CMDA-3654

Homework 10

Eduardo Salvador

Due as a .pdf upload

Problem 1: [100 pts] More PCA

Food technicians are interested in comparing pizza doughs from Naples, made with traditional methods, with doughs from other places. Even though they taste different, previous research has not found good univariate characteristics to distinguish between the different types of doughs. Therefore, the food technician collected a multivariate data set.

Doughs from six restaurants are investigated, with five samples from each restaurant. The first four restaurants are famous Naples restaurants and the last two are from other Italian cities. For each dough two mechanical test (measuring pressure load and deformation volume) and four microbiological/chemical tests (counting the bacteria in the dough, counting the yeast, measuring pH and measuring total titratable acidity) are performed. The data is given in `doughs.csv`.

a. Produce a principal component analysis object called myPCAfit for the first 6 columns of the doughs dataset only (the last column is just a restaurant ID). Remember that you have to do either `scale. = TRUE` if using `prcomp()` or `cor = TRUE` if using `princomp()`.

```
#Reading csv file first 6 columns
doughs<-read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /Assignments/HW10/doughs.
doughs[,1:6]
```

```
   PressureLoad   DefVol Bacteria    Yeast       pH       TTA
1     130.02568 35.52756 5.757972 6.542892 6.097234 1.1110839
2     123.86851 40.34567 6.034986 6.317691 6.063427 1.1380931
3     115.46597 26.58752 5.410269 6.668726 5.784212 1.1991341
4     127.58992 41.19118 5.870146 6.534598 6.148323 0.9865285
5     132.26647 45.74070 5.709261 6.838558 5.837073 1.6194396
6     122.15912 56.40890 7.318165 6.382213 5.865991 1.2675547
7     120.19586 33.66885 5.467638 6.436234 5.811155 1.3182378
8     126.89051 46.73313 7.784137 7.380511 5.743203 1.5392976
9     115.51803 35.78070 6.235717 7.306624 5.511193 1.4904791
10    113.86556 50.23146 7.928312 7.503896 5.718868 1.3443432
11    103.65008 28.95107 4.732192 6.990209 5.528908 1.6071463
12    108.79718 35.04796 8.945713 7.221850 5.728546 1.4261939
13    115.14925 46.68753 6.096369 7.935406 5.671798 1.6282453
14    106.21542 34.87870 6.184451 7.346561 5.524667 1.4803353
15    109.43033 44.43017 5.293796 7.108836 5.854408 1.5413321
16    113.28606 37.58062 5.599988 5.434055 5.918052 1.2092090
17    114.27267 31.29681 5.271806 6.440685 5.998186 0.8948270
18    112.45257 30.09542 6.410794 6.071087 5.840823 0.8860809
19    132.13855 53.15689 7.256938 6.436790 5.904885 1.3014732
20    125.52017 52.56363 7.022526 5.907304 5.991819 1.1720496
21     97.05126 22.09524 7.141561 6.147562 6.075787 1.1939264
22     94.34469 40.03719 4.623785 7.830586 6.153740 1.0555925
23    107.19449 34.88862 6.691160 7.366451 6.104683 1.2118887
24    102.57482 46.07382 8.085106 7.755833 6.092374 1.3194596
25     93.07118 37.95756 6.063649 6.874420 6.204745 0.9398126
26    111.32353 36.67969 5.238570 6.921054 6.075195 1.1870484
27     96.58765 36.64397 6.341320 7.444028 5.858414 1.2855114
28    105.08312 44.87899 5.951783 7.109943 6.009198 1.2542399
29    103.23041 47.07211 6.545814 7.401993 5.847526 1.2924900
30    106.80567 26.13801 6.504370 6.443620 6.075960 1.2830062
```

