A5P2

Dawu Liu

In this assignment, principal component is written as **PC** sometimes for short.

Also, here are the keys for Flea Beetle data that will be analyzed below:

Distance TG = distance of transverse groove from posteriori border of prothorax

Elytra = length of elytra

Second Antenna = length of second antennal joint

Third Antenna = length of third antennal joint

(a)
PCA on Haltica oleracea group

i. Sample covariance matrix **S**

```
##                 Distance TG   Elytra Second Antenna Third Antenna
## Distance TG        187.5965 176.8626        48.3713      113.5819
## Elytra             176.8626 345.3860        75.9795      118.7807
## Second Antenna      48.3713  75.9795        66.3567       16.2427
## Third Antenna      113.5819 118.7807        16.2427      239.9415
```

ii. The eigenvalues are:

```
## [1] 561.3057 168.9858  65.2771  43.7120
```

The first two eigenvalues are relatively large compared to the rest. The first eigenvalue accounts for the majority of the total variance. (table shown in iii)

iii.

Criteria 1, eigenvalues and their cumulative proportions table

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   561.3057          66.8794                     66.8794
## Dim.2   168.9858          20.1346                     87.0140
## Dim.3    65.2771           7.7777                     94.7917
## Dim.4    43.7120           5.2083                    100.0000
```

This method suggests to keep **2** principal component which gives 87.0% of the total variance.

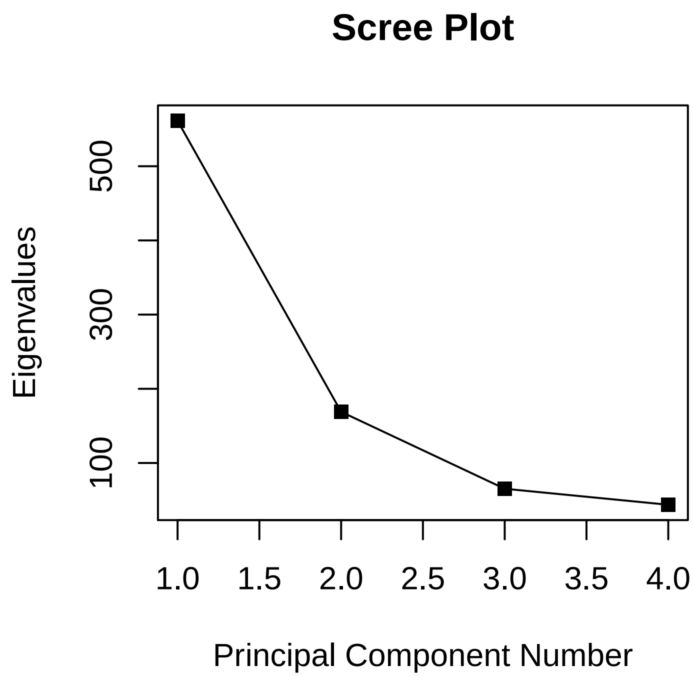
Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

The mean of the eigenvalues is:

```
## [1] 209.8202
```

This method suggests to keep **1** principal component.

Criteria3, scree plot



The "bend" occurs at PC2 is the strongest(the bend at PC3 is only slight weaker), indicating from PC2 and on, the the eigenvalues are relatively small. This method suggests to keep **1** principal component.

Overall, the criteria suggest **1** principal component should be retained. This is a very close call, since criteria 1 suggests to keep 2, criteria 2 suggests to keep 1, and in criteria 3 the angles at PC2 and PC3 are nearly identical.

But in order to make a scatter plot, we will use 2 principal components and proceed.

iv. The eigenvectors for the principal components:

$(\tilde{\hat{e_1}})^T =$

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.4997         0.7187         0.1740         0.4511
```

$(\tilde{\hat{e_2}})^T =$

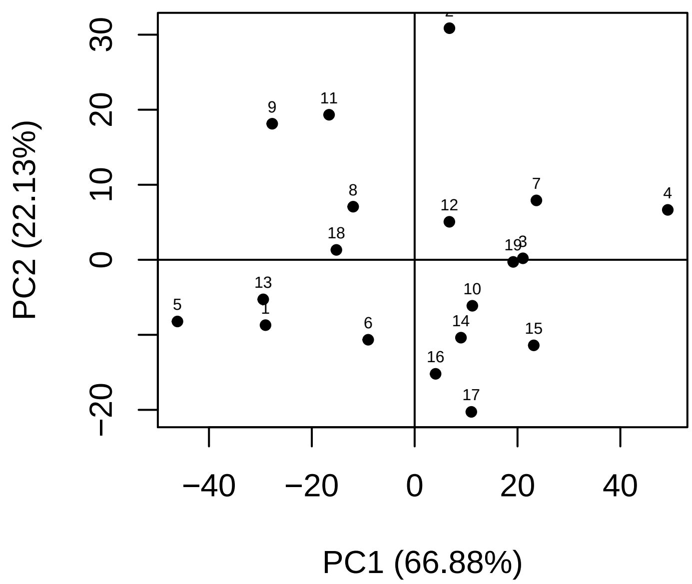
```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.0092        -0.4844        -0.2203         0.8466
```

v.

