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.

```
# calculate PCA scores
mydata_2 =
  mydata%>%
  mutate(math = scale(math) - mean(scale(math)),
         bio = scale(bio) - mean(scale(bio)),
         chem = scale(chem) - mean(scale(chem)))
scores_2 = as.matrix(mydata_2)%*%loadings
scores_2
```

```
##           PC1          PC2         PC3
## 1  -1.6579369  0.571682476  0.27659132
## 2  -2.5759191  0.171154076 -0.39465359
## 3  -1.8500938  0.561523155 -0.32895929
## 4  -2.0029634 -0.279232617  0.20909576
## 5  -1.7019579 -0.553776982  0.53149553
## 6  -1.0237890  0.746472175  0.42698709
## 7  -0.9343334  0.599025432  0.24836634
```

```
## 8  -2.0297283 -0.031871212 -0.13372583
## 9  -2.2666815  0.246059593 -0.25427541
## 10 -2.6579342 -0.005870397 -0.16437182
## 11  1.3079701  0.359969524  0.41720262
## 12  0.9980941  1.269880082 -0.28446005
## 13  1.3110519  1.182549991 -0.11825137
## 14  1.6046169  1.232476842  0.15248740
## 15  1.2290368  1.005525517  0.11203039
## 16  0.8027869  0.651507049  0.04113399
## 17  0.8884538  0.880662669  0.23237339
## 18  1.5032773  0.978342557  0.09160858
## 19  1.1241139  0.484893331 -0.76606024
## 20  0.7141229  1.472874891 -0.01392787
## 21  1.1276809 -1.336079451  0.05473737
## 22  0.3206321 -1.229551905 -0.01435777
## 23  0.9474764 -1.159024498 -0.70700433
## 24  0.5478545 -1.331670861  0.35630217
## 25  1.3013734 -0.943475598 -0.26024587
## 26  0.6022446 -1.366839327  0.41257208
## 27  1.0307847 -0.864161491  0.18169711
## 28  0.4413646 -1.112674562 -0.05800904
## 29  0.4666995 -1.049140991 -0.04278934
## 30  0.4317024 -1.151229468 -0.20358933
```

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

```
##              PC1           PC2           PC3
## 1  -4.017848e-16  3.884055e-16  8.027895e-16
## 2  -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
## 11  6.790510e-16  1.542107e-16  2.661112e-16
## 12  5.561715e-16  0.000000e+00 -1.951457e-16
## 13  5.080911e-16  0.000000e+00 -7.041502e-16
## 14  5.535143e-16  0.000000e+00  1.820188e-16
## 15  5.419966e-16  0.000000e+00  2.477504e-16
## 16  8.297767e-16  1.704084e-16  3.542491e-15
## 17  7.497675e-16  2.521335e-16  5.972193e-16
## 18  4.431211e-16  0.000000e+00  9.089403e-16
## 19  5.925857e-16  1.144812e-16 -2.898527e-16
## 20  9.328000e-16  1.507559e-16 -1.046224e-14
## 21  7.876151e-16  0.000000e+00  3.929778e-15
## 22  2.077564e-15  0.000000e+00 -6.041060e-15
## 23  7.030611e-16  0.000000e+00 -1.570320e-16
## 24  1.418545e-15  0.000000e+00  4.673939e-16
## 25  6.824931e-16  0.000000e+00 -6.399082e-16
## 26  1.290433e-15  0.000000e+00  4.036469e-16
```

```
## 27  8.616527e-16 -1.284740e-16  1.222059e-15
## 28  1.635031e-15  0.000000e+00 -2.392349e-15
## 29  1.546273e-15  0.000000e+00 -3.243282e-15
## 30  1.671626e-15  0.000000e+00 -1.363312e-16
```

