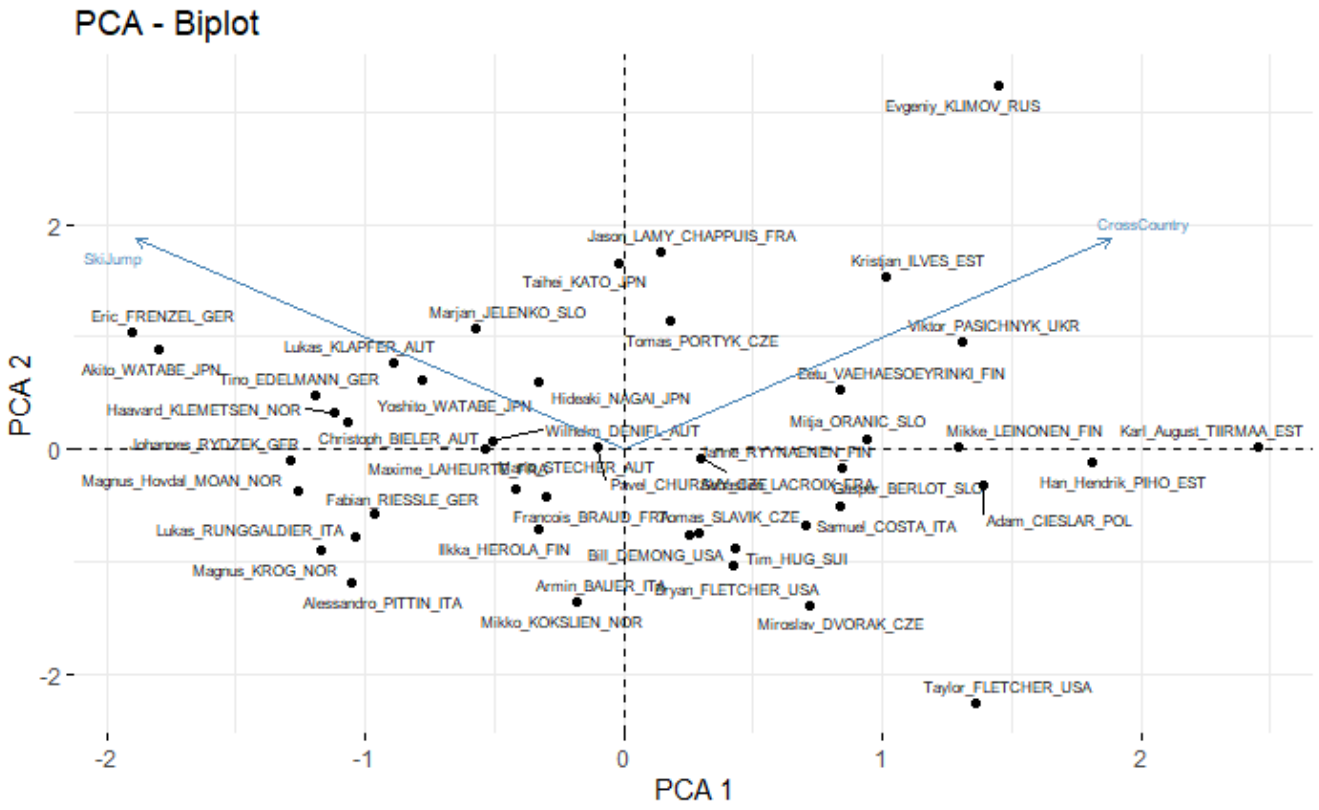
## Assignment 3

- All code found in appendix

A) Nordic Combination (max. 4 marks)

The file Nordic.txt contains the result of the Sochi 2014 Nordic Combined 10k/Normal Hill event. The competition is decided by who performs the best in a combination of ski jumping and cross-country skiing. The variable SkiJump is the ski jump score and CrossCountry is the cross-country time in seconds Source: http://www.sochi2014.com/en/nordic-combined-ind-gund-nh-10-km-cross-c-free-race

1. Perform the principal component analysis on the correlation matrix.




2. One way of combining the scores is to use the first principal component. Why might this be a good idea?
   - The first principal component PC1 best accounts for the shape of the point swarm and it explains the greatest amount of variance in the original features meaning that we preserve the greatest amount of information of our dataset

