Nordic Combination - PCA

1. One way of combining the scores is to use the first principal component. Why might this be a good idea?

This is a good idea as the first principal component represents if the athletes perform well in the two sports, or if they perform badly in one sport and better in the other sport. Overall, it gives a general sense of the variation in the data as first principal component accounts for the highest variation.

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

```
nordic <- read.csv("Nordic.csv")
head(nordic)
```

```
##   Nat         first    Name SkiJump CrossCountry
## 1 GER          Eric FRENZEL   131.5       1430.2
## 2 JPN         Akito  WATABE   130.0       1428.4
## 3 NOR        Magnus    KROG   115.8       1375.3
## 4 ITA    Alessandro  PITTIN   113.4       1367.5
## 5 NOR Magnus_Hovdal    MOAN   119.4       1394.9
## 6 GER      Johannes  RYDZEK   121.2       1406.5
```

```
mat <- as.matrix(nordic[, -c(1:3)])
rownames(mat) <- nordic[,3]

pca <- prcomp(mat, center = TRUE, scale = TRUE)

pca$rotation
```

```
##                     PC1       PC2
## SkiJump      -0.7071068 0.7071068
## CrossCountry  0.7071068 0.7071068
```

```
pca$x
```

```
##                         PC1          PC2
## FRENZEL         -1.88249626  1.023188804
## WATABE          -1.77599130  0.877788113
## KROG            -1.15734985 -0.888277292
## PITTIN          -1.04009940 -1.174075891
## MOAN            -1.24787101 -0.374224881
## RYDZEK          -1.27368372 -0.097750812
## RUNGGALDIER     -1.02362232 -0.771343471
## RIESSLE         -0.95358168 -0.566953147
## EDELMANN        -1.17710155  0.469055251
## KLEMETSEN       -1.10899903  0.318839533
## BIELER          -1.05262856  0.240860323
## KLAPFER         -0.88214228  0.764014516
## KOKSLIEN        -0.18215144 -1.340544255
```

```
## BAUER            -0.17999057 -1.338383381
## WATABE           -0.77156221  0.605895223
## HEROLA           -0.32638892 -0.712271060
## LAHEURTE         -0.52591643 -0.004938219
## STECHER          -0.41372765 -0.346179616
## DENIFL           -0.49708515  0.074274190
## BRAUD            -0.29435132 -0.411534100
## JELENKO          -0.56366004  1.065703043
## NAGAI            -0.32252377  0.584709780
## CHURAVY          -0.10074147  0.017187690
## DEMONG            0.25332350 -0.753925919
## SLAVIK            0.28981172 -0.734231415
## FLETCHER          0.41653945 -1.027346432
## HUG               0.42416140 -0.868581097
## LACROIX           0.29510892 -0.073979509
## DVORAK            0.71293078 -1.369116101
## COSTA             0.69328406 -0.666633280
## KATO             -0.02078745  1.642163053
## PORTYK            0.18037266  1.121193626
## FLETCHER          1.34464414 -2.232042952
## BERLOT            0.82451020 -0.501819718
## LAMY_CHAPPUIS     0.13921108  1.734986746
## RYYNAENEN         0.83892029 -0.168329135
## ORANIC            0.93214338  0.092831063
## VAEHAESOEYRINKI   0.82875086  0.526837268
## CIESLAR           1.37442867 -0.321362869
## LEINONEN          1.28144112  0.022286042
## ILVES             1.00182131  1.522799516
## PASICHNYK         1.29475967  0.942464946
## PIHO              1.79367317 -0.120436603
## TIIRMAA           2.42397837  0.022850998
## KLIMOV            1.43063867  3.194351434
```

The first principal component represents athletes who scored poorly as they covered less distance in their jump and took more time to finish cross country. According to the first PCA, Evgeniy KLIMOV would get the bronze medal as he has the third highest score which is 1.43063867.

3. What do you think the second principal component represents?

```
pca$rotation
```

```
##                     PC1       PC2
## SkiJump      -0.7071068 0.7071068
## CrossCountry  0.7071068 0.7071068
```

The second principal component shows that both scores have the same value and are positive. It represents the athletes that scored higher in ski jump as their jump distance is large, but lower in cross country as they took more time.

4. Are the data adequately summarized by one principal component?

```
pca <- prcomp(mat, center = TRUE, scale = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2
## Standard deviation     1.0053 0.9947
## Proportion of Variance 0.5053 0.4947
## Cumulative Proportion  0.5053 1.0000
```

No because the first principal component only explains roughly 51% of variation. We need the second principal component as well as it explains another roughly 49%. Using two principal components explain 100% of total variation within the sample.

5. 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?

Cross-country skiing and ski jumping are equivalent due to their rotations having the same value. This means that both variables can be explained similarly, so it is reasonable to drop them.

6. 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 medallist in that case based on the scores for the first principal component?