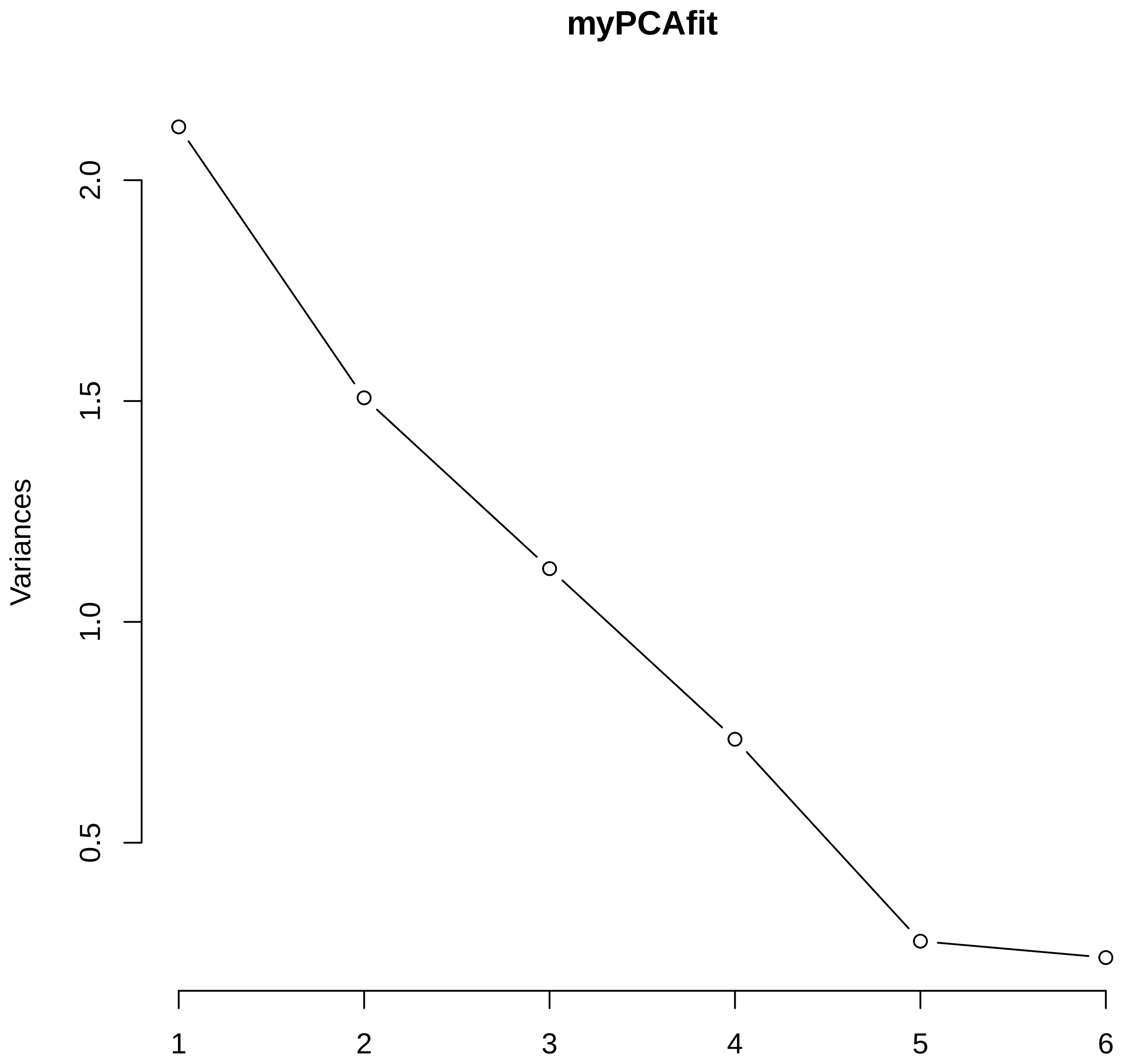
```
#using princomp to standarize the data
myPCAfit<-prcomp(doughs[,-7],scale=T)
myPCAfit
```

```
Standard deviations (1, .., p=6):
[1] 1.4562678 1.2277267 1.0585905 0.8569041 0.5263959 0.4898777

Rotation (n x k) = (6 x 6):
                   PC1        PC2        PC3         PC4        PC5
PressureLoad -0.1171101  0.6838728 -0.3728393  0.12391859  0.5569172
DefVol       -0.3326713  0.4931852  0.3674845  0.48092876 -0.5285789
Bacteria     -0.2791707  0.2758539  0.5205766 -0.74978124  0.1123330
Yeast        -0.4108043 -0.4192826  0.4097314  0.35965118  0.4658433
pH            0.5121716  0.1548093  0.4767097  0.24759902  0.3974856
TTA          -0.6054740 -0.1150096 -0.2445692  0.02292519  0.1510763
                    PC6
PressureLoad -0.23255427
DefVol        0.01910856
Bacteria     -0.01333191
Yeast        -0.37577017
pH            0.51687727
TTA           0.73280986
```