3. If the competitors were ranked based on the first principal component, who would have won the bronze medal?

   The Top three based on PCA1 is

   1st: Eric frenzel

   2nd: Akido Watabe

   3rd: Johannes Rydzek

   The bronze medal goes to Johannes Rydzek

4. What do you think the second principal component represents?
   - The second principal component (PC2) is oriented such that it reflects the second largest source of variation in the data while being orthogonal to the first Principal component in this case it reflects just under 50% of the variance. PC2 also passes through the average point. Specifically for our data the loadings show that with both our values of PCA2 being -0.7071068 this implies on our graph the more negative the value the lower ski jump score and cross country time.

| PCA1 | PCA2 |
|---|---|
| Ski jump -7071 | ski jump - 0.7071 |
| cross country 0.7071 | cross country -0.7071 |

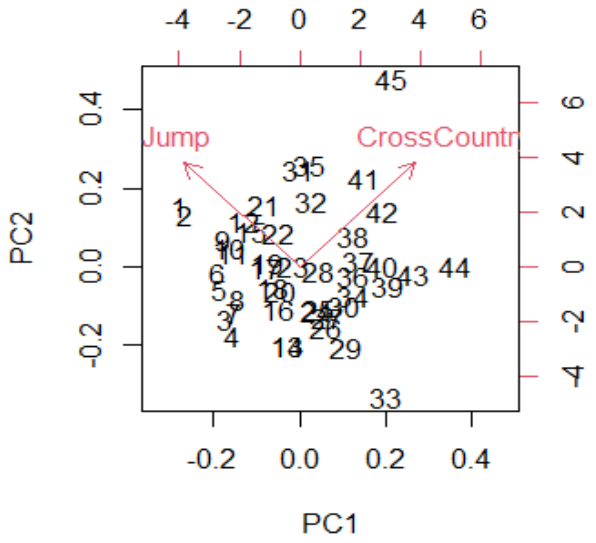
5. Are the data adequately summarized by one principal component?
   - No as one PCA1 only summarizes around 50.5% of the proportional variance and we cant make many correct inferences based on this.

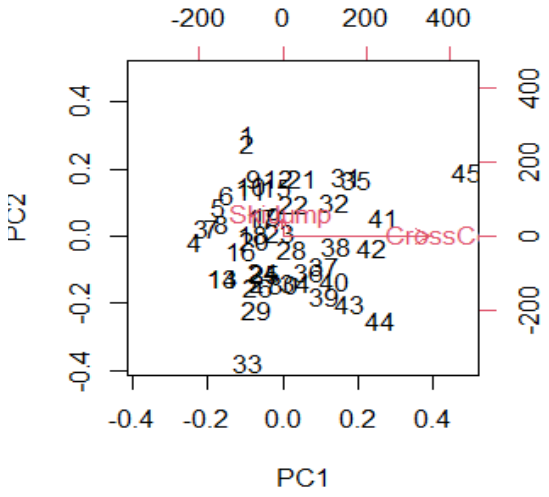
6. The IOC wants to introduce a new snowmobile half-pipe event and is considering dropping the Nordic combined on the grounds that ability in cross-country skiing and ski jumping are more or less equivalent. Do you think this is reasonable in terms of correlation?
    - I think this is an unreasonable assumption to be made as the correlation between the two variables is -0.01059985 this is an incredibly weak negative relationship meaning the two variables are almost not linearly related at all so they are in fact not more of less equivalent

7. Would it be better to run a PCA on the covariance matrix instead of the correlation matrix in this example? Who would be the gold medal list in that case (first PCA component on the covariance matrix mean?

    - Using the correlation matrix is equivalent to standardizing each of the variables (to mean 0 and standard deviation 1) whereas covariance requires scaling if we want to standardize our results assuming that we don't manually scale then the Correlation matrix is better as its less work required and has better readability overall along with the main factor being because we have two different unit measurements and without scaling, the weighting on values with a high standard deviation will skew the axis leading to incorrect information in the end. This is evident in the change of gold medalists from Eric frenzel to Alessandro PITTIN because Cross country time being larger has a higher impact on the final results
    - The gold medalist is the person with the most negative PC1 value for covariance matrix: Alessandro PITTIN

This is evident in the below graphs correlation to the left and covariance on the right



