Exercise 6

Tobias Raidl, 11717659

2023-11-30

Contents

1 Principal Component Analysis

```
load("darwinM.RData")
df = darwinM
```

a

Perform classical PCA on the whole data (except "class"), and try to explain based on the biplot the main differences between the patient and the healthy group.

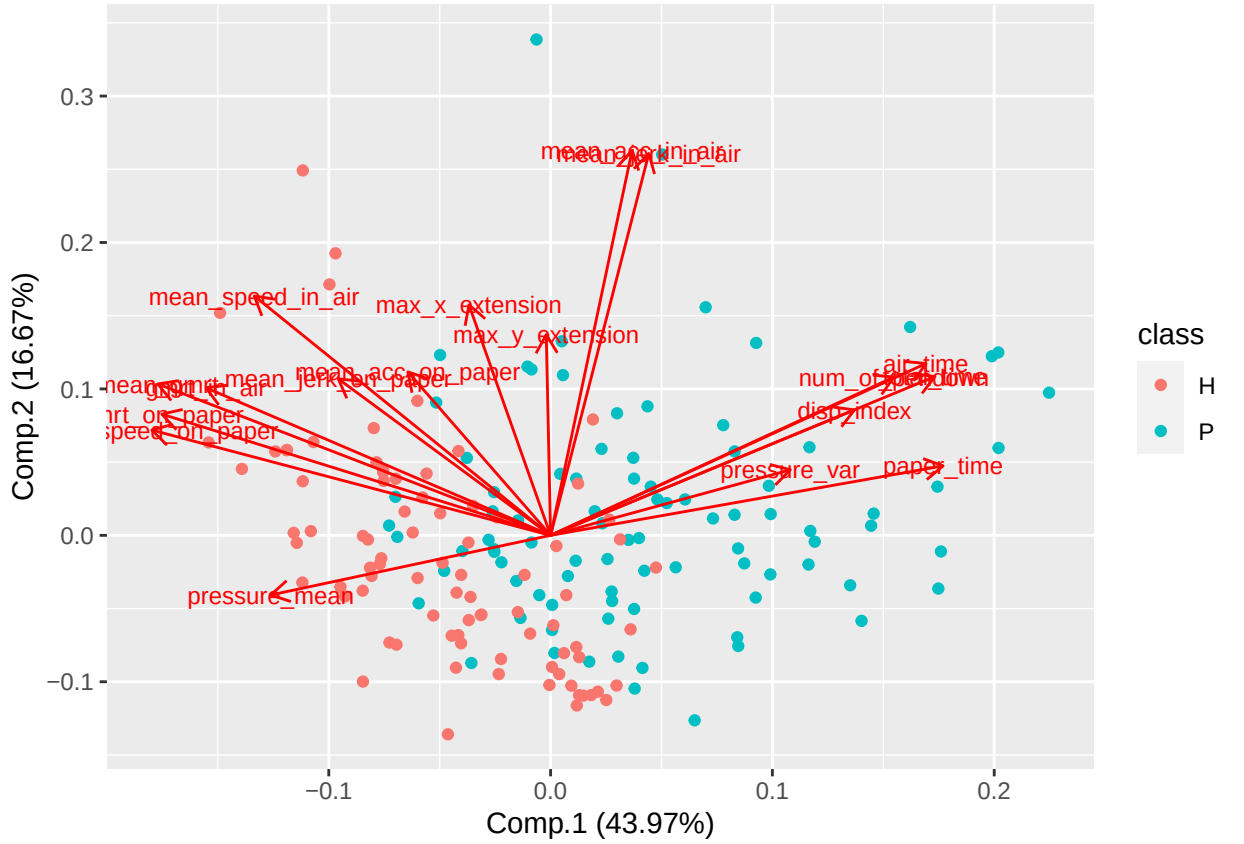
```
library(ggplot2)
```

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.3.2 erstellt
```

```
library(ggfortify)
```

```
## Warning: Paket 'ggfortify' wurde unter R Version 4.3.2 erstellt
```

```
set.seed(11717659)
pca = princomp(~. - class, df, cor = TRUE)
autoplot(pca, data = df, colour = "class", loadings = TRUE, loadings.colour = "red",
    loadings.label = TRUE, loadings.label.size = 3)
```