PC1 depends on the length of elytra the most, followed by the distance of transverse groove from posteriori border of prothorax and the length of third antennal joint. All four variables contribute a considerable amount to PC1, indicating PC1 is related to the size of the beetles.

PC2 depends on the length of third antennal joint the most, then the length of elytra but in opposite direction, showing there is some contrast between those two. This indicats PC2 is related to the shape of the beetles.

vi. Scatter plot for PC2 vs PC1



The data appear to have a weak linear relationship. The data can somewhat fit into an ellipse shape. Observation number 4 on the far right for example, has the largest PC1 value, indicating it is the largest beetle in size.

PCA on Haltica carduorum group

i. Sample covariance matrix **S**

```
##                Distance TG   Elytra Second Antenna Third Antenna
## Distance TG       101.8395 128.0632        36.9895       32.5921
## Elytra            128.0632 389.0105       165.3579       94.3684
## Second Antenna     36.9895 165.3579       167.5368       66.5263
## Third Antenna      32.5921  94.3684        66.5263      177.8816
```

ii. The eigenvalues are:

```
## [1] 555.6931 145.4463  93.4637  41.6652
```

The first two eigenvalues are relatively large compared to the rest. The first eigenvalue accounts for the majority of the total variance. (table shown in iii)

iii.

Criteria 1, eigenvalues and their cumulative proportions table

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   555.6931          66.4491                     66.4491
## Dim.2   145.4463          17.3923                     83.8414
## Dim.3    93.4637          11.1763                     95.0177
## Dim.4    41.6652           4.9823                    100.0000
```

This method suggests to keep **2** principal component which gives 83.8% of the total variance.

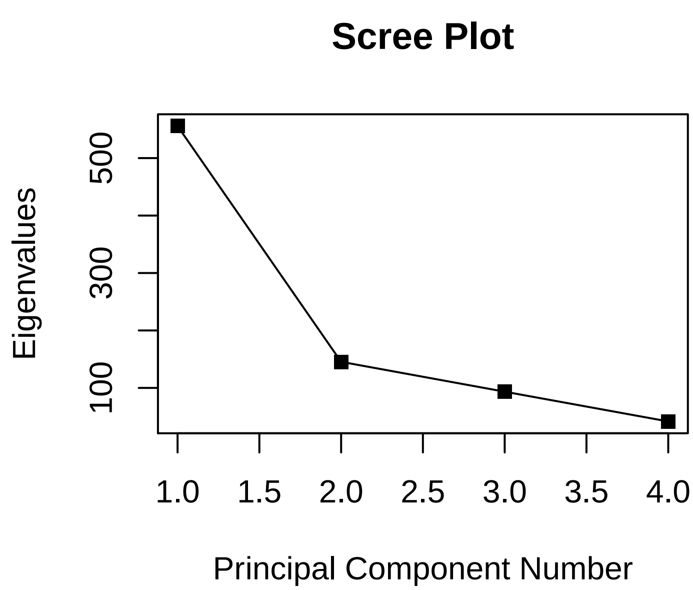
Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

The mean of the eigenvalues is:

```
## [1] 209.0671
```

This method suggests to keep **1** principal component.

Criteria3, scree plot



The "bend" occurs at PC2, indicating from PC2 and on, the the eigenvalues are relatively small. This method suggests to keep **1** principal component.

Overall, the criteria suggest **1** principal component should be retained.

But in order to make a scatter plot, we will use 2 principal components.

iv. The eigenvectors for the principal components:

$(\tilde{e_1})^T =$

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.2837         0.8069         0.4222         0.3004
```