## B) Police Applicants (max. 3 marks)

To obtain a simplified rating scheme of police applicants, the variables should be categorised into groups that characterise different aspects of the applicants' abilities.
Perform Factor Analysis to allocate the variables into several groups

1. How many factors can be found? (using hypotheses testing with p <= 0:05)

    - There are 5 factors that can be found from this data set.
      These are all displayed in appendix

2. Which variables are grouped by the first two factors? (e.g. threshold loading >= 0.5)
    - We can see that for factor 1 "WEIGHT","THIGH" and "FAT" are grouped together and for factor 2 "HEIGHT","WEIGHT","SHLDR","PELVIC" and "BREATH" are grouped together.

3. To reduce the time and effort of obtaining so many variables, we would rather not measure the
   diastolic blood pressure. Just measuring the resting pulse rate should be sufficient. Do you agree? (Why or why not...)

    - I disagree that just reading pulse will be sufficient as there is no factor that contains both pulse rate and diastolic blood pressure. As factor analysis groups correlated variables together based on common variance if they are both within the same factor it implies, they share an underlying relationship in this case they do not. therefore, just measuring resting pulse rate is not sufficient.

4. When we want to separate huge athletic applicants from huge non-athletic applicants, which factor scores can be used?

    - We can use factor 1 and factor 2 to determine which applicants are considered huge based on variables such as weights heights shoulder and pelvic size and then we can cross reference with applicants in factor 3 to determine if they are athletic or not based on RECVR and speed variables.

## C) Singular Value Decomposition (max. 3 marks)

### 1) Compute $A^TA$

$$A^TA \begin{bmatrix} 10 & 3 & 4 \\ 10 & 0 & -5 \end{bmatrix} \begin{bmatrix} 10 & 10 \\ 3 & 0 \\ 4 & -5 \end{bmatrix} = \begin{bmatrix} (10\cdot10)+(3\cdot3)+(4\cdot4) & (10\cdot10)+(3\cdot0)+(4\cdot(-5)) \\ (10\cdot10)+(0\cdot3)+(-5)\times4 & (10\cdot10)+0+(-25) \end{bmatrix}$$

$$= \begin{bmatrix} 125 & 80 \\ 80 & 125 \end{bmatrix} = B$$

### 2) eigenvalue

$$\begin{bmatrix} 125-\lambda & 80 \\ 80 & 125-\lambda \end{bmatrix} = 0 \qquad (125-\lambda)\times(125-\lambda) - (80\times80) = 0$$

$$15625 - 125\lambda - 125\lambda - \lambda^2 - 6400 = 0$$

$$\lambda^2 - 250\lambda + 9225 = 0$$

$$\frac{-(-250) \pm \sqrt{(-250)^2 - 4(1)(9225)}}{2(1)} \qquad \lambda_1 = 205 \qquad \lambda_2 = 45$$

205 45

$$\begin{bmatrix} 205 \\ 45 \end{bmatrix}$$

### eigenvectors $B-\lambda I$ for $\lambda = 205$

$$\begin{bmatrix} 125-\lambda & 80 \\ 80 & 125-\lambda \end{bmatrix} = \begin{bmatrix} 125-205 & 80 \\ 80 & 125-205 \end{bmatrix} = \begin{bmatrix} -80 & 80 \\ 80 & -80 \end{bmatrix}$$ RREF

$$\begin{bmatrix} -80 & 80 \\ 80 & -80 \end{bmatrix} R_1 \leftarrow -80R_1 \quad \begin{bmatrix} 1 & -1 \\ 80 & -80 \end{bmatrix} R_2 \leftarrow R_2 - 80R_1 \quad \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = 0 \qquad \lambda_2 \text{ free variable} = t$$

$$\lambda_1 - t = 0 \quad \lambda_1 = t \quad \Rightarrow V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} t$$ → convert to unit vector later.

$$\|V\| = \sqrt{1^2 + 1^2} = \sqrt{2}$$

$$\text{Null} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

### for $\lambda = 45$

$$\begin{bmatrix} 125-45 & 80 \\ 80 & 125-45 \end{bmatrix} = \begin{bmatrix} 80 & 80 \\ 80 & 80 \end{bmatrix} R_1 \leftarrow \frac{1}{80}R_1 \quad \begin{bmatrix} 1 & 1 \\ 80 & 80 \end{bmatrix} R_2 \leftarrow R_2 - 80R_1$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = 0$$