The pressure mean is higher for healthy people, while the pressure variance is higher patients. Both air and paper time are higher for patients while mean speed in air and on paper both tend to be higher for healthy people.

b

For the following tasks, use the function PcaHubert() from the package rrcov, which performs a robust PCA. Apply PCA separately on the healthy and on the patient group, and show the PCA diagnostics plots with orthogonal and score distances. What could be the reason for the clear outlyingness of some observations?

```
library(rrcov)
```

```
## Warning: Paket 'rrcov' wurde unter R Version 4.3.2 erstellt

## Lade nötiges Paket: robustbase

## Warning: Paket 'robustbase' wurde unter R Version 4.3.2 erstellt

## Scalable Robust Estimators with High Breakdown Point (version 1.7-4)
```

```r
library(dplyr)
```

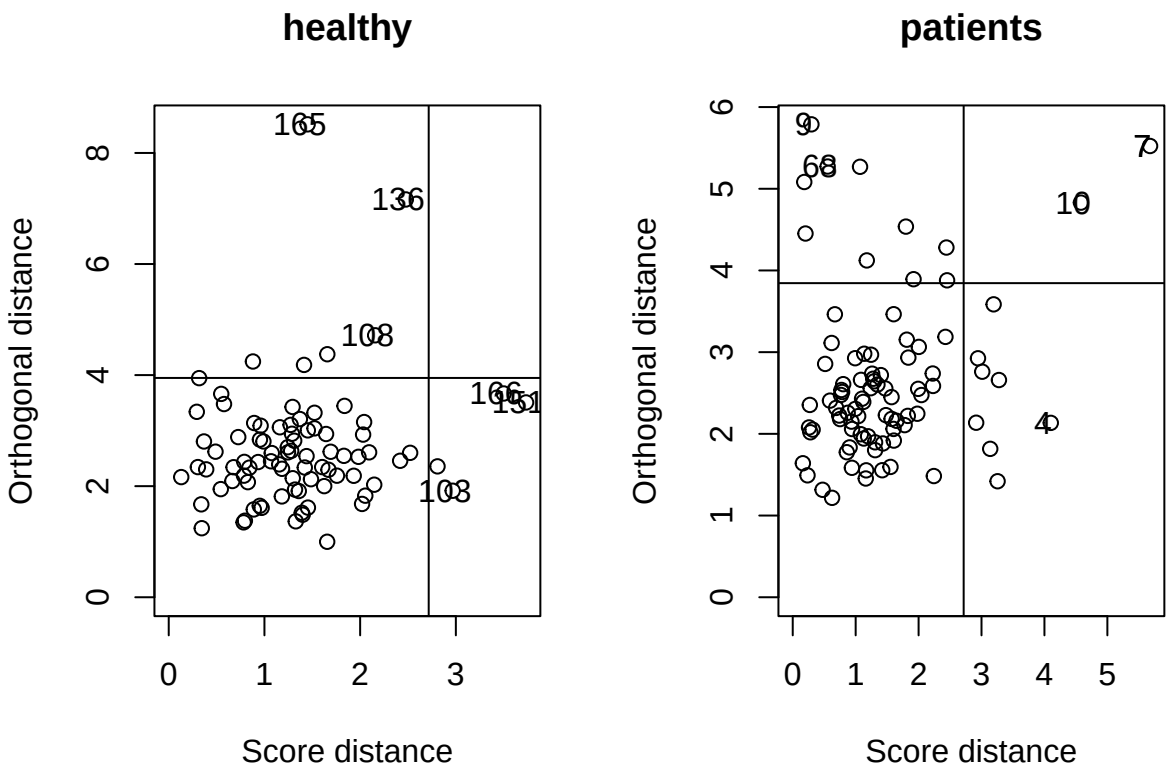
```
## Warning: Paket 'dplyr' wurde unter R Version 4.3.2 erstellt
```