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
```

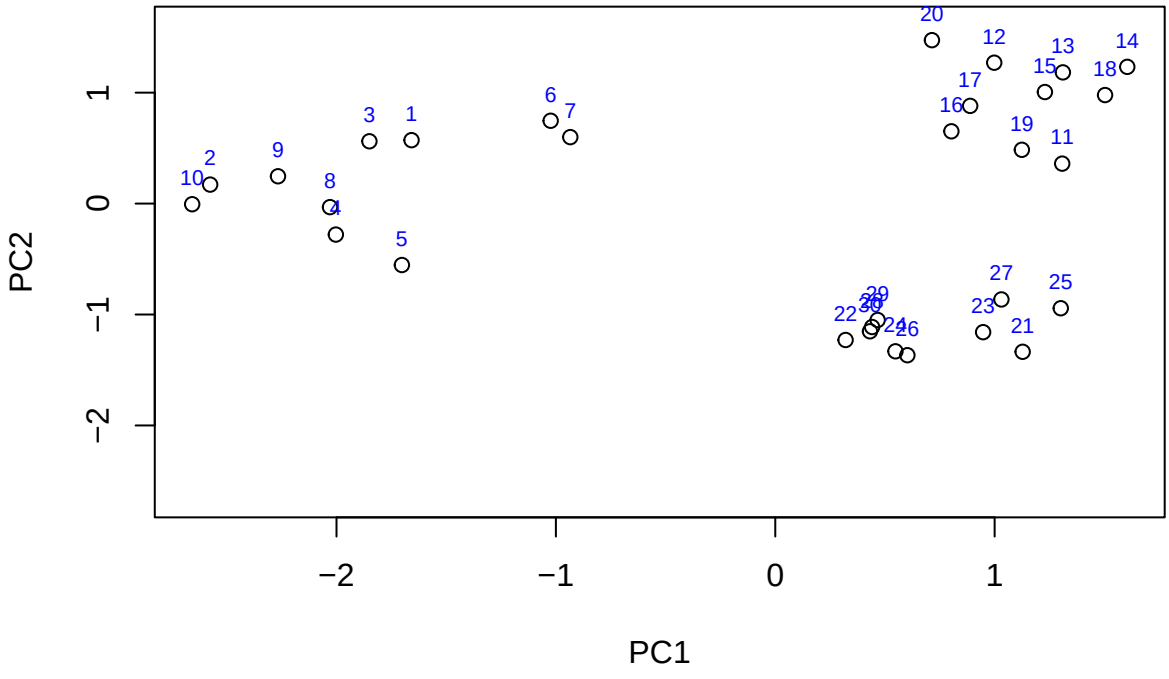
```
## [1] 66.93% 29.71% 3.36%
```

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

c)

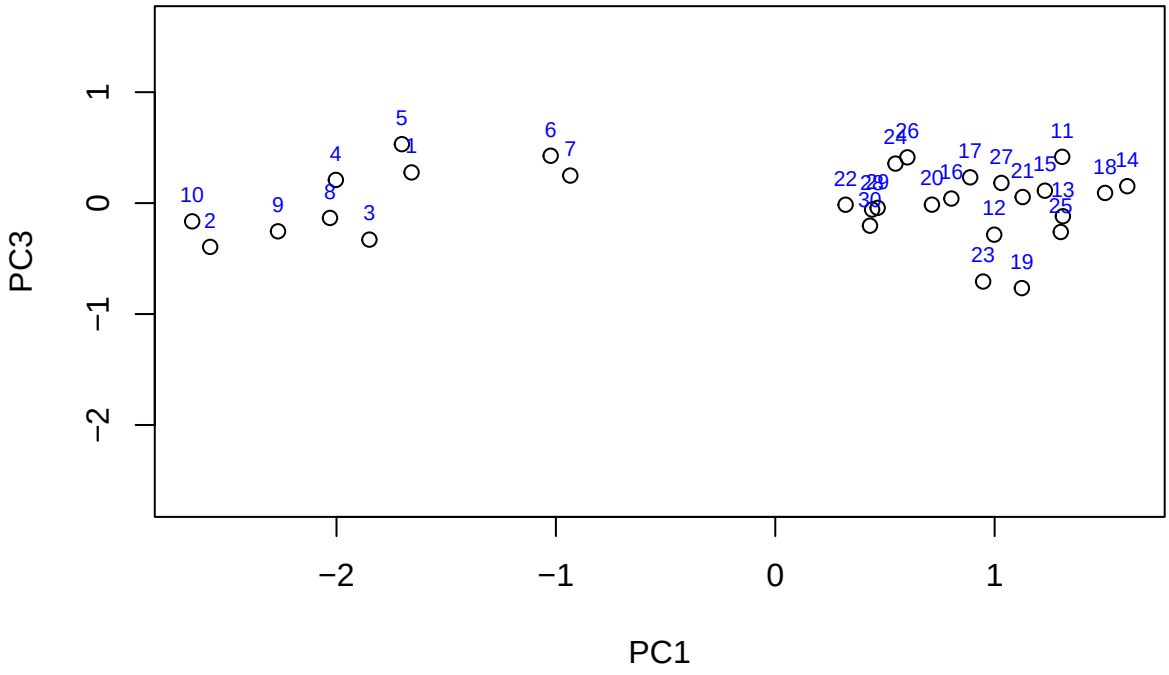
i) PC1 vs PC2