It is not better to run PCA on a covariance matrix because the the scaling for SkiJump and CrossCountry are different. SkiJump is measured in metres whereas CrossCountry is measured in seconds. Correlation matrix accounts for the differences in the scaling of SkiJump and CrossCountry thus giving more accurate analysis and results, whereas covariance matrix does not.

```
pca$x
```

```
##                          PC1          PC2
## FRENZEL          -1.88249626  1.023188804
## WATABE           -1.77599130  0.877788113
## KROG             -1.15734985 -0.888277292
## PITTIN           -1.04009940 -1.174075891
## MOAN             -1.24787101 -0.374224881
## RYDZEK           -1.27368372 -0.097750812
## RUNGGALDIER      -1.02362232 -0.771343471
## RIESSLE          -0.95358168 -0.566953147
## EDELMANN         -1.17710155  0.469055251
## KLEMETSEN        -1.10899903  0.318839533
## BIELER           -1.05262856  0.240860323
## KLAPFER          -0.88214228  0.764014516
## KOKSLIEN         -0.18215144 -1.340544255
## BAUER            -0.17999057 -1.338383381
## WATABE           -0.77156221  0.605895223
```

```
## HEROLA           -0.32638892 -0.712271060
## LAHEURTE         -0.52591643 -0.004938219
## STECHER          -0.41372765 -0.346179616
## DENIFL           -0.49708515  0.074274190
## BRAUD            -0.29435132 -0.411534100
## JELENKO          -0.56366004  1.065703043
## NAGAI            -0.32252377  0.584709780
## CHURAVY          -0.10074147  0.017187690
## DEMONG            0.25332350 -0.753925919
## SLAVIK            0.28981172 -0.734231415
## FLETCHER          0.41653945 -1.027346432
## HUG               0.42416140 -0.868581097
## LACROIX           0.29510892 -0.073979509
## DVORAK            0.71293078 -1.369116101
## COSTA             0.69328406 -0.666633280
## KATO             -0.02078745  1.642163053
## PORTYK            0.18037266  1.121193626
## FLETCHER          1.34464414 -2.232042952
## BERLOT            0.82451020 -0.501819718
## LAMY_CHAPPUIS     0.13921108  1.734986746
## RYYNAENEN         0.83892029 -0.168329135
## ORANIC            0.93214338  0.092831063
## VAEHAESOEYRINKI   0.82875086  0.526837268
## CIESLAR           1.37442867 -0.321362869
## LEINONEN          1.28144112  0.022286042
## ILVES             1.00182131  1.522799516
## PASICHNYK         1.29475967  0.942464946
## PIHO              1.79367317 -0.120436603
## TIIRMAA           2.42397837  0.022850998
## KLIMOV            1.43063867  3.194351434
```

Based on first principal component, Karl August TIIRMAA would be the gold medalist as he has the highest rank which is 2.42397837

Police Applicants - Factor Analysis

1. How many factors can be found? (using hypotheses testing with p ≤ 0.05)

```
police <- read.csv("police.csv")
head(police)
```

```
##   REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE DIAST CHNUP BREATH
RECVR
## 1 0.310  179.6  74.20  41.7   27.3  82.4  19.0    64    64     2    158
108
## 2 0.345  175.6  62.04  37.5   29.1  84.1   5.5    88    78    20    166
108
## 3 0.293  166.2  72.96  39.4   26.8  88.1  22.0   100    88     7    167
116
## 4 0.254  173.8  85.92  41.2   27.6  97.6  19.5    64    62     4    220
120
```

```
## 5 0.384  184.8  65.88  39.8   26.1  88.2  14.5    80    68     9    210
120
## 6 0.406  189.1 102.26  43.3   30.1 101.2  22.0    60    68     4    188
91
##   SPEED ENDUR   FAT
## 1   5.5     4 11.91
## 2   5.5     4  3.13
## 3   5.5     4 16.89
## 4   5.5     4 19.59
## 5   5.5     5  7.74
## 6   6.0     4 30.42

fa <- factanal(police, factors=2)
fa

##
## Call:
## factanal(x = police, factors = 2)
##
## Uniquenesses:
##  REACT HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP
BREATH
##  0.957  0.522  0.005  0.429  0.527  0.210  0.205  0.879  0.972  0.494
0.784
##  RECVR  SPEED  ENDUR    FAT
##  0.886  0.890  0.832  0.028
##
## Loadings:
##        Factor1 Factor2
## REACT           0.206
## HEIGHT  0.692
## WEIGHT  0.920   0.384
## SHLDR   0.753
## PELVIC  0.686
## CHEST   0.824   0.334
## THIGH   0.244   0.858
## PULSE  -0.332   0.105
## DIAST  -0.109   0.125
## CHNUP  -0.370  -0.608
## BREATH  0.458
## RECVR  -0.233   0.244
## SPEED          -0.321
## ENDUR  -0.275  -0.304
## FAT     0.533   0.829
##
##                Factor1 Factor2
## SS loadings      3.991   2.389
## Proportion Var   0.266   0.159
## Cumulative Var   0.266   0.425
##
```

```
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 140.55 on 76 degrees of freedom.
## The p-value is 9.66e-06
```