$(\tilde{e_2})^T =$

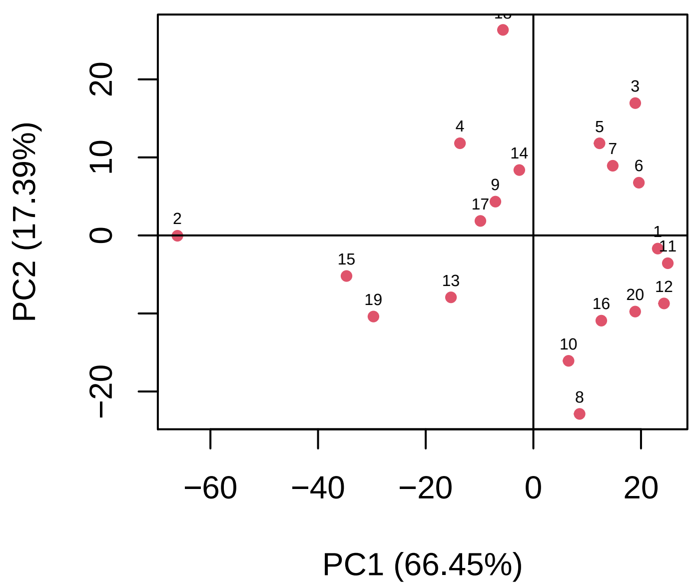
```
##    Distance TG         Elytra Second Antenna  Third Antenna
##        -0.2007        -0.3390         0.1360         0.9090
```

v.

PC1 depends on the length of elytra the most, then depends on the rest of the three variables almost evenly. All four variables contribute a considerable amount to PC1 in the same direction, indicating PC1 is related to the size of the beetles.

PC2 primarily represents the length of third antennal joint. Distance TG and the length of elystra have a opposite sign to it, show there are some contrast between those three.

vi. Scatter plot for PC2 vs PC1



Observation number 2 on the far left has the smallest PC1 values, indicating it is the smallest beetle in size, appears to be an outlier. Observation number 18 on the top for example, has long third antenna, but probably short elytra. There is a gap around the origin which in the the middle between the observations, which might due to small sample size.

PCA using data from both groups

i. Sample covariance matrix **S**

```
##                  Distance TG   Elytra Second Antenna Third Antenna
## Distance TG         196.8880  56.9372       -34.4798      -19.0715
## Elytra               56.9372 502.7085       239.4251      245.3401
## Second Antenna      -34.4798 239.4251       216.0445      159.4514
## Third Antenna       -19.0715 245.3401       159.4514      341.8313
```

ii. The eigenvalues are:

```
## [1] 818.2734 238.2294 144.9609  56.0086
```

The first eigenvalue is relatively large and contributes to the majority of the total variance. (table shown in iii)

iii.

Criteria 1, eigenvalues and their cumulative proportions table

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   818.2734          65.0729                     65.0729
## Dim.2   238.2294          18.9451                     84.0180
## Dim.3   144.9609          11.5280                     95.5459
## Dim.4    56.0086           4.4541                    100.0000
```

This method suggests to keep **2** principal component which gives 84.02% of the total variance.

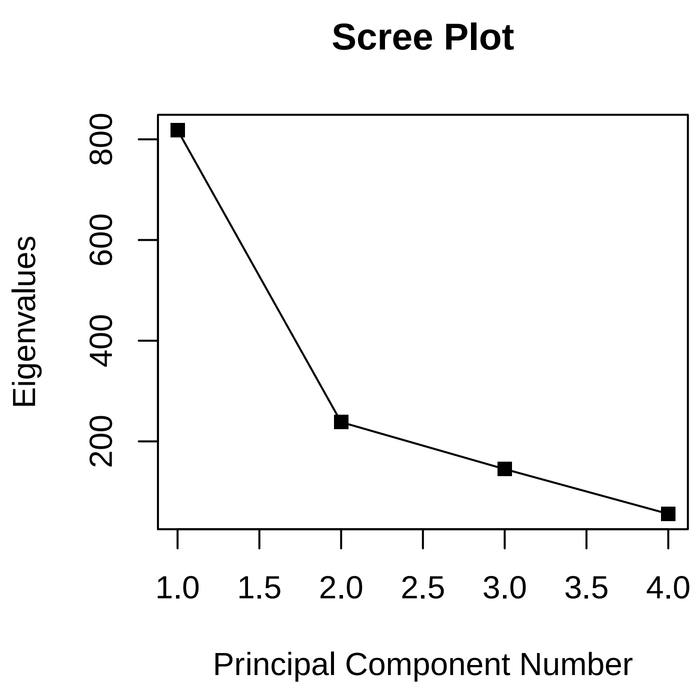
Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

The mean of the eigenvalues is:

```
## [1] 314.3681
```

This method suggests to keep **1** principal component.

Criteria3, scree plot



The "bend" occurs at PC2, indicating from PC2 and on, the the eigenvalues are relatively small. This method suggests to keep **1** principal component.

Overall, the criteria suggest **1** principal component should be retained.

But in order to make a scatter plot, we will use 2 principal components.

iv. The eigenvectors for the principal components:

$(\tilde{\hat{e}}_1)^T =$

