HW3

Zirui Zhang

2023-03-27

```
library(dplyr)
library(formattable)
```

QUESTION 1

```
mydata = read.table(file="./PCA_data.csv", header=TRUE, row.names=1, sep=",")
mydata.pca = prcomp(mydata, retx=TRUE, center=TRUE, scale=TRUE)
# variable means set to zero, and variances set to one "scale=TRUE"
# PCA scores for each sample store in mydata.pca$x
# loadings stored in mydata.pca$rotation
# square roots of eigenvalues store in mydata.pca$sdev (note that eigenvalues are variances of principa
# variable means stored in mydata.pca$center
# variable standard deviations stored in mydata.pca$scale
sd = mydata.pca$sdev
loadings = mydata.pca$rotation
rownames(loadings) = colnames(mydata)
scores = mydata.pca$x
```

a) Calculate PCA scores using loadings and original math/chem/bio scores, and compare to output PCA scores from the R package prcomp.

```
## PC1 PC2 PC3
## 1 -1.6579369 0.571682476 0.2765913
## 2 -2.5759191 0.171154076 -0.3946536
## 3 -1.8500938 0.561523155 -0.3289593
## 4 -2.0029634 -0.279232617 0.2090958
## 5 -1.7019579 -0.553776982 0.5314955
## 6 -1.0237890 0.746472175 0.4269871
## 7 -0.9343334 0.599025432 0.2483663
```

```
## 8 -2.0297283 -0.031871212 -0.1337258
## 9 -2.2666815 0.246059593 -0.2542754
## 10 -2.6579342 -0.005870397 -0.1643718
```

```
# compare with the results from package
head(abs(scores_2-scores)/scores, 10)
```

```
##
               PC1
                             PC2
                                           PC3
## 1
                    3.884055e-16 8.027895e-16
     -4.017848e-16
     -3.448006e-16 8.108359e-16 -4.219738e-16
## 3 -3.600541e-16 1.977163e-16 -3.374956e-16
## 4 -4.434322e-16 -5.963969e-16 5.309639e-16
## 5
     -3.913927e-16 -2.004820e-16 2.088866e-16
## 6 -6.506553e-16 1.487293e-16 3.900199e-16
## 7 -7.129509e-16 1.853382e-16 5.587628e-16
## 8 -4.375849e-16 -1.524017e-15 -8.302233e-16
## 9 -3.918409e-16 4.512009e-16 -4.366223e-16
## 10 -3.341612e-16 -2.762959e-14 -1.350868e-15
```

The difference between the two PCA scores are rather small.

b) Calculate percent variance explained of each component.

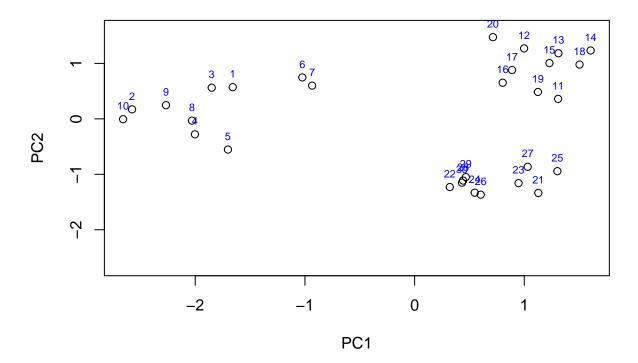
```
lamda = percent(mydata.pca$sdev^2/sum(mydata.pca$sdev^2))
lamda
```

```
## [1] 66.93% 29.71% 3.36%
```

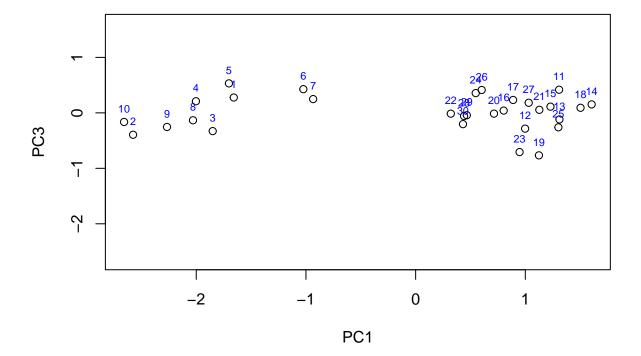
The variance explained by PC1 is 66.93%, by PC2 is 29.71%, by PC3 is 3.36%.

c)

(1) PC1 vs PC2



(2) PC1 vs PC3



(3) PC2 vs. PC3.

