Multivariate Statistical Analysis

Homework 3

Lucas Fellmeth, Helen Kafka, Sven Bergmann

02/28/24

Problem 1

```
swissbanknotes <- read.csv("Data_csv/SwissBankNotes.csv")
swissbanknotes$class <- factor(swissbanknotes$class)
banknote_type <- swissbanknotes$class
```

(a)

```
swissbanknotes2 <- scale(swissbanknotes[, -7])
summary(swissbanknotes2)
```

```
##      length          height_left       height_right     inner_frame_lower
##  Min.   :-2.91060   Min.   :-3.1064   Min.   :-2.3672   Min.   :-1.5350
##  1st Qu.:-0.78608   1st Qu.:-0.6135   1st Qu.:-0.6348   1st Qu.:-0.8428
##  Median : 0.01062   Median : 0.2174   Median : 0.1077   Median :-0.2198
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.54175   3rd Qu.: 0.7714   3rd Qu.: 0.6645   3rd Qu.: 0.8186
##  Max.   : 3.72855   Max.   : 2.4333   Max.   : 2.8299   Max.   : 2.2723
##  inner_frame_upper     diagonal
##  Min.   :-3.67459   Min.   :-2.32889
##  1st Qu.:-0.68560   1st Qu.:-0.85354
##  Median :-0.06289   Median :-0.02907
##  Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.68435   3rd Qu.: 0.88218
##  Max.   : 2.05431   Max.   : 1.66324
```

```
pca_banknotes <- princomp(swissbanknotes2)
summary(pca_banknotes)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.7119668 1.1276938 0.9298857 0.66896923 0.51704305
## Proportion of Variance 0.4909264 0.2130140 0.1448388 0.07496145 0.04477948
## Cumulative Proportion  0.4909264 0.7039403 0.8487791 0.92374054 0.96852002
##                            Comp.6
## Standard deviation     0.43351526
## Proportion of Variance 0.03147998
## Cumulative Proportion  1.00000000
```

The cumulative proportion of the first two components is 0.704.

(b)

```
pca_banknotes$loadings[, 1]
```

```
##            length       height_left      height_right inner_frame_lower
##       0.006987029      -0.467758161      -0.486678705      -0.406758327
## inner_frame_upper          diagonal
##      -0.367891118       0.493458317
```

Component 1

The largest elements (in absolute value) correspond to height_left, height_right, diagonal length. So it's an index of height / diagonal length, where diagonal length has a negative sign. Observations with large negative component scores ksi_{i1} are shorter banknotes with longer diagonal whereas observations with large positive component scores ksi_{i1} are taller banknotes with a shorter diagonal.

```
pca_banknotes$loadings[, 2]
```

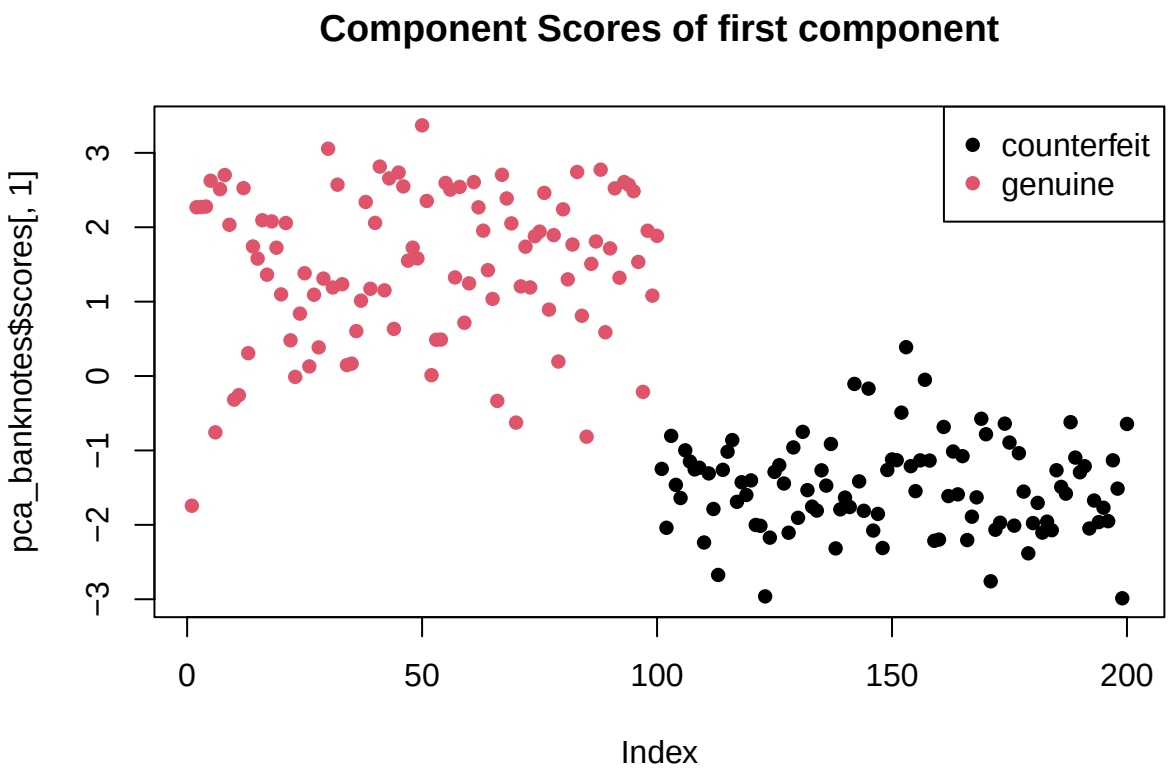
```
##            length       height_left      height_right inner_frame_lower
##        0.81549497        0.34196711        0.25245860       -0.26622878
## inner_frame_upper          diagonal
##       -0.09148667        0.27394074
```