```
##    Distance TG          Elytra Second Antenna  Third Antenna
##        -0.0276         -0.7365        -0.4294        -0.5219
```

$(\tilde{\hat{e}}_2)^T =$

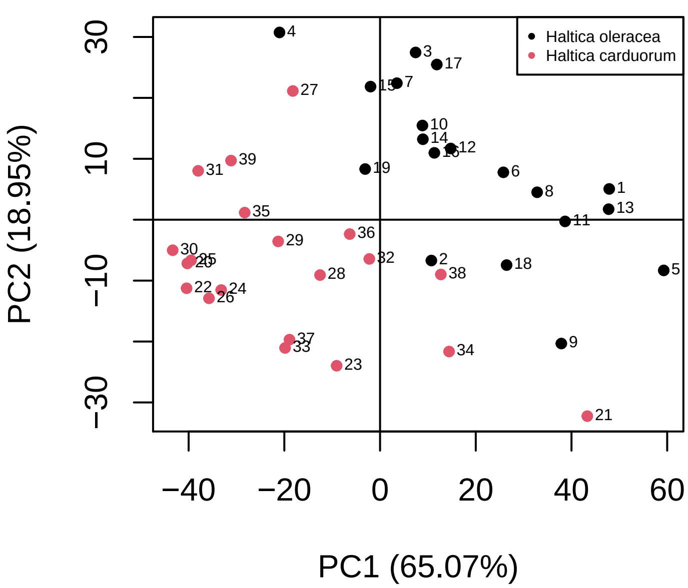
```
##    Distance TG          Elytra Second Antenna  Third Antenna
##         0.8303          0.3548        -0.1991        -0.3808
```

v.

PC1 strongly depends on the length of elytra, followed by the length of third and second antenna. There is almost no dependency on Distance TG. Since all the coefficients are in the same directions, PC1 might be an indication of the size of the beetles.

PC2 primarily depends on distance TG. It also shows opposite relationship between distance TG, the length of elytra on one hand, and the length of second, third antenna on another hand.

vi. Scatter plot for PC2 vs PC1



Haltica oleracea have higher average values in both PC1 and PC2 than Haltica carduorum, it appear those two group belongs to two different clusters. And each group data appears to somewhat fit in an ellipse. Since the values of the eigenvectors for PC1 are all negative, this shows Haltica oleracea tends to be slightly smaller than Haltica carduorum in size. Both groups have a negative relationship between PC1 and PC2.

(b)

Each type of the analysis has their own advantages. Since all four variables are the measurements of the length on certain body parts, the first two analyses give more accurate information on the relationships between body sizes within each group. The third analysis (using combined data) shows how one group is related to another, telling us there is a clear divide between the two groups.

Code used to solve the questions(graphs are hidden):

```
rm(list = ls())
library(readxl)
library(factoextra)
data <- read_excel("C:/Users/John/Desktop/STAT 445/Data/fleabeetledata.xlsx")
```

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ...
```

```
data <- as.data.frame(data[-(1:2),-1])

# group 1
X1 <- sapply(data[-20,1:4], as.numeric)
colnames(X1) <- c("Distance TG", "Elytra", "Second Antenna", "Third Antenna")
X1_pr = prcomp(X1, center = TRUE, scale. = FALSE)
X1_eigen_table = get_eigenvalue(X1_pr)
S1 <- cov(X1)
round(S1,4)
```

```
##                 Distance TG   Elytra Second Antenna Third Antenna
## Distance TG        187.5965 176.8626        48.3713      113.5819
## Elytra             176.8626 345.3860        75.9795      118.7807
## Second Antenna      48.3713  75.9795        66.3567       16.2427
## Third Antenna      113.5819 118.7807        16.2427      239.9415
```

```
round(X1_eigen_table[,1], 4)
```

```
## [1] 561.3057 168.9858  65.2771  43.7120
```

```
round(X1_eigen_table,4)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   561.3057          66.8794                     66.8794
```

```
## Dim.2   168.9858          20.1346                      87.0140
## Dim.3    65.2771           7.7777                      94.7917
## Dim.4    43.7120           5.2083                     100.0000
```

```r
mean(X1_eigen_table[,1])
```

```
## [1] 209.8202
```

```r
plot(X1_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
```

```r
e_X1_1 <- X1_pr$rotation[,1]
e_X1_2 <- X1_pr$rotation[,2]
round(e_X1_1, 4)
```

