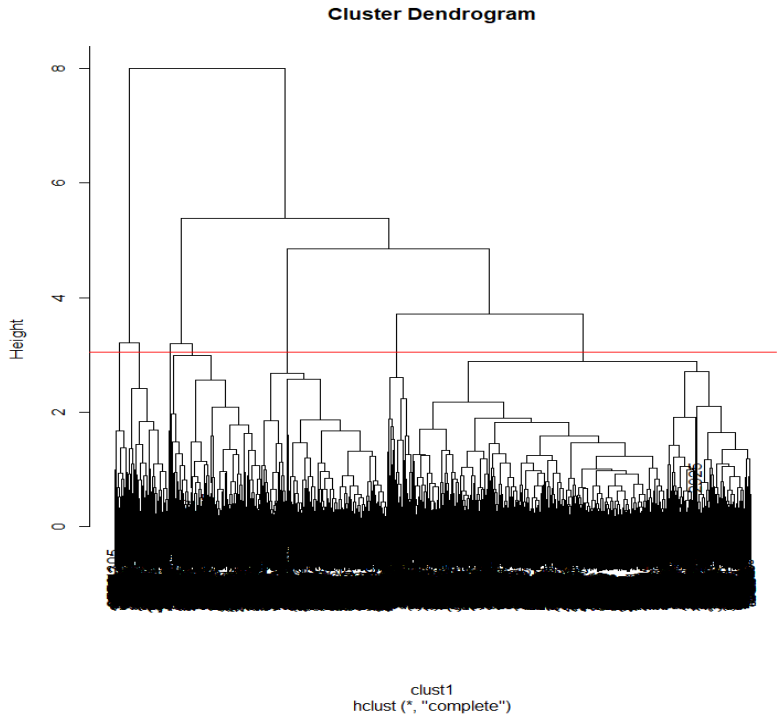
As seen above the Principal Component plot showcases the similarities between the protein markers. The clusters of the cells do not dictate much other than the similarities although where the high overlap is, illustrates the presence of lymphocyte cells due to the nature of lymphocyte cells and the protein markers of the antibodies fighting the infection.

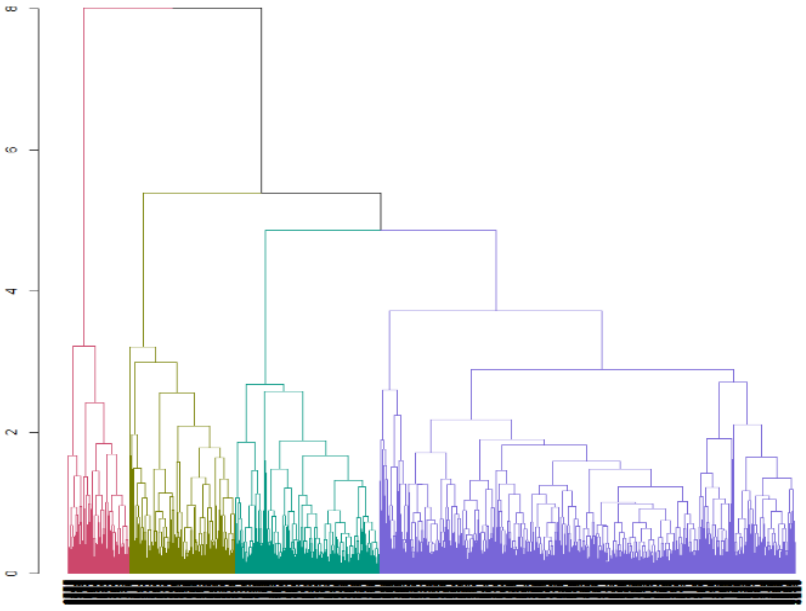
Hierarchical Clustering

The analysis technique Hierarchical Clustering (HC) is the application of clustering/grouping similar observations in the data structure into groups called clusters. This results in group structure becoming evident in the data, the clusters are also predominant ordering from top to bottom.