Component 2

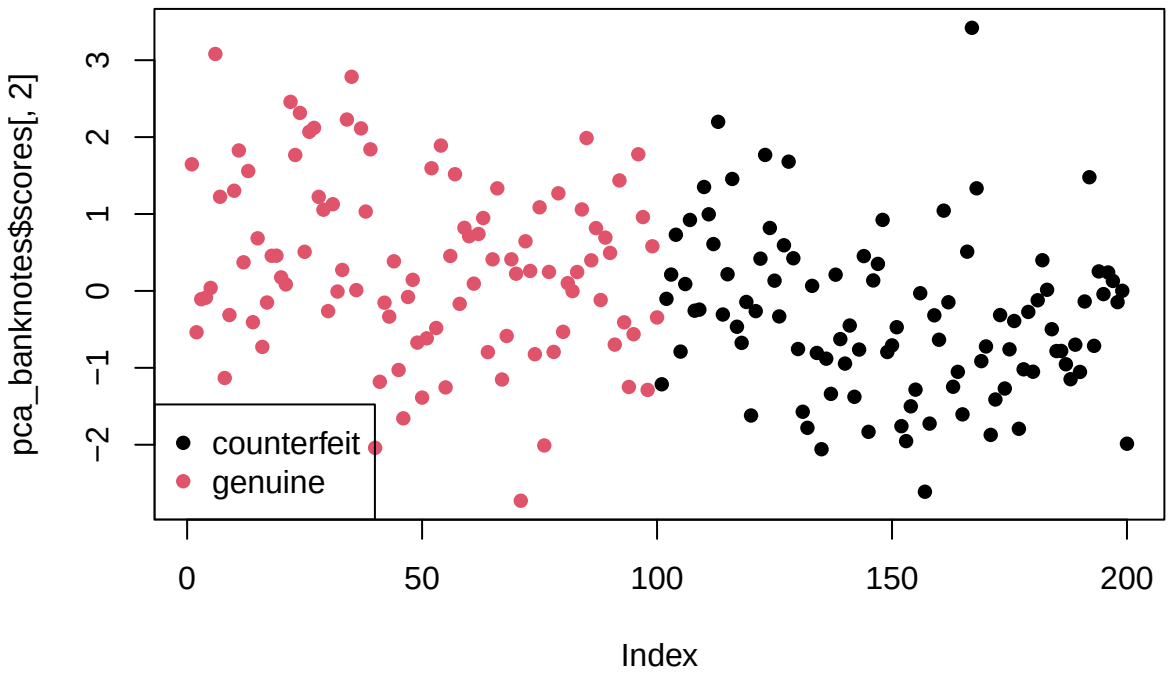
The largest element (in absolute value) corresponds to length. So it's an index of banknote length. Observations with large negative component scores ksi_{i2} are shorter banknotes whereas Observations with large positive component scores ksi_{i2} are longer banknotes.

(c)

```
plot(pca_banknotes$scores[, 1], main = "Component Scores of first component",
    col = banknote_type, pch = 16)
legend("topright", legend = levels(banknote_type), col = seq_along(banknote_type),
    pch = 16)
```




```
plot(pca_banknotes$scores[, 2], main = "Component Scores of second component",
    col = banknote_type, pch = 16)
legend("bottomleft", legend = levels(banknote_type), col = seq_along(banknote_type),
    pch = 16)
```



The two groups for counterfeit banknotes and genuine banknotes are pretty well separated for both components.

