Decathlon PCA Exercise

Julianne Quinn

Source files & load and clean data

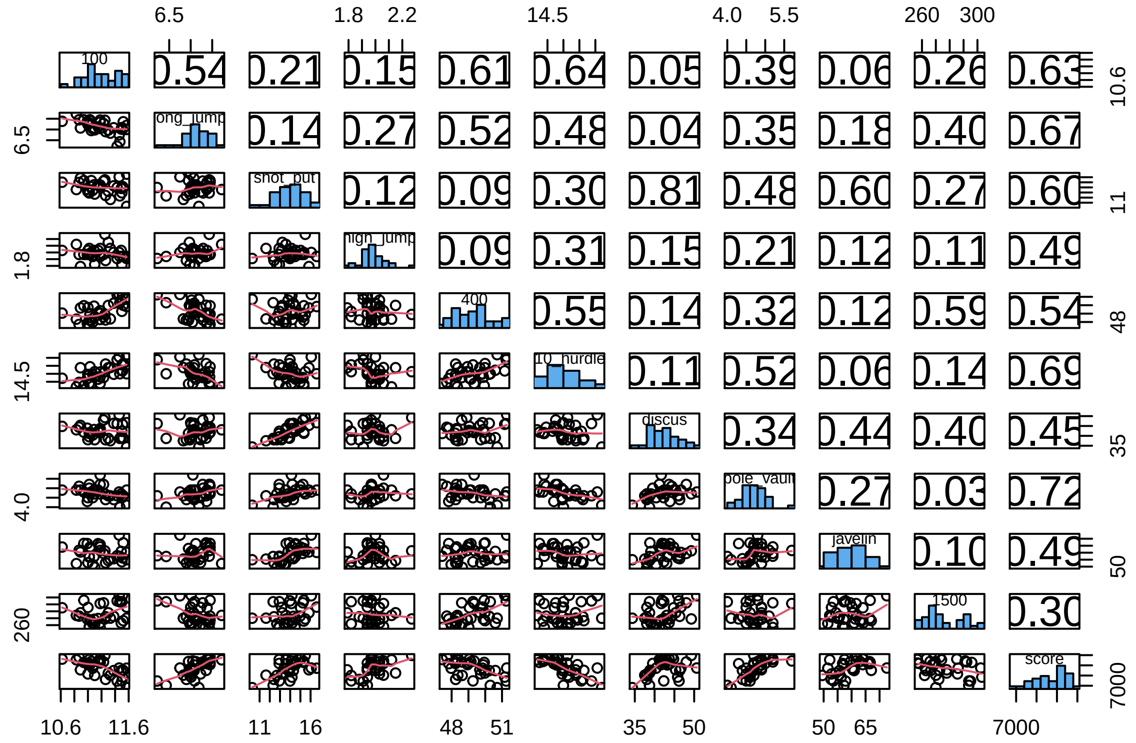
```r
source("SPM_Panel.R")
source("PCAplots.R")

setwd(sourcedir)

library(ade4)

# load olympic dataset with decathlon data
data(olympic)

# create a data frame with the decathlon data
decathlon.data <- olympic$tab
decathlon.data$score <- olympic$score

# change French event names to English
names(decathlon.data)[names(decathlon.data) == "long"] <- "long_jump"
names(decathlon.data)[names(decathlon.data) == "poid"] <- "shot_put"
names(decathlon.data)[names(decathlon.data) == "haut"] <- "high_jump"
names(decathlon.data)[names(decathlon.data) == "110"] <- "110_hurdles"
names(decathlon.data)[names(decathlon.data) == "disq"] <- "discus"
names(decathlon.data)[names(decathlon.data) == "perc"] <- "pole_vault"
names(decathlon.data)[names(decathlon.data) == "jave"] <- "javelin"
```

1. Describe this dataset. What are the variables and what are the different observations? What relationships would you expect from this dataset?

The dataset contains 33 observations of decathlon performances on the 10 events and their corresponding scores. We would expect jump scores to be correlated with each other (long jump, high jump, pole vault) and perhaps also with the sprinting events (100, 110 hurdles, 400). We would expect throwing events to be correlated with each other, but less so with the jumping and sprinting events (shot put, javelin, discus). Finally, the distance event is probably unlike the others (1500).