As I previously mentioned within the Healthy Flow data set there is continuous variables, for this reason I chose to apply the Euclidean or Gower's general dissimilarity measure for the dissimilarity matrix when creating the visualisation plot. For the linkage method in the clustering, I chose the complete linkage as for this data set it will group the last clusters at a large measure of dissimilarity. This will then establish an acceptable level of internal similarity within the clusters too.

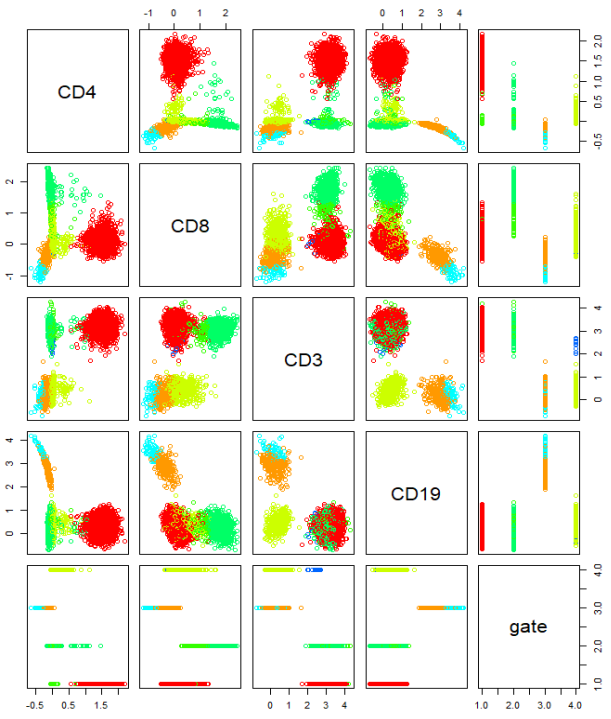
In a dendrogram, like the one attached below, observations that are connected at the bottom share the most similarities and then as you continue up the graph the clusters that are most dissimilar are those being joined. This creates a good visual display of the hierarchical relationship between observations within a data set

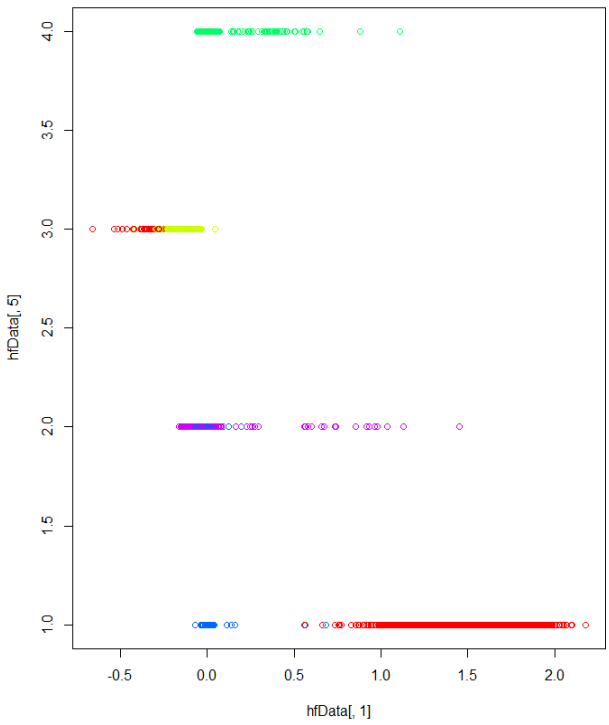




The two dendrograms above are of the same analysis to the data set, I have included the second plot for clearer viewing with the coloured visual, it illustrates the group structure present in the data. On the y-axis we can see the height of the dendrogram, this depicts the dissimilarity between the groups from clustering.

The cut-off height which determines how many clusters are present in the above data set is 3.0474. It is visualised by the orange line in the first dendrogram, it is derived using the formula [h +3sh], where sh is standard deviation and h is the mean height at which the clusters are joined.





Above are two plots which have been used to visually interpret the data. Figure 1 is on the left and figure 2 on the right. In figure 2 I have taken the comparison of CD4 and the Gate variable. The clusters in the figures are signified by the different colours. From this analysis not much can be deduced visually, other than correlation of which distinct populations of cells are similar to the identified gated populations, 1, 2, 3 or 4.