$\lambda_2$ free variable $= t$

$\lambda_1 + t = 0 \quad \lambda_1 = -t \quad \Rightarrow V_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} t$

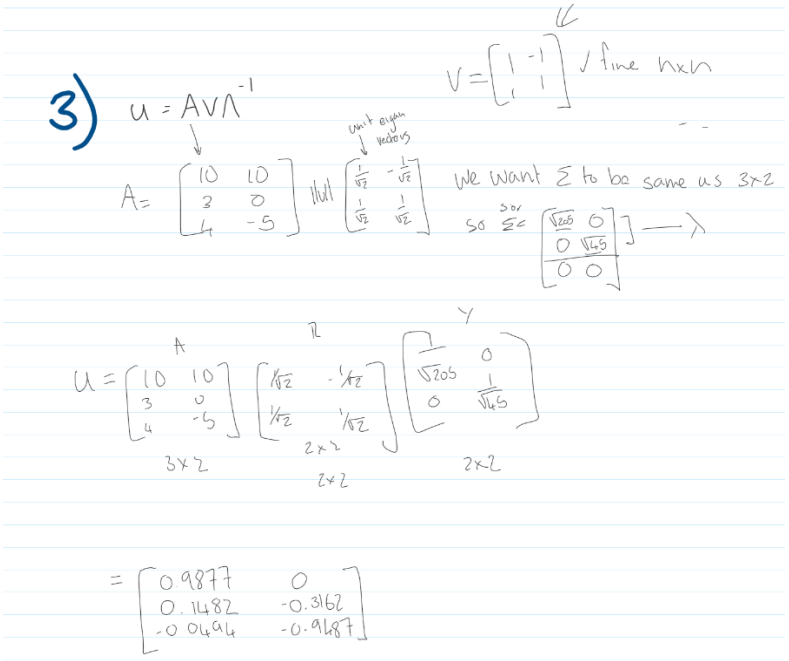
$$\|V\| = \sqrt{(-1)^2 + 1^2}$$

### eigen vectors

$$\left\{45 \begin{bmatrix} -1 \\ 1 \end{bmatrix} t\right\}, \left\{205 \begin{bmatrix} 1 \\ 1 \end{bmatrix} t\right\}$$

$$\text{Null} = \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

The Eigan values are 45 and 205 with normalized vectors as [-0.7071 0.7071] [0.7071  0.7071]




U =

| | |
|---|---|
| 0.9877 | 0.0000 |
| 0.1482 | 0.3162 |
| -0.0494 | 0.9487 |

4. What does the the eigenvectors V and eigenvalues $\Lambda$ tell you about the variation in the 3 observation in X (think about the principal components).

The total variance is the sum of the variances of all individual principal components

The first eigenvector displays the greatest variance in the dataset and the direction. The second line eigenvector is the direction of the second greatest variance. The second greatest variance also explains all the remaining variance. The Eigenvalues are the total amount of variance in the data set explained by the common factors. Proportionally the first eigenvalue explains 205/250 % of variance and the second eigenvalue explains remaining 45/250 %

# Appendix – All code in order

```
#install.packages('tidyverse')
#install.packages("tidyr")
#install.packages('factoextra')
#install.packages('FactoMineR')
#install.packages('dplyr')

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library("tidyr")
library("factoextra")

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

library("FactoMineR")
library("dplyr")

#read all the files in to dataframe for easy access
header<-read.table('Nordic.txt',nrows = 1, header = FALSE, stringsAsFactors =
FALSE)
nordic<-read.table('Nordic.txt', skip = 1, header = FALSE)
colnames(nordic)<-unlist(header)
nor<-unite(nordic,"first_Name_Nat",first,Name,Nat)
```

A)

1)

```
#Completes PCA using an inbuilt function so in looks nice other ways are also
completed in code found at bottom
attach(nor)
final_data2 <- nor[, c(1,2,3)]
```

```
rownames(final_data2) <- nor[,1]
df <- final_data2[, c(2,3)]
res.pca <- PCA(df, graph = FALSE)
fviz_pca_biplot(res.pca, repel = TRUE, labelsize=2, xlab='PCA 1', ylab='PCA 2
')
```