Problem 2

```
turtles <- read.csv("Data_csv/turtles.csv")
turtle_sex <- factor(turtles$sex)
turtles <- turtles[-1]
```

(a)

Before applying log we check if there are any zero values.

```
which(turtles == 0)
```

```
## integer(0)
```

```
turtles2 <- log(turtles)
turtles2 <- scale(turtles2)
pca_turtles <- princomp(turtles2)
summary(pca_turtles)
```

```
## Importance of components:
##                            Comp.1     Comp.2      Comp.3
## Standard deviation     1.6942473 0.21117868 0.149765480
## Proportion of Variance 0.9771826 0.01518177 0.007635642
## Cumulative Proportion  0.9771826 0.99236436 1.000000000
```

The cumulative proportion of the first two components is 0.992.

(b)

```
pca_turtles$loadings[, 1]
```

```
##    length     width    height
## 0.5785513 0.5784099 0.5750829
```

Component 1

The largest elements (in absolute value) correspond to length, width and height So it's a positive index of length, width and height. Observations with large negative component scores ksi_{i1} are turtles with smaller length, width and height whereas Observations with large positive component scores ksi_{i1} are turtles with larger length, width and height.

```
pca_turtles$loadings[, 2]
```

```
##     length      width     height
##  0.3936563  0.4195001 -0.8179575
```

Component 2

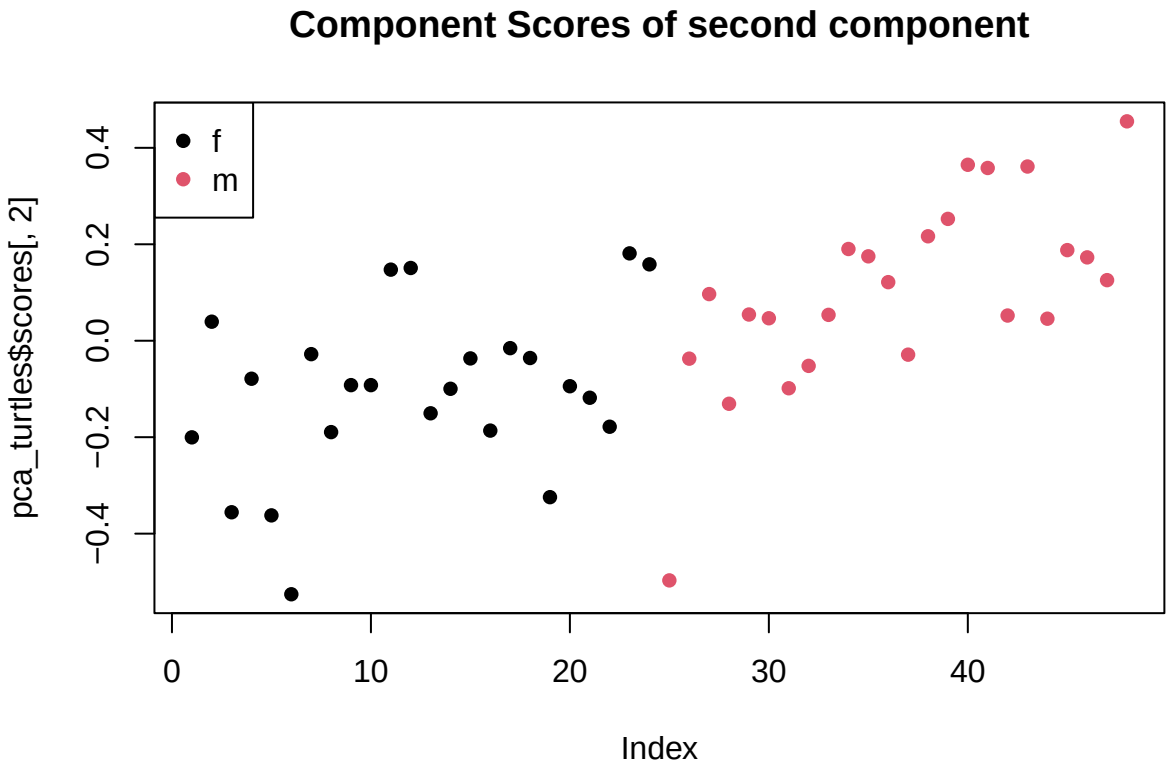
The largest element (in absolute value) corresponds to height. So it's a negative index of turtle height. Observations with large negative component scores ksi_{i2} are taller turtles whereas observations with large positive component scores ksi_{i2} are shorter turtles.

(c)

```
plot(pca_turtles$scores[, 1], main = "Component Scores of first component",
    col = turtle_sex, pch = 16)
```

```r
legend("topright", legend = levels(turtle_sex), col = seq_along(turtle_sex),
    pch = 16)
```


```r
plot(pca_turtles$scores[, 2], main = "Component Scores of second component",
    col = turtle_sex, pch = 16)
legend("topleft", legend = levels(turtle_sex), col = seq_along(turtle_sex),
    pch = 16)
```



The two groups for female turtles and male turtles are pretty well separated for both components.