```
##
## Attache Paket: 'dplyr'
```

```
## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

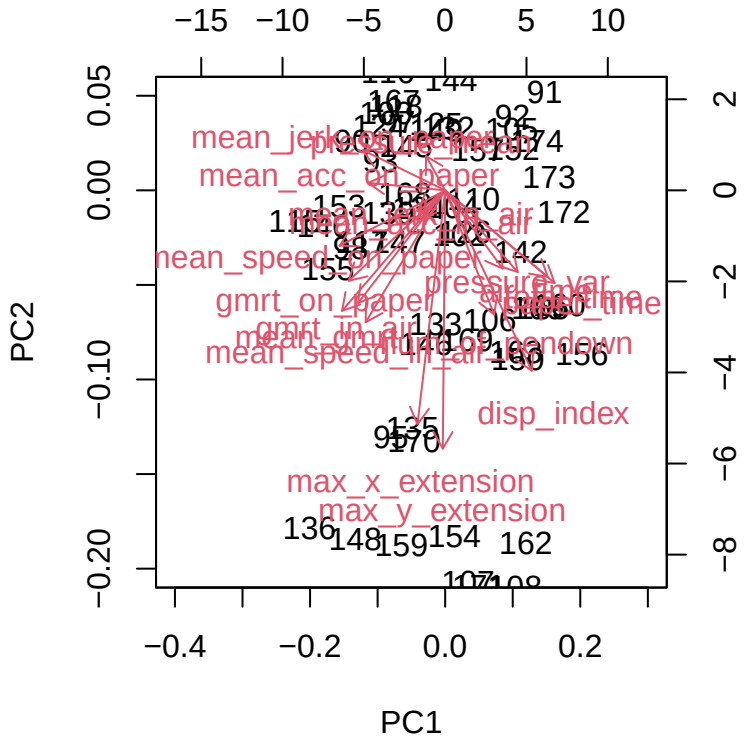
```r
pca_hubert = PcaHubert(~. - class, df, k = 2, cor = TRUE)
df_h = select(df[df$class == "H", ], -class)
df_p = select(df[df$class == "P", ], -class)

pca_hubert_h = PcaHubert(~., df_h, k = 2, scale = TRUE)
pca_hubert_p = PcaHubert(~., df_p, k = 2, scale = TRUE)

par(mfrow = c(1, 2))
plot(pca_hubert_h, main = "healthy")
plot(pca_hubert_p, main = "patients")
```