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

```
attach(nor)
```

```
## The following objects are masked from nor (pos = 3):
##
##     CrossCountry, first_Name_Nat, SkiJump
```

```
# mean-adjusted values # we need to do this as we want a s.d. of one and a me
an of zero for PCA #its value - mean / Sd? and we want to scale as they are d
iffernt measures
nor$SkiJump_adj = (nor$SkiJump - mean(nor$SkiJump))/sd(SkiJump)
nor$CrossCountry_adj = (nor$CrossCountry - mean(nor$CrossCountry))/sd(CrossCo
untry)
```

```
# calculate correlation matrix and eigenvectors/values
(cm = cor(nor[,2:3]))         #as we are using correlation matrix this is equ
ilivant of standardizing each of the variables to mean of 0 and sd of 1
```

```
##                  SkiJump CrossCountry
## SkiJump       1.00000000  -0.01059985
## CrossCountry -0.01059985   1.00000000
```

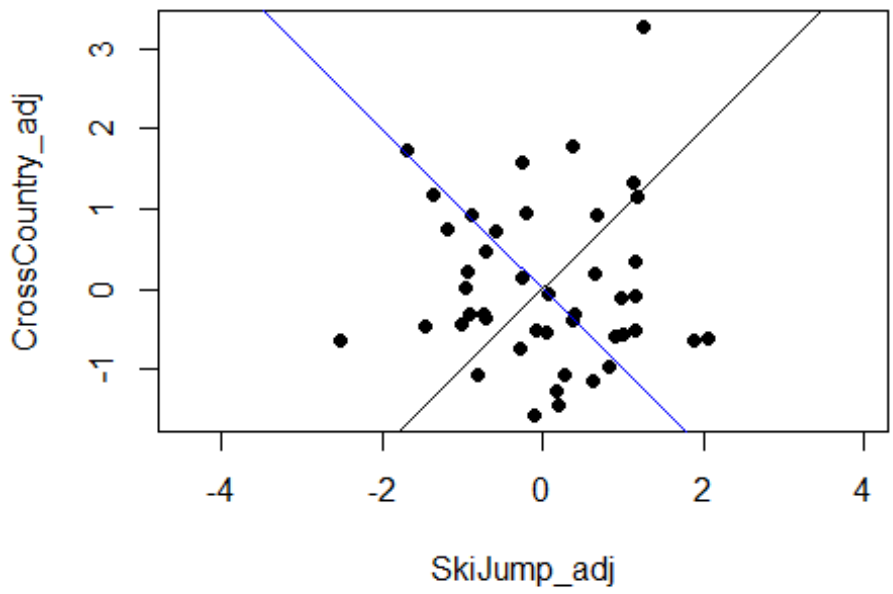
```
#find the eigan directions
(e = eigen(cm))
```

```
## eigen() decomposition
## $values
## [1] 1.0105999 0.9894001
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.7071068
## [2,]  0.7071068 -0.7071068
```

```
s1 = e$vectors[2,1] / e$vectors[1,1] # PC1
s2 = e$vectors[2,2] / e$vectors[1,2] # PC2
```

```
#pca1 -variable 1 skijump is negatively correlated and crosscountry time is p
ositive correlated
#this means that for PCA the lower its value the better(higer) skijump score
they got and lower cross country time
```

```
plot(nor$SkiJump_adj, nor$CrossCountry_adj, asp=T, pch=16, xlab='SkiJump_adj'
, ylab='CrossCountry_adj')
abline(a=0, b=s1, col='blue') #pca 1st dimension
abline(a=0, b=s2) #pca 2nd dimension
```