Note, in figure 2, the x-axis is depicting CD4 and the y-axis the Gate variable.

Rand Index

	GateVariable			
Cells	1	2	3	4
1	2112	0	0	95
2	0	327	0	0
3	0	1	562	0
4	0	0	0	767

In the above table it is illustrated what clusters fall into each of the gates. Not much more useful insights can be gained from this table.

The rand index can be found below which is a more precise computation of how accurately the clusters reflect the true structured groups of the data.

RI	ARI
0.9632343	0.9224079

The Rand index can be interpreted as follows; It can have a value between 0 and 1, where 0 is indicative that the two data clusterings do not agree on any pair of points in the structure and 1 is indicative that the data clusterings are the same.

The Rand Index, 0.9632343, represents the hierarchical data clustering results with usage of Glover's dissimilarity measure and complete linkage. This is representative of the unadjusted Rand Index as it was not calculated with respect to agreement by chance.

The adjusted Rand Index, 0.9224079, was calculated with respect to agreement by chance. It depicts a very high score which is indicative that there is quite significant agreement between the clusters and the real data in terms of connection between the protein markers and gate variable.

CLASSIFICATION

The aim of classification in this report is to try and use the known data to determine what cells will be assigned to what gates within the Healthy Flow data set.

The data in this section will not be standardise for any of the analysis methods. This is due to wanting to ensure that the output is understandable and easily interpreted.

Predicting the assignment of Cells to Gates.

Section aim:

- This section aims to determine accurately the prediction of what cell populations will be assigned to the identified gates.

K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a non-parametric model that classifies data points based on the points most similar to it as a method of assigning group membership. A benefit to this method of analysis is that it has a quick calculation time and does not make any assumptions on the data set it is applied to. However, a consequence of this, is that it is therefore not possible to classify data points in a boundary where there it can be assigned one way or another.

KNN compares and observation to the k nearest datapoint in the dataset and then assigns that observation to the group which contains the most k datapoints that it is located nearby.

An important factor in using the KNN analysis is choosing an accurate and appropriate value for k. Choosing small values for K can have a high influence on the results and create a lot of noise, larger values for K will result in smoother decision boundaries meaning lower variance but increased bias. In this report I will follow the cross-validation method to choose K. I will split the known data into a training, test and validation set. The training set classifies the “unlabelled” points in the test set and then the test set is used to find the value of k that fits best for the classification. The validation set is then used to estimate the classification error of the best-chosen k identified within the test set.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are dimensionality reduction analysis methods, used if it is known or assumed that there is group structure internally in the data set. Within this data set there is 4 gates all correlating to a different result of the gating analysis performed on the cells.

This analysis reveals structure within the data set that then allows the classification of future observations i.e. the prediction of cells to identified gates. The method behind it is that LDA & QDA uses characteristics of the already labelled data to then classify the group membership of unlabelled data.

What makes LDA and QDA different to KNN is that they assume use of a distribution over the data which then results in the ability to quantify the given uncertainty over the structure of the data set. This allows us to consider the probability whether an observation (cell) is likely to be assigned to a gate or not.

CONCLUSION

In conclusion, the Principal Component Analysis that was performed was not the most informative method in this report. When applied to this dataset it did not result in many principal components which was what would be required to indicate an acceptable level of variance. However, the graph plots were a good visual depiction of the variance between the principal components found within the dataset.

The Hierarchical Clustering analysis that was performed resulted in showing what protein markers of CD4, CD8, CD3 and CD19 were present in the identified gates of 1 through to 4. The graphs clustered the cells accordingly illustrating the internal structure of the data set. The Rand Index above is indicative of the strong agreement between the clusters within the internal structure of the dataset further alluding to the connection between the protein markers and their specified gates.

Unfortunately, I was unable to gain any applicable output in regards any classification analysis methods from R studio. Therefore above I went onto explain the use of K-Nearest Neighbour and Linear Discriminant Analysis instead and how I intended to apply these methods to the Healthy Flow Data Set.