MAST90138 Assignment 2

Haonan Zhong

Question 1a

```r
wheat <- read.csv("Wheat data.txt", sep = "", header = F)
X <- scale(wheat[, -c(8)], scale = FALSE)
PCX <- prcomp(X, retx = TRUE)
(lambda <- PCX$sdev^2)
```

```
## [1] 1.079333e+01 2.129455e+00 7.363003e-02 1.288749e-02 2.748227e-03
## [6] 1.570450e-03 2.965544e-05
```

```r
(gamma <- PCX$rotation)
```

```
##             PC1          PC2         PC3         PC4          PC5          PC6
## V1 -0.884228505  0.100805775  0.26453354  0.19944949 -0.137172970  0.280639558
## V2 -0.395405417  0.056489625 -0.28251995 -0.57881686  0.574756029 -0.301558638
## V3 -0.004311324 -0.002894744  0.05903584  0.05776023 -0.053104536 -0.045229054
## V4 -0.128544478  0.030621731 -0.40014946 -0.43610024 -0.786997760 -0.113437606
## V5 -0.111059139  0.002372293  0.31923869  0.23416358 -0.144802899 -0.896267845
## V6  0.127615624  0.989410476  0.06429754 -0.02514736 -0.001575639  0.003287998
## V7 -0.128966499  0.082233392 -0.76193973  0.61335659  0.087653609 -0.109923643
##             PC7
## V1 -0.025398239
## V2  0.065839904
## V3  0.994125646
## V4  0.001431435
## V5 -0.081549900
## V6  0.001142692
## V7  0.008971926
```

```r
(fracvar <- lambda/sum(lambda))
```

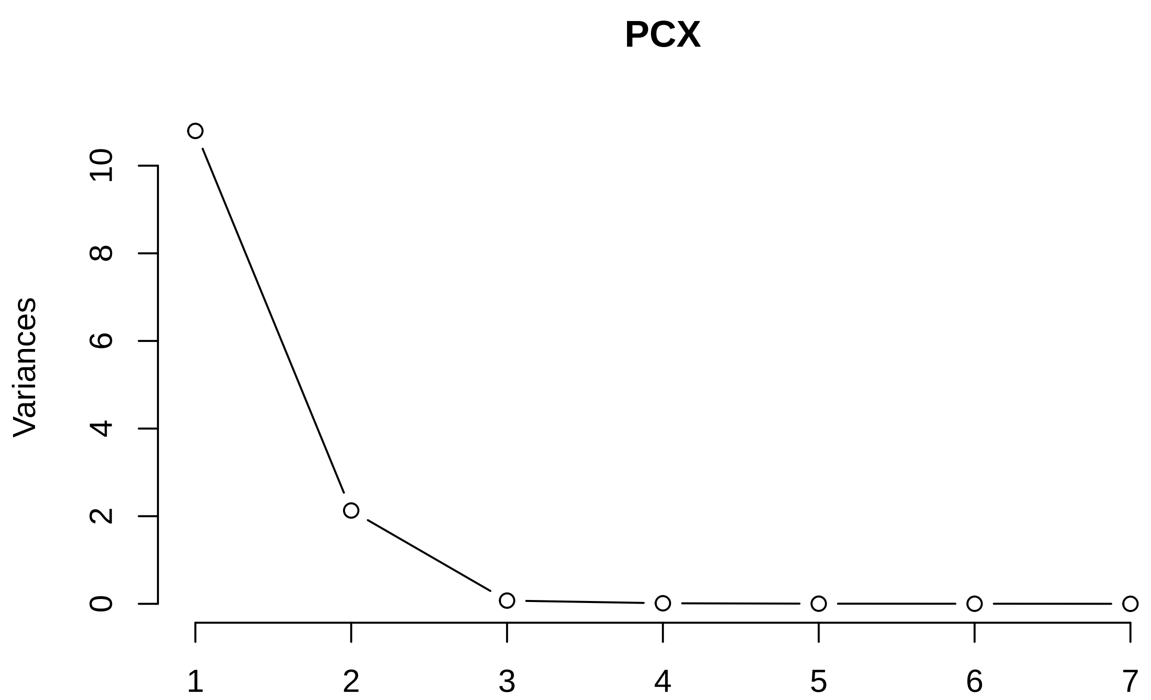
```
## [1] 8.293852e-01 1.636325e-01 5.657909e-03 9.903061e-04 2.111803e-04
## [6] 1.206771e-04 2.278796e-06
```

As we can see from the output, the first and the second principle components was able to explain about 82.9% and 16.4% of the variability of the data, respectively. And the third component explains around 0.57% of variances. Whilst, the rest of the PCs only explains a tiny portion of variances.

```r
(cumuprop <- cumsum(lambda)/sum(lambda))
```

```
## [1] 0.8293852 0.9930176 0.9986756 0.9996659 0.9998770 0.9999977 1.0000000
```

```
screeplot(PCX, type = "line")
```



Visually, one can look for an elbow in the screeplot and stop there. In our case, we should keep the first three principle components, with the first three PCs we explain almost 100% of the variability of the data.

