Maria Emma Arcia, Corinne Scales, Olin Yoder

Assignment - 9

BSAN 450 (Spring 2023)

This assignment is due on April 27, 2022 at 9 PM Central. The total points possible are 100 and there is one (1) problem with several sub-parts. Each sub-part carries equal points. You can form groups to attempt these problems. Each group will submit one copy of the assignment on Canvas, either in word or pdf, and the assignment should clearly include the names of the group members.

1. In this example we will analyze the NCI60 cancer cell line microarray data, which consists of 6,830 gene expression measurements on 64 cancer cell lines. To read this data in R, you will need the following code:

```
library(ISLR) nci.labs=NCI60$labs
nci.data=NCI60$data
```

The data nci.data has 64 rows and 6,830 columns. Each row is a cell line for which we have gene expressions recorded for 6,830 genes. The nci.labs data has the cancer type for each of these 64 cell lines. Your analysis will rely primarily on nci.data and not on nci.labs as the latter are just labels.

a. Perform PCA on nci.data with scale=TRUE in the prcomp function.

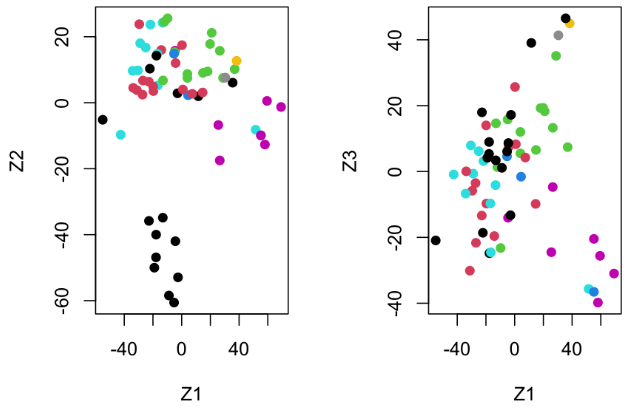
pr.out1 =prcomp(nci.data, scale=TRUE)

b. Plot the first two principal components. You may require the following code wherein pr.out is the name of the output from the prcomp function that you used in part (a).

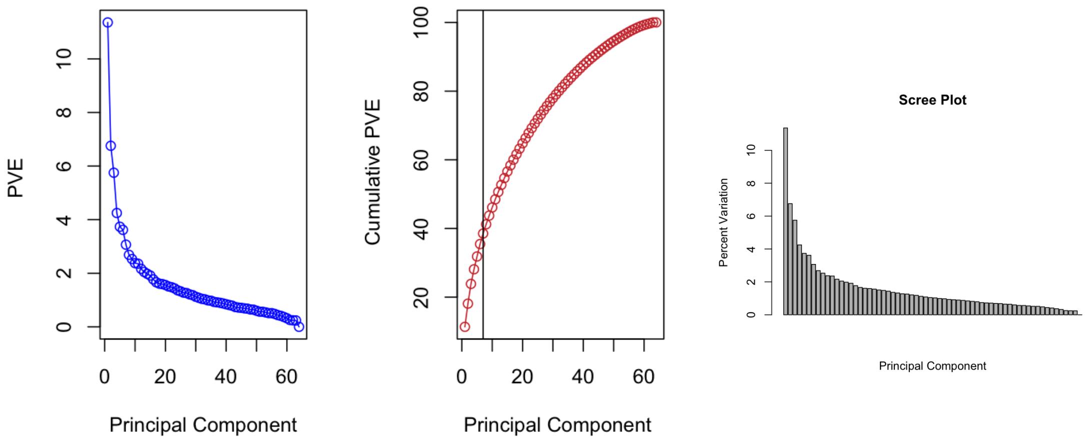
```
par(mfrow =c(1,2))


plot(pr.out$x [,1:2], col=as.factor(nci.labs), pch =19, xlab ="Z1",ylab="Z2")
plot(pr.out$x[,c(1,3)], col=as.factor(nci.labs), pch =19,
xlab="Z1",ylab="Z3")
```

You will notice from your plot that observations belonging to a single cancer type tend to lie near each other in this low-dimensional space.



c. As demonstrated in the lecture, plot the PVE of each principal component (i.e. a scree plot) and the cumulative PVE of each principal component. What is the % of variance explained by the first seven principal components? Based on the scree plot, how many principal components should we choose?



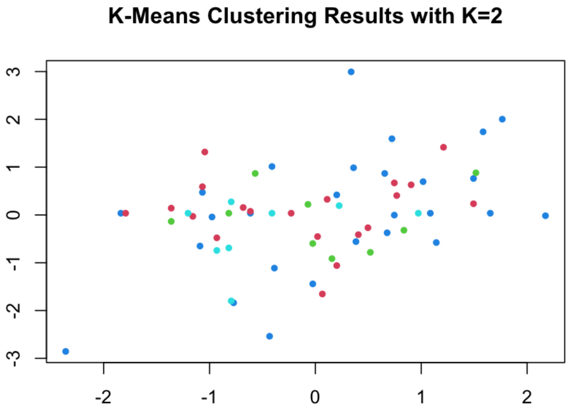
About 40% of the variance is explained by the first seven principal components. Based on the scree plot, we would recommend using the first four principal components.

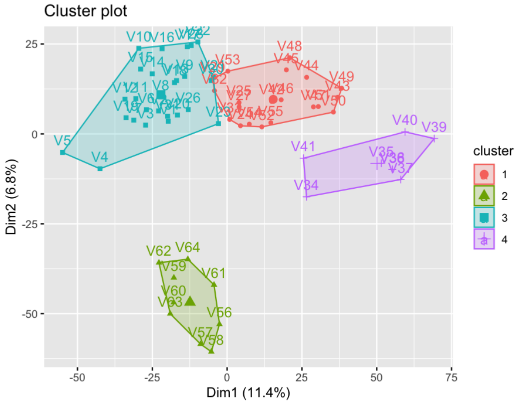
d. Scale the nci.data such that the features have zero mean and 1 standard deviation. You will have to use the scale() function in R for this purpose. Then, on the scaled data, perform K-means clustering with K = 4. What are the cluster sizes that you obtain? Calculate the Between Cluster sum of squares as a % of total sum of squares.

Maria Emma Arcia, Corinne Scales, Olin Yoder

Cluster sizes: 20, 9, 27, 8

Between cluster ss: 19.92218%


K-Means Clustering Results with K=2


Cluster plot
Dim2 (6.8%)
Dim1 (11.4%)
cluster
V10
V16
V15
V14
V9
V8
V11
V6
V26
V3
V23
V5
V4
V53
V48
V45
V44
V49
V43
V25
V27
V46
V50
V55
V52
V40
V39
V41
V35
V36
V37
V34
V62
V64
V59
V61
V60
V63
V56
V57
V58