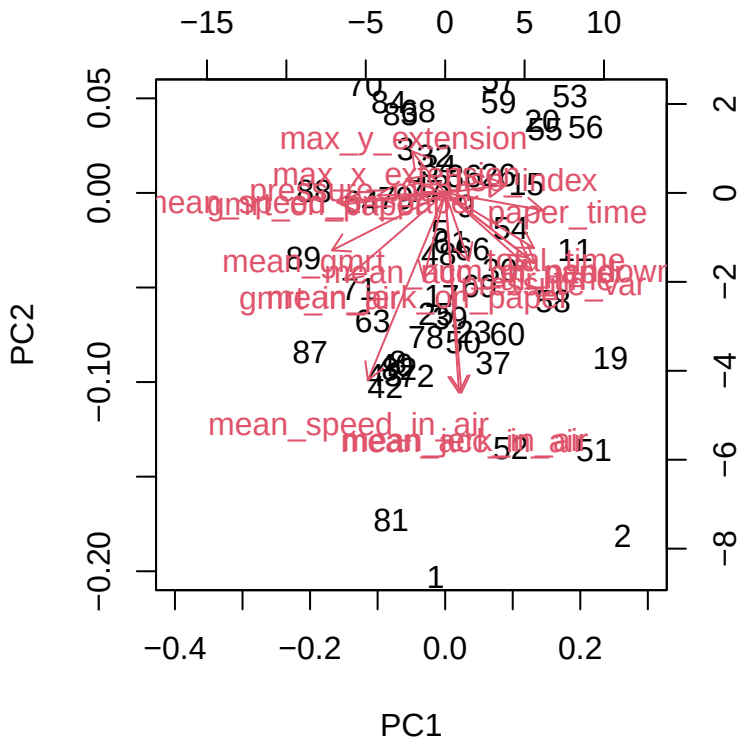
c

Present both biplots and try to identify main differences in the data structure of both groups.

```
biplot(pca_hubert_h, xlim = c(-0.4, 0.3), ylim = c(-0.2, 0.05))
```




```
biplot(pca_hubert_p, xlim = c(-0.4, 0.3), ylim = c(-0.2, 0.05))
```



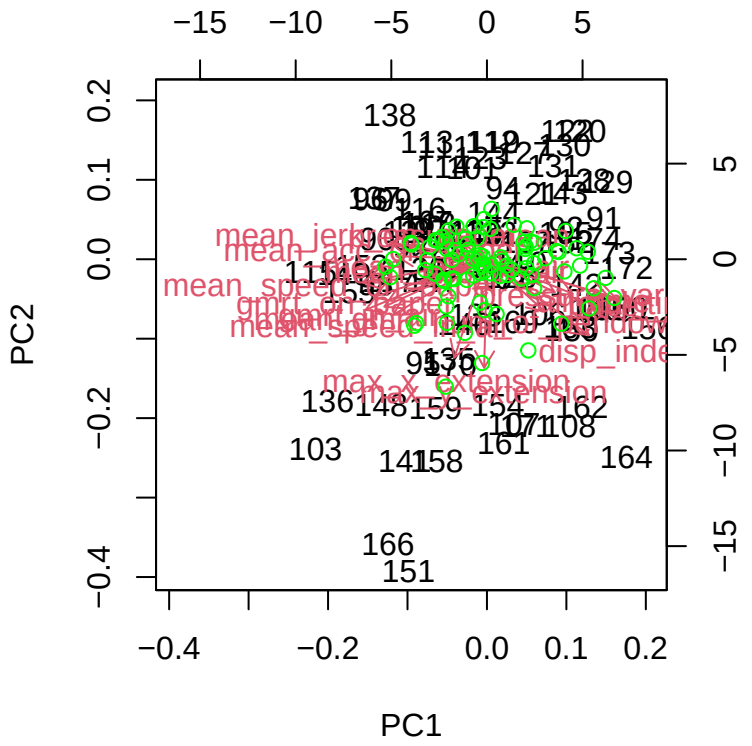
For the patient group the loadings point in directions within circa 180 degrees, whereas the loadings from the healthy group have a higher variance in direction.

Ask Prof. Filzmoser: Cannot see much here. Any better way for visualization?

d

Take the PCA result for the healthy group, and project the patient data into the plane of the first 2 healthy group PCs. This allows for a better comparison of the main differences in the data structure of the two data sets. What are these main differences?

```
biplot(pca_hubert_h)
points(pca_hubert_p$scores, col = "green")
```



Sadly I cannot she anything in this plot.

Ask Prof. Filzmoser: Any better visualization method?

2

a

Compute a factor analysis model for the complete data (without "class") using an appropriate number k of factors.

```
fa <- factanal(scale(select(df, -class)), factors = 8, scores = "regression")
```