```
plot(scores[,1:2],
     xlim=c(min(scores[,1:2]),max(scores[,1:2])),
     ylim=c(min(scores[,1:2]),max(scores[,1:2])))
text(scores[,1], scores[,2], rownames(scores), col="blue", cex=0.7, pos=3)
```



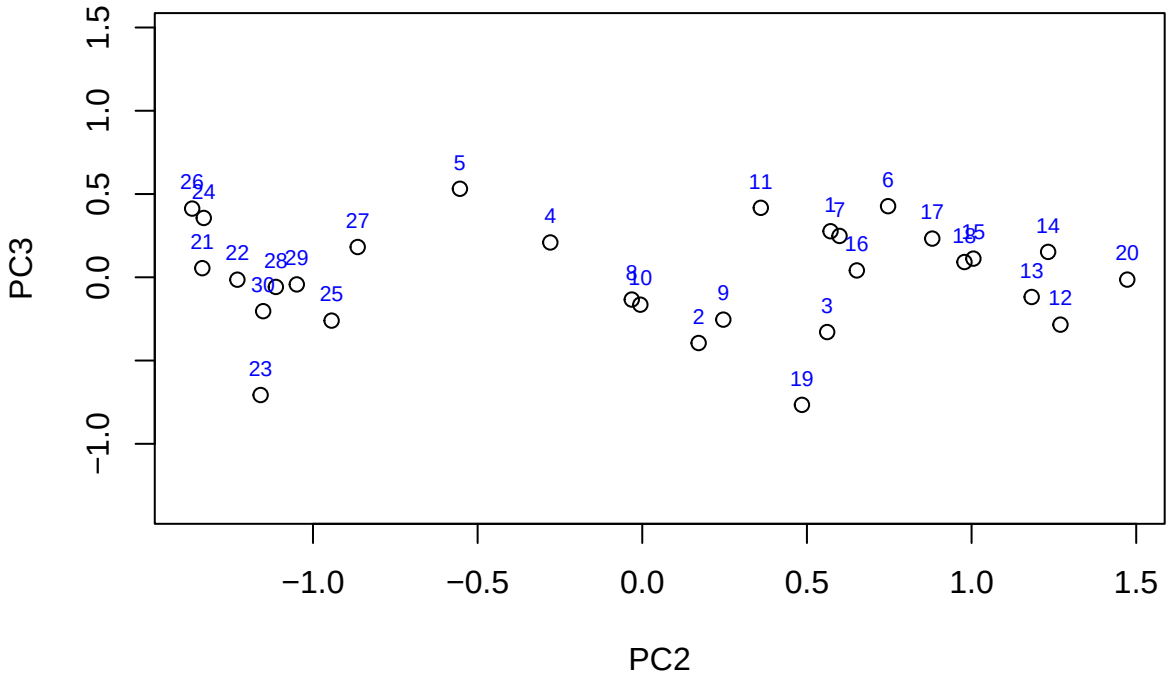
ii) PC1 vs PC3

```
plot(scores[,c(1,3)],
     xlim=c(min(scores[,c(1,3)]),max(scores[,c(1,3)])),
     ylim=c(min(scores[,c(1,3)]),max(scores[,c(1,3)])))
text(scores[,1], scores[,3], rownames(scores), col="blue", cex=0.7, pos=3)
```



iii) PC2 vs. PC3.

```
plot(scores[,2:3],
     xlim=c(min(scores[,2:3]),max(scores[,2:3])),
     ylim=c(min(scores[,2:3]),max(scores[,2:3])))
text(scores[,2], scores[,3], rownames(scores), col="blue", cex=0.7, pos=3)
```


PC3
PC2
1.5
1.0
0.5
0.0
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
26
24
21
22
28
29
30
27
25
23
5
4
8
10
9
2
11
1
7
16
3
19
6
17
18
15
14
13
12
20