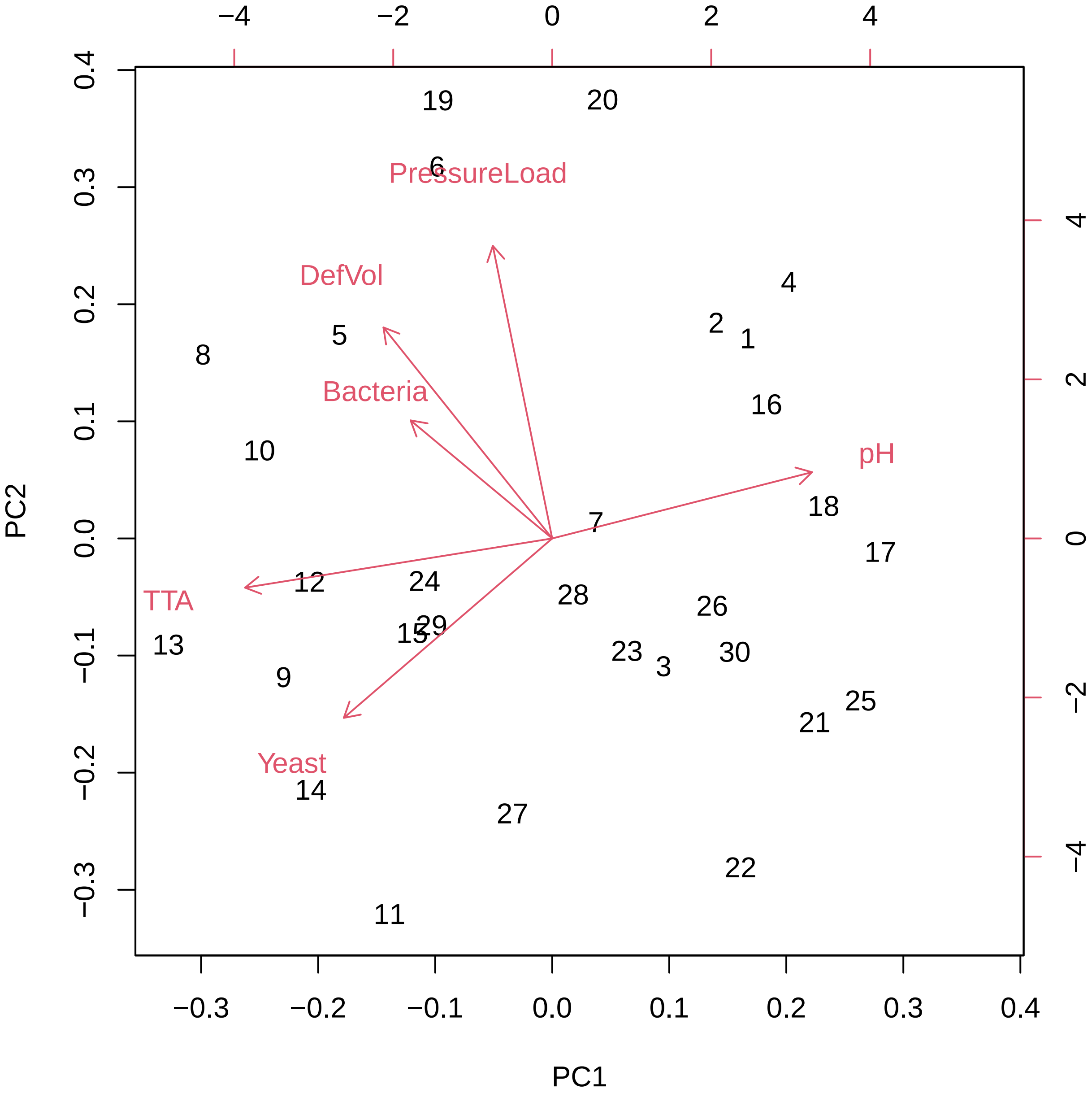
i. Plot an associated screeplot using `screeplot(myPCAfit, type = "lines")`. You can use the elbow method to determine how many principal components seem sufficient for capturing the majority of the variation of the data.

```
#Plotting screeplot
screeplot(myPCAfit, type="lines")
```