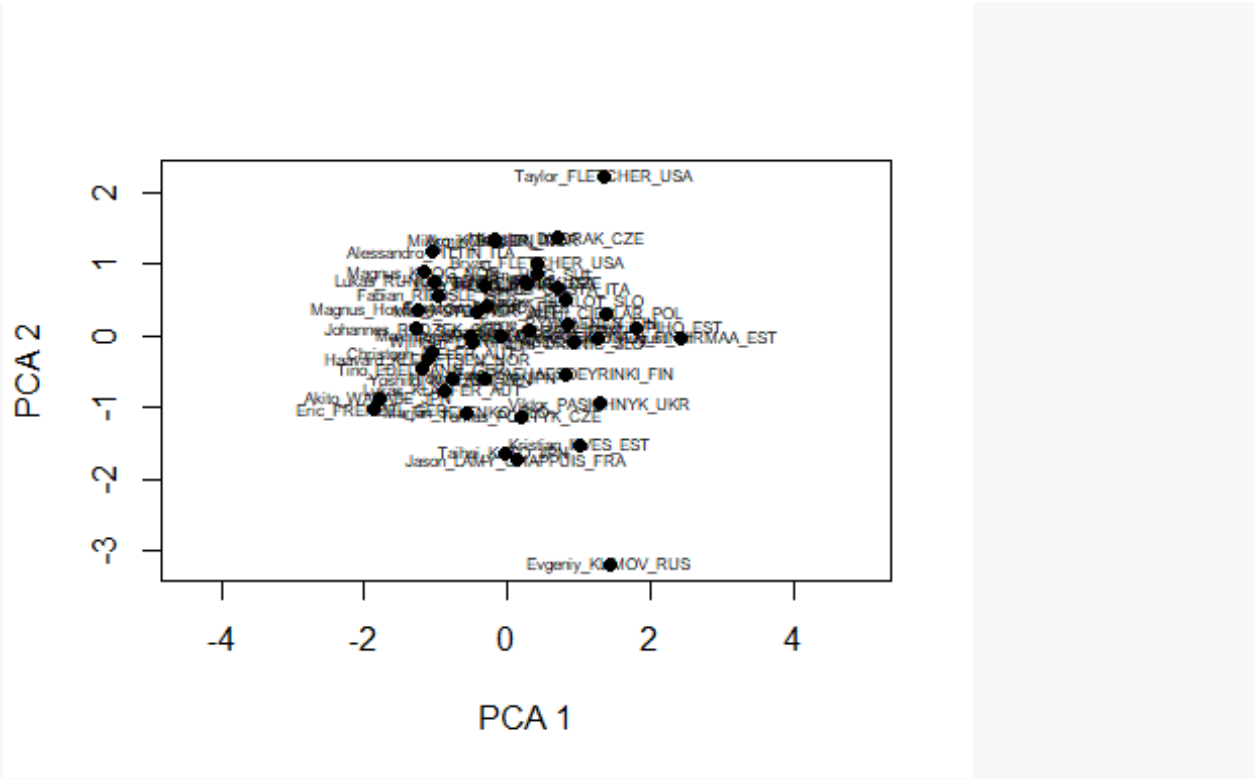



```
#as this is currently based on Correlation matrix why is the mean and SD not
0 and 1 respectivly

# PCA data = rowFeatureVector (transposed eigenvectors) * RowDataAdjust (mean
adjusted, also transposed)
feat_vec = t(e$vectors)
row_data_adj = t(nor[,4:5])
final_data = data.frame(t(feat_vec %*% row_data_adj))
names(final_data) = c('PCA1','PCA2')


#final_data

plot(final_data, asp=T, xlab='PCA 1', ylab='PCA 2', pch=16)
text(final_data, labels=c(nor$first_Name_Nat), cex=0.5)
```