b

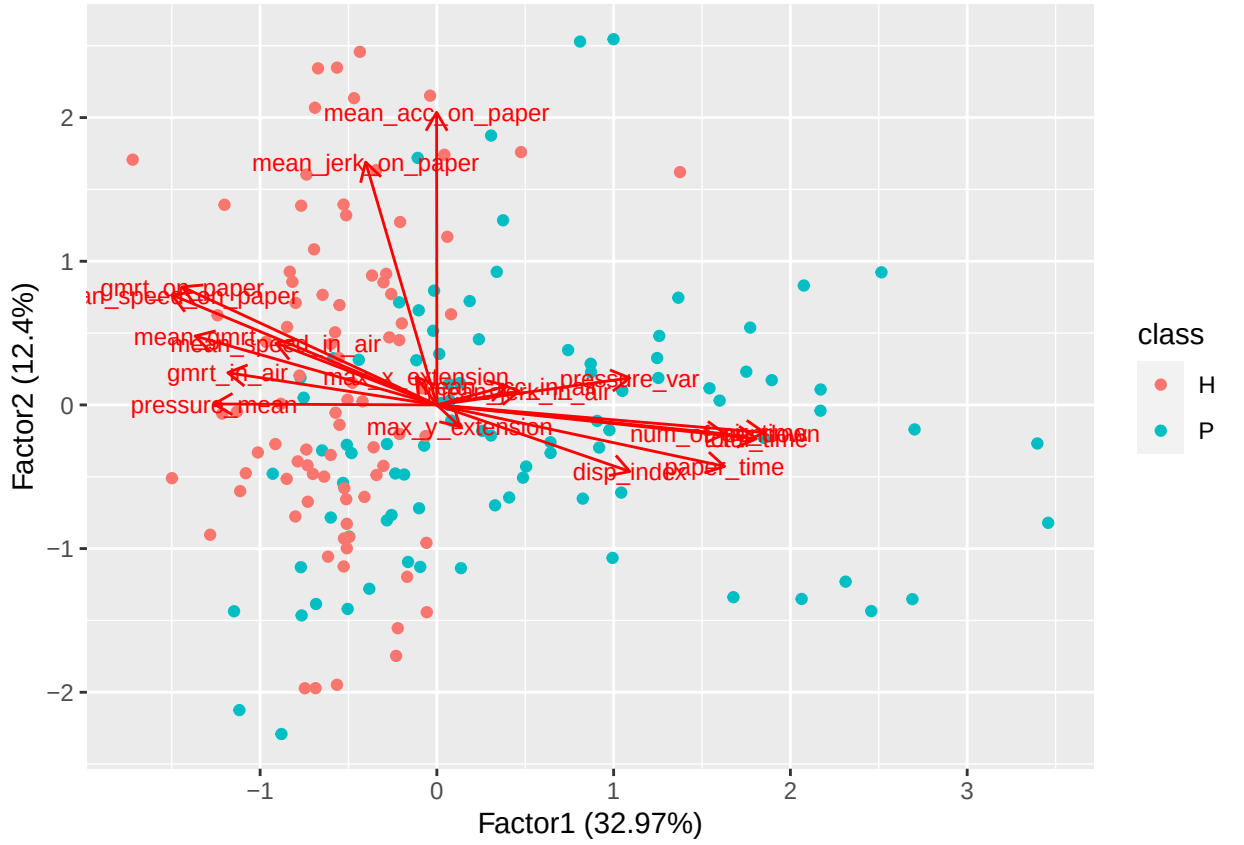
What is the maximum value of k you could consider? k=8 is the largest k to consider. (formula in script)