ii. Construct a biplot for PC2 versus PC1. Based upon the loadings and the results of seen in biplot can you determine which variables are the most important for PC1? What about PC2?

```
biplot(myPCAfit)
```



The function `biplot(myPCAfit)` will easily make this plot for you.

Additionally if you install and enable the `ggfortify` library then you can also make this plot doing

```
library(ggfortify)
autoplot(myPCAfit, loadings = TRUE, loadings.label = TRUE)
```

b. Find the first four principal components of the data (don't forget to scale the data – it's an option in the R functions). Do they seem to be sufficient to describe most of the variation in the data set (specifically report how much variation they describe individually and together)?

```
#Using summary function to find the four principal principal components
summary(myPCAfit)
```

```
Importance of components:
                          PC1     PC2     PC3     PC4      PC5     PC6
```

```
Standard deviation     1.4563 1.2277 1.0586 0.8569 0.52640 0.4899
Proportion of Variance 0.3534 0.2512 0.1868 0.1224 0.04618 0.0400
Cumulative Proportion  0.3534 0.6047 0.7914 0.9138 0.96000 1.0000

#PC1 has 33% variation, PC2 has  25%, PC3 has 19% and PC4 has around 12% variation. In total,
#all PC's together have around 91% variation suggesting that it is sufficient to describe most of the variatio
```

c. Use `grid.arrange()` to plot scatter plots of the first three principal components versus each other. Judging from the plots, can the first three PCs be used to discriminate doughs from Naples and doughs from other places? Does this agree with your conclusion in (a)?