The p-value is less than 0.05 so we need to change the number of factors.

```
fa4 <- factanal(police, factors=4)
fa4

##
## Call:
## factanal(x = police, factors = 4)
##
## Uniquenesses:
##  REACT HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP
BREATH
##  0.883  0.231  0.039  0.318  0.487  0.026  0.146  0.694  0.933  0.484
0.636
##  RECVR  SPEED  ENDUR    FAT
##  0.005  0.657  0.821  0.024
##
## Loadings:
##        Factor1 Factor2 Factor3 Factor4
## REACT   0.199           -0.119  -0.250
## HEIGHT  0.212   0.828   -0.165  -0.106
## WEIGHT  0.667   0.601   -0.212   0.332
## SHLDR   0.158   0.785            0.182
## PELVIC  0.275   0.596   -0.261   0.122
## CHEST   0.569   0.415   -0.184   0.666
## THIGH   0.890            0.169  -0.163
## PULSE           -0.145   0.524
## DIAST           -0.174   0.153
## CHNUP  -0.696  -0.178
## BREATH  0.192   0.569
## RECVR                    0.992
## SPEED  -0.232   0.195   -0.416  -0.279
## ENDUR  -0.364  -0.203
## FAT     0.955   0.209            0.140
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      3.378   2.729   1.685   0.823
## Proportion Var   0.225   0.182   0.112   0.055
## Cumulative Var   0.225   0.407   0.519   0.574
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 74.47 on 51 degrees of freedom.
## The p-value is 0.0177
```

The p-value is still less than 0.05.

```
fa5 <- factanal(police, factors=5)
fa5

##
## Call:
## factanal(x = police, factors = 5)
##
## Uniquenesses:
##  REACT HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP
BREATH
##  0.370  0.109  0.028  0.313  0.485  0.081  0.055  0.621  0.870  0.465
0.587
##  RECVR  SPEED  ENDUR    FAT
##  0.005  0.522  0.826  0.058
##
## Loadings:
##        Factor1 Factor2 Factor3 Factor4 Factor5
## REACT                                   0.782
## HEIGHT  0.176   0.888  -0.164           0.189
## WEIGHT  0.614   0.615  -0.187   0.424
## SHLDR   0.193   0.747  -0.146   0.100  -0.247
## PELVIC  0.238   0.585  -0.272   0.195
## CHEST   0.488   0.458  -0.112   0.666  -0.121
## THIGH   0.957           0.104  -0.117
## PULSE          -0.114   0.575           0.146
## DIAST          -0.166   0.230   0.166   0.142
## CHNUP  -0.690  -0.175          -0.109  -0.124
## BREATH  0.166   0.598                   0.145
## RECVR   0.102           0.948  -0.127  -0.258
## SPEED  -0.191   0.166  -0.534  -0.327  -0.147
## ENDUR  -0.354  -0.198
## FAT     0.895   0.245           0.273
##
##                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
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 53.8 on 40 degrees of freedom.
## The p-value is 0.0712
```

By adding another factor, the p-value is now greater than 0.05. This means that 5 factors are sufficient in analysing the variables. Therefore, the total number of factor that can be found is 5.

2. Which variables are grouped by the first two factors? (e.g. threshold |loading| ≥ 0.5)

```
print(loadings(fa5), 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
```

The groups involved in factor 1 are WEIGHT, THIGH, CHNUP and FAT. The groups involved in factor 2 are HEIGHT, WEIGHT, SHLDR, PELVIC and BREATH which is mostly the body measurement of police applicants.

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…)

Diastolic blood pressure is the pressure in the arteries when the heart rests between beats. The definition of PULSE in the sample is the resting pulse rate. They are similar by definition.

According to the the results above (cutoff >= 0.5), DIAST has no value greater than 0.5 in all factors meaning that its significance in the data is not very beneficial. Even the results in question 1 above (without cutoff) shows that DIAST has very low values.

Therefore, I agree for it not needing to be measured and to use the resting pulse rate instead.

4. When we want to separate huge athletic people from big 'doughnut loving' police applicants, which factor scores can be used?

```
print(loadings(fa5), 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
```

The big 'doughnut loving' police applicants are grouped in factor 1. This is because it has large weight value therefore large total body fat value and large thigh thickness value. They also perform less chin ups per minute as it is negative value that contributes to their weight. Overall, they are heavier.

The huge athletic police applicants are grouped in factor 2. This can be seen due to a very large value in height hence taller, and their weight is also large. Due to their huge body, their shoulder width and pelvic width are also large. They are also athletic as their max breathing capacity is also large; usually athletic people have large max breathing capacity.