Question 1b

According to the eigenvectors given in the question 1a, we have

$$Y_1 = -0.884X_1 - 0.395X_2 - 0.004X_3 - 0.129X_4 - 0.111X_5 + 0.127X_6 - 0.129X_7$$

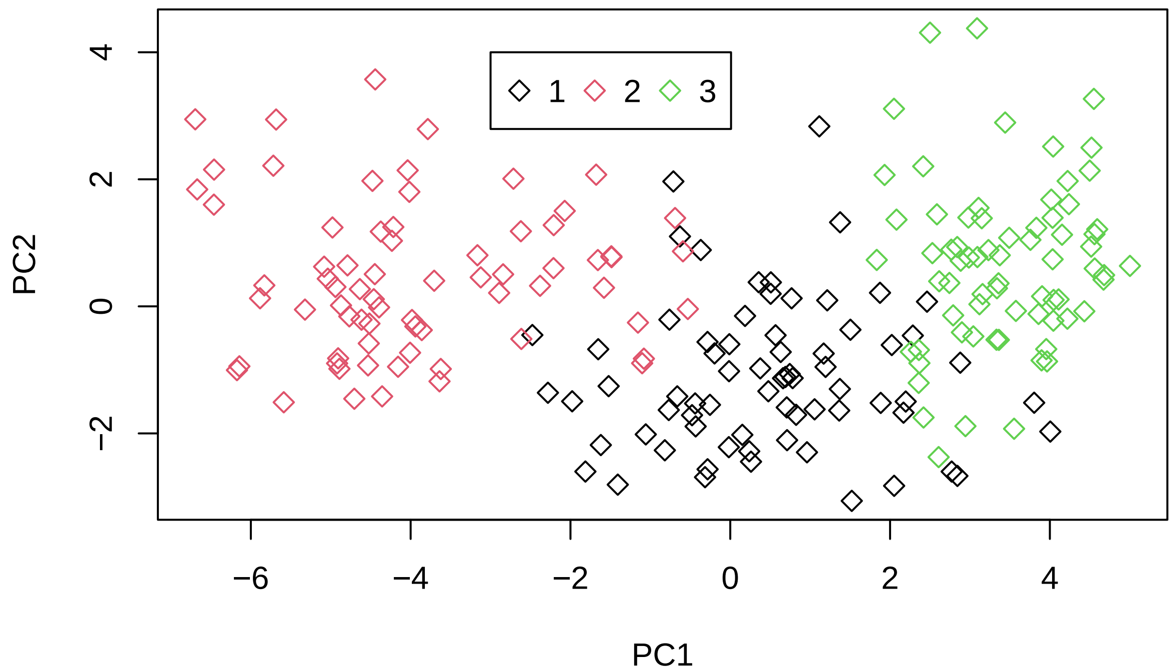
Thus, the first PC is mostly dominated by the first variable X_1, followed by X_2. On the other hand, X_4, X_5, X_6, and X_7 are quite similar in terms of contribution, whilst X_3 plays the smallest role in the construction of the first PC.

$$Y_2 = 0.101X_1 + 0.056X_2 - 0.003X_3 + 0.031X_4 + 0.002X_5 + 0.989X_6 + 0.082X_7$$

Thus, X_6 plays the major role in the construction of the second PC, followed by X_1.

Question 1c

```
PC1 <- X %*% as.matrix(gamma[,1])
PC2 <- X %*% as.matrix(gamma[,2])
plot(x = PC1, y = PC2, col = wheat$V8, pch=5)
legend(-3, 4, horiz = TRUE, unique(wheat$V8), col=1:length(wheat$V8), pch=5)
```



Based on the scatter plot between the first two principle components, we can see that PC1 captured most of the variation driven by varieties of wheat the kernel comes from, as it divides the data points into three clusters, where black points indicates data point from group 1, red points indicates data point from group 2, and green points represents data point from group 3.