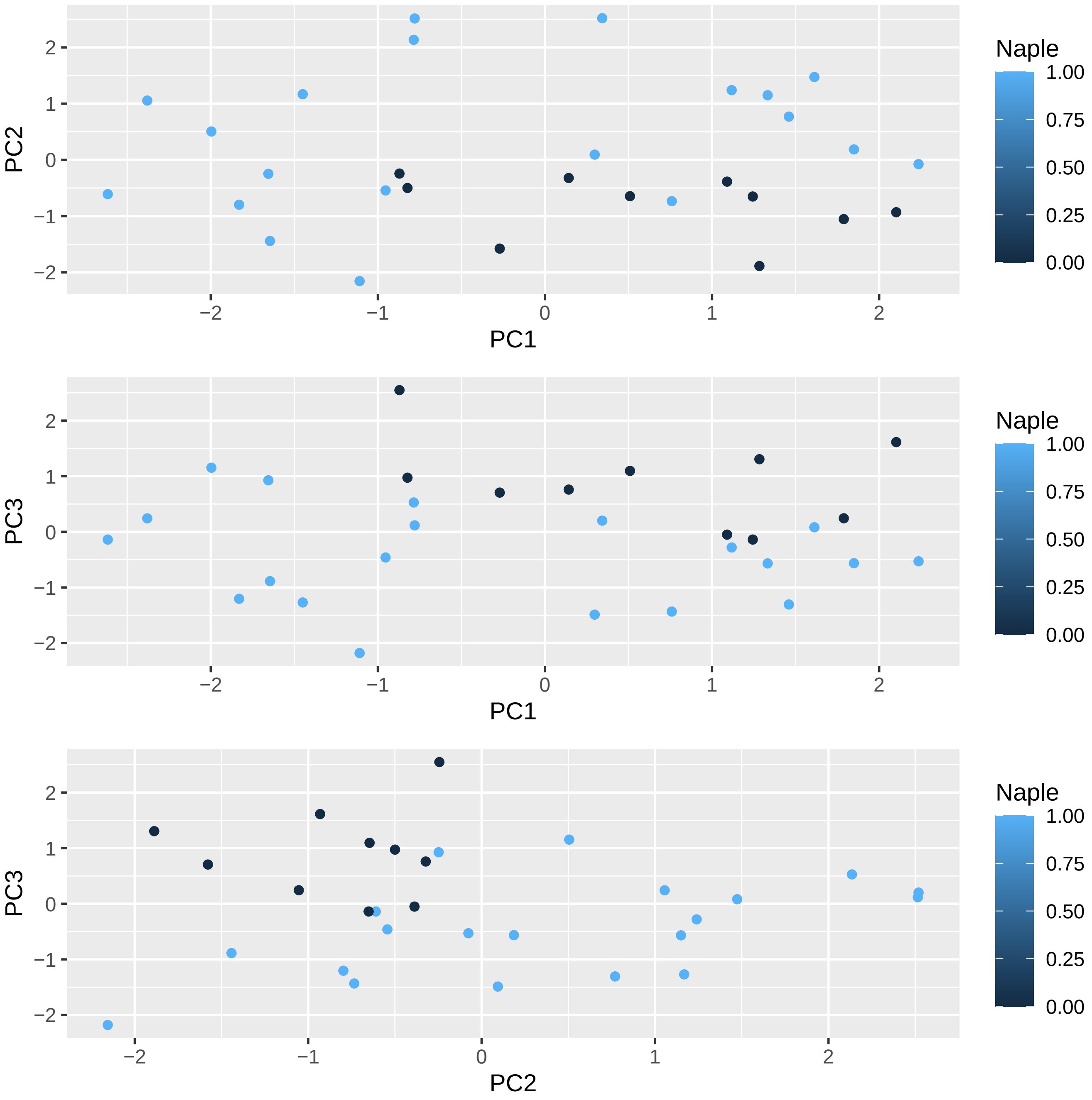
```
library(gridExtra)
library(tidyverse)

#Making changes to doughs dataframe using mutate
mdoughs<-mutate(doughs,Naple=
                  case_when(Restaurant==1~1,
                            Restaurant==2~1,
                            Restaurant==3~1,
                            Restaurant==4~1,
                            Restaurant==5~0,
                            Restaurant==6~0))

#Assigning first three principal components to mdoughs variable
mdoughs$PC1<-myPCAfit$x[,1]
mdoughs$PC2<-myPCAfit$x[,2]
mdoughs$PC3<-myPCAfit$x[,3]

#Using qplot to wrap for creating a number of different types of plots
#Assigning variable for PC1 vs PC2
scp1<-qplot(PC1,PC2,data=mdoughs,col=Naple)
#Assigning variable for PC1 vs PC3
scp2<-qplot(PC1,PC3,data=mdoughs,col=Naple)
#Assigning variable for PC1 vs PC2
scp3<-qplot(PC2,PC3,data=mdoughs,col=Naple)

#Using grid.arrange to plot scatter plots
grid.arrange(scp1,scp2,scp3)
```



```
#Judging from the plots, the first 3 can be used to discriminate doughs from Naples and others
#It does agree
```

Hint: You will need to use ggplot multiple times to get plot objects p1, p2, etc for plotting the principal components versus each other and you need to be plotting the component scores for the different restaurant. You need to colorize the restaurants from Naples to be the same color and the restaurants that aren't from Naples a different color.

d. If you had to perform a specific statistical learning method to classify the doughs based upon these features, which method would it be and why?

```
#If I had to perform a specific statistical method to classify the doughs based upon the features,
#I would use K means clustering because it can divide the dataset into non-overlapping data points.
```