c

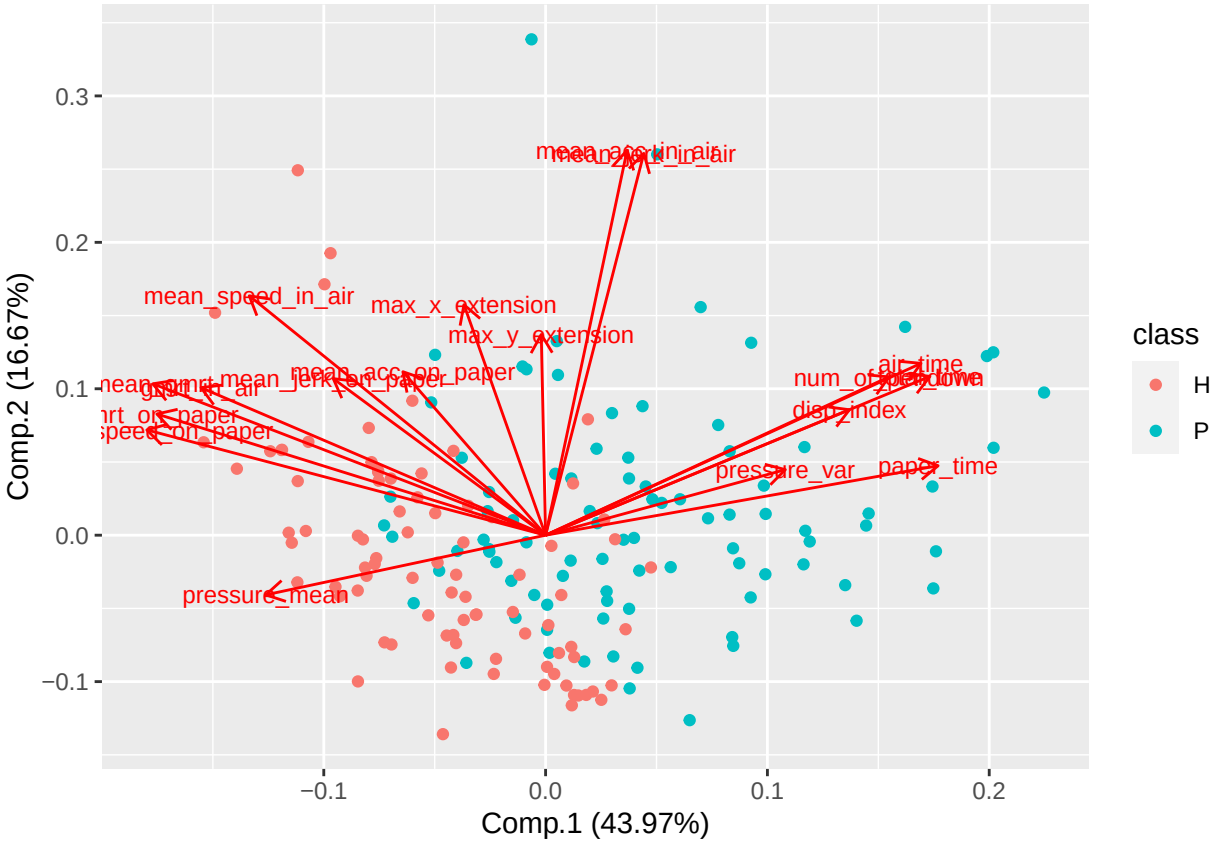
What are the differences to PCA (loadings, scores)? What are the uniquenesses? The uniquenesses are the error term e. There is one uniqueness value per variable. The higher it is the less variance of it is explained by the factors.

d Show loadings and scores in a biplot and compare with the PCA biplot. How can you interpret the first two factors? The first factor focuses on all the horizontally displayed loading vectors like paper_time and pressure_mean. The second one on the mean_acc_on_paper and mean_jerk_on_paper. Variables whose loading vectors are rather short are hardly covered by the first 2 factors (max_y_extension, max_x_extension). They seem to be less relevant than others.

```
autoplot(fa, data = df, colour = "class", loadings = TRUE, loadings.colour = "red",
    loadings.label = TRUE, loadings.label.size = 3)
```




```
autoplot(pca, data = df, colour = "class", loadings = TRUE, loadings.colour = "red",
    loadings.label = TRUE, loadings.label.size = 3)
```


Comp.2 (16.67%)
Comp.1 (43.97%)
0.3
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
mean_acc_in_air
mean_speed_in_air
max_x_extension
max_y_extension
mean_acc_on_paper
mean_jerk_on_paper
num_of_pendown
air_time
disp_index
pressure_var
paper_time
pressure_mean
class
H
P