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.4997         0.7187         0.1740         0.4511
```

```r
round(e_X1_2, 4)
```

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.0092        -0.4844        -0.2203         0.8466
```

```r
plot(X1_pr$x[,1], X1_pr$x[,2], xlab = "PC1 (66.88%)", ylab = "PC2 (22.13%)", pch = 20)
text(X1_pr$x[,1], X1_pr$x[,2], pos = 3, offset = 0.3, cex = 0.5)
abline(v = 0, h = 0)
```

```r
# group2
X2 <- sapply(data[,5:8], as.numeric)
colnames(X2) <- c("Distance TG", "Elytra", "Second Antenna", "Third Antenna")
X2_pr = prcomp(X2, center = TRUE, scale. = FALSE)
X2_eigen_table = get_eigenvalue(X2_pr)
S2 <- cov(X2)
round(S2,4)
```

```
##                 Distance TG   Elytra Second Antenna Third Antenna
## Distance TG        101.8395 128.0632        36.9895       32.5921
## Elytra             128.0632 389.0105       165.3579       94.3684
## Second Antenna      36.9895 165.3579       167.5368       66.5263
## Third Antenna       32.5921  94.3684        66.5263      177.8816
```

```r
round(X2_eigen_table[,1], 4)
```

```
## [1] 555.6931 145.4463  93.4637  41.6652
```

```r
round(X2_eigen_table,4)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   555.6931          66.4491                     66.4491
## Dim.2   145.4463          17.3923                     83.8414
## Dim.3    93.4637          11.1763                     95.0177
## Dim.4    41.6652           4.9823                    100.0000
```

```r
mean(X2_eigen_table[,1])
```

```
## [1] 209.0671
```

```r
plot(X2_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
```

```r
e_X2_1 <- X2_pr$rotation[,1]
e_X2_2 <- X2_pr$rotation[,2]
round(e_X2_1, 4)
```

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.2837         0.8069         0.4222         0.3004
```

```r
round(e_X2_2, 4)
```

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##        -0.2007        -0.3390         0.1360         0.9090
```

```r
plot(X2_pr$x[,1], X2_pr$x[,2], xlab = "PC1 (66.45%)", ylab = "PC2 (17.39%)", pch = 20, col = 2)
text(X2_pr$x[,1], X2_pr$x[,2], pos = 3, offset = 0.3, cex = 0.5)
abline(v = 0, h = 0)
```

```r
# combined
X <-rbind(X1,X2)
X_pr = prcomp(X, center = TRUE, scale. = FALSE)
X_eigen_table = get_eigenvalue(X_pr)
S <- cov(X)
round(S,4)
```

```
##                 Distance TG   Elytra Second Antenna Third Antenna
## Distance TG        196.8880  56.9372       -34.4798      -19.0715
## Elytra              56.9372 502.7085       239.4251      245.3401
## Second Antenna     -34.4798 239.4251       216.0445      159.4514
## Third Antenna      -19.0715 245.3401       159.4514      341.8313
```

```r
round(X_eigen_table[,1], 4)
```

```
## [1] 818.2734 238.2294 144.9609  56.0086
```

```r
round(X_eigen_table,4)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   818.2734          65.0729                     65.0729
## Dim.2   238.2294          18.9451                     84.0180
## Dim.3   144.9609          11.5280                     95.5459
## Dim.4    56.0086           4.4541                    100.0000
```

```r
mean(X_eigen_table[,1])
```

```
## [1] 314.3681
```

```r
plot(X_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
```

```r
e_X_1 <- X_pr$rotation[,1]
e_X_2 <- X_pr$rotation[,2]
round(e_X_1, 4)
```

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##        -0.0276        -0.7365        -0.4294        -0.5219
```

```r
round(e_X_2, 4)
```

```
##    Distance TG         Elytra Second Antenna  Third Antenna
##         0.8303         0.3548        -0.1991        -0.3808
```

```r
PR <- cbind(X_pr$x[,1], X_pr$x[,2])
PR <- cbind(PR,c(1))
PR[20:39,3] <- 2
plot(PR[,1], PR[,2], xlab = "PC1 (65.07%)", ylab = "PC2 (18.95%)", pch = 20, col = PR[,3])
text(PR[,1], PR[,2], pos = 4, offset = 0.2, cex = 0.5)
abline(v = 0, h = 0)
par(xpd = TRUE)
legend("topright", legend=c("Haltica oleracea", "Haltica carduorum"),
       lty= c(0,0), pch=c(20, 20),col = c(1,2), cex= 0.5)
```