7)

```
hep.PC.cor = prcomp(nor[,2:3], scale=TRUE) #correlation this means that each value is standardized i.e mean 0 sd 1 so we set Scale=True
hep.PC.cov = prcomp(nor[,2:3], scale=FALSE) #covaraince w/0 standardization so Scale=False
biplot(hep.PC.cov) #covariance #both preform centering

hep.PC.cov$x
```

```
##                PC1         PC2
##  [1,]  -39.790624  17.2470549
##  [2,]  -41.588542  15.7445600
##  [3,]  -94.668798   1.4709315
##  [4,] -102.465462  -0.9398836
##  [5,]  -75.073809   5.0981105
##  [6,]  -63.476317   6.9141963
##  [7,]  -83.068670   1.3870192
##  [8,]  -70.369792   2.2046315
##  [9,]  -32.780229   9.7567701
## [10,]  -36.578423   8.4515013
## [11,]  -37.577312   7.6501152
## [12,]   -5.480256   9.7946313
## [13,]  -70.457033  -6.9954983
## [14,]  -70.257033  -6.9952210
## [15,]   -7.678035   8.1915818
## [16,]  -48.063434  -2.3644372
```

```
## [17,]  -24.570945   3.0681488
## [18,]  -35.167191   0.3534507
## [19,]  -19.571366   3.3750828
## [20,]  -32.665668  -0.7430811
## [21,]   23.219855   9.7344342
## [22,]   12.125830   5.4190442
## [23,]   -3.867637   0.6968590
## [24,]  -23.158326  -6.0299009
## [25,]  -20.558190  -6.1262949
## [26,]  -28.254716  -8.6369713
## [27,]  -20.555971  -7.7262934
## [28,]   10.236371  -2.1835835
## [29,]  -30.349444 -12.4398801
## [30,]    1.244563  -8.0960595
## [31,]   75.019528  10.0062732
## [32,]   60.225506   5.6857519
## [33,]  -41.037090 -21.3547109
## [34,]   14.944272  -7.8770598
## [35,]   86.720072   9.6224998
## [36,]   31.041622  -5.9547332
## [37,]   47.440219  -4.9319897
## [38,]   62.735766  -1.7107739
## [39,]   48.747291 -10.0301819
## [40,]   60.343674  -7.4140969
## [41,]  116.828919   3.2642504
## [42,]  103.536143  -1.9541898
## [43,]   77.449066 -11.2903778
## [44,]  113.253053 -14.1407256
## [45,]  214.018562  10.7990458

biplot(hep.PC.cor)  #correlation #best one

nor2 <- cbind(nor, hep.PC.cov$x)
head(arrange(nor2, (PC1)), 3) #want the most negative PCA1 for gold medal

##          first_Name_Nat SkiJump CrossCountry        PC1        PC2
## 1 Alessandro_PITTIN_ITA   113.4       1367.5 -102.46546 -0.9398836
## 2       Magnus_KROG_NOR   115.8       1375.3  -94.66880  1.4709315
## 3 Lukas_RUNGGALDIER_ITA   115.7       1386.9  -83.06867  1.3870192
```

Question 2:

To obtain a simplified rating scheme of police applicants, the variables should be categorised into groups that characterise different aspects of the applicants abilities.

Perform Factor Analysis to allocate the variables into several groups:

```
FA <- read.csv("Police.csv", header=TRUE) #Input the dataset into R
sFA <- scale(FA, center=TRUE, scale=TRUE) # then center and scale the factors
.
```

1)

```
round(sapply(1:9, function(i) factanal(sFA, factors=i)$PVAL), 3) < 0.05   # T
hese are the P values we want to keep all below 0.05 and then 1 that is just
above

## objective objective objective objective objective objective objective obje
ctive
##      TRUE      TRUE      TRUE      TRUE     FALSE     FALSE     FALSE
FALSE
## objective
##     FALSE
```

```
2) fa <- factanal(sFA, factors = 5) #scaled used for factor analysis #3 facto
rs
apply(fa$loadings[,c(1,2,3,4,5)] >= 0.5 , 2, function(x) names(FA)[x]) #keeps
fist three factors greater than 0.5

## $Factor1
## [1] "WEIGHT" "THIGH"  "FAT"
##
## $Factor2
## [1] "HEIGHT" "WEIGHT" "SHLDR"  "PELVIC" "BREATH"
##
## $Factor3
## [1] "PULSE" "RECVR"
##
## $Factor4
## [1] "CHEST"
##
## $Factor5
## [1] "REACT"
```

```
print(fa$loadings, cutoff= 0.5)

##
## Loadings:
##        Factor1 Factor2 Factor3 Factor4 Factor5
## REACT                                   0.782
## HEIGHT          0.888
## WEIGHT  0.614   0.615
## SHLDR           0.747
## PELVIC          0.585
## CHEST                           0.666
## THIGH   0.957
## PULSE                   0.575
## DIAST
```

```
## CHNUP  -0.690
## BREATH          0.598
## RECVR                   0.948
## SPEED                  -0.534
## ENDUR
## FAT     0.895
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings      3.149   2.843   1.760   0.948   0.905
## Proportion Var   0.210   0.190   0.117   0.063   0.060
## Cumulative Var   0.210   0.399   0.517   0.580   0.640
```

3)