2. Now look at your data using scatter plot matrices.

```r
uva.pairs(decathlon.data)
```



3. Which features are strongly correlated? Which are most predictive of score?

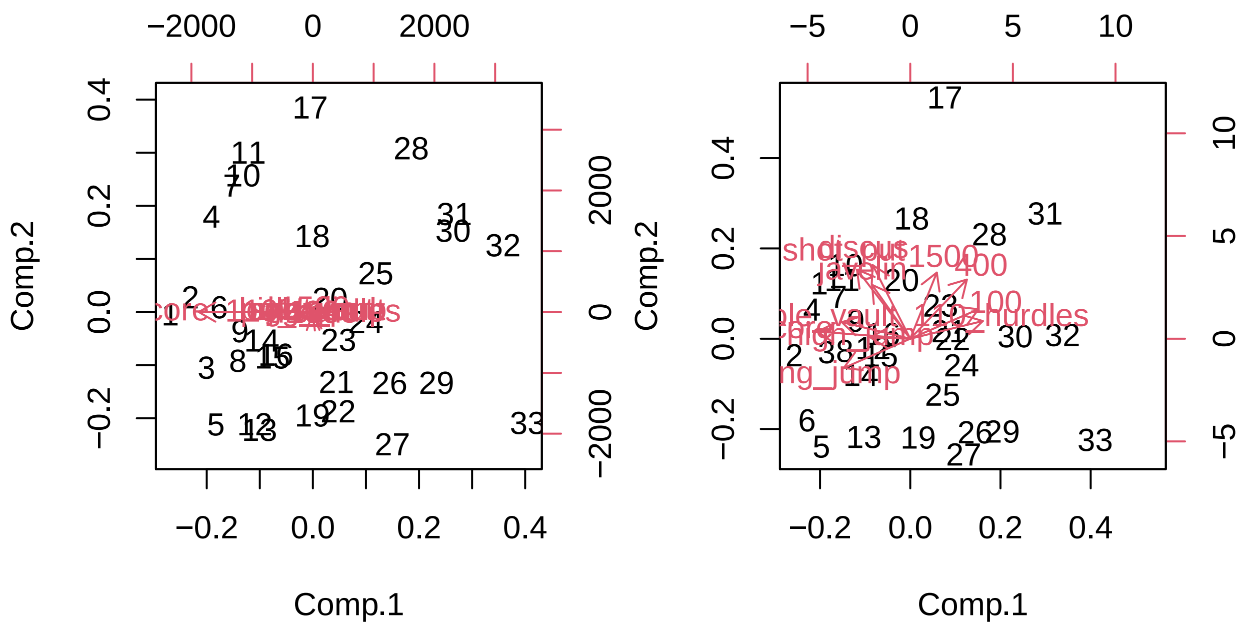
Pole valut, the 110 hurdles and long jump are the most predictive events in that order. The 1500 is the least predictive. 100, 400 and 110 hurdles are highly correlated. Long jump is most correlated with the 400. Shot put and discus are highly correlated. Javelin is next most correlated with them. High jump isn't as strongly correlated with the other events. The 400 and 1500 are fairly correlated. Pole vault is most correlated with the 100 hurdles.

4. Now create principal components with both the covariance and correlation matrices.

```r
decathlon.pca.cov <- princomp(decathlon.data)
decathlon.pca.corr <- princomp(decathlon.data, cor=T)
```

5. Create a biplot of your data for both the covariance and correlation matrices

```r
par(mfrow=c(1,2))
biplot(decathlon.pca.cov)
biplot(decathlon.pca.corr)
```



```
par(mfrow=c(1,1))
```

6. What do you notice about the biplots with the 2 methods?

Score dominates everything with the covariance matrix and you can't see how the other variables relate to it and one another. With the correlation matrix, the importance is much more even across events and score.

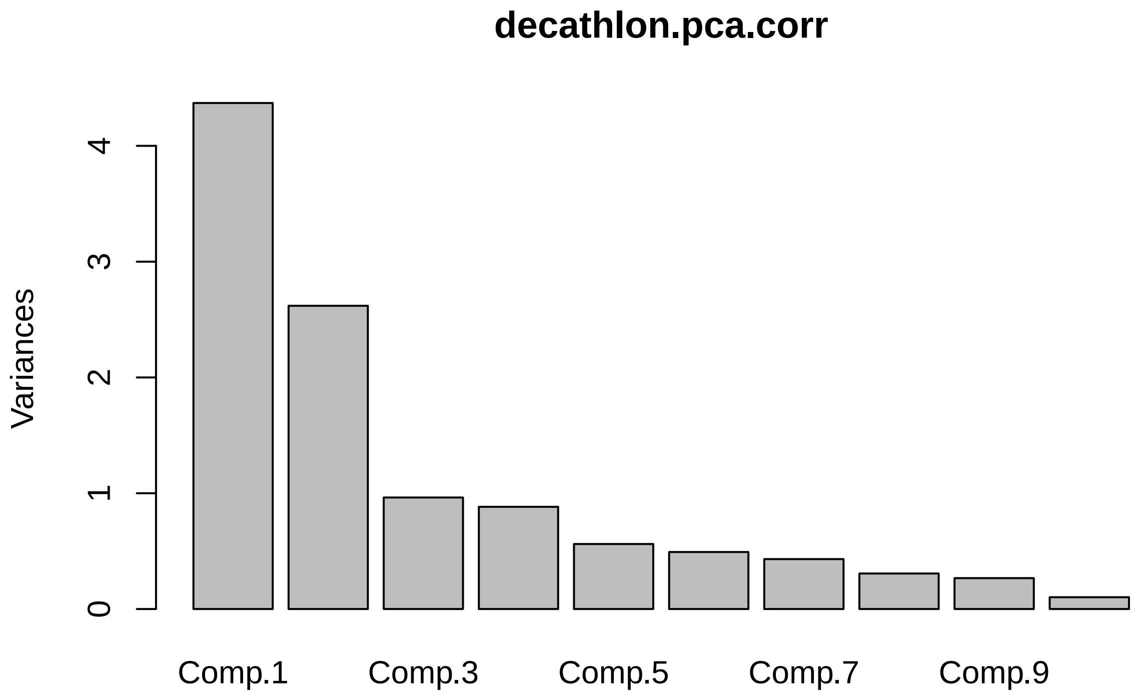
From here on out, use the correlation matrix.

7. Describe the relationships between the different variables. What do they imply about which decathlon events require similar or different strengths? (Note: In some events you want high scores, and in some events you want low scores.)

Running events are correlated with each other, throwing events are correlated with each other, and jumping events are correlated with each other. Jumping events appear anti-correlated with running events, with throwing events in between. However, this is because high values of jumping events are good, while low values of running events are good. If you negate the running times, running and jumping events would appear positively correlated, with throwing events being independent.

8. Create a screeplot.

```
screeplot(decathlon.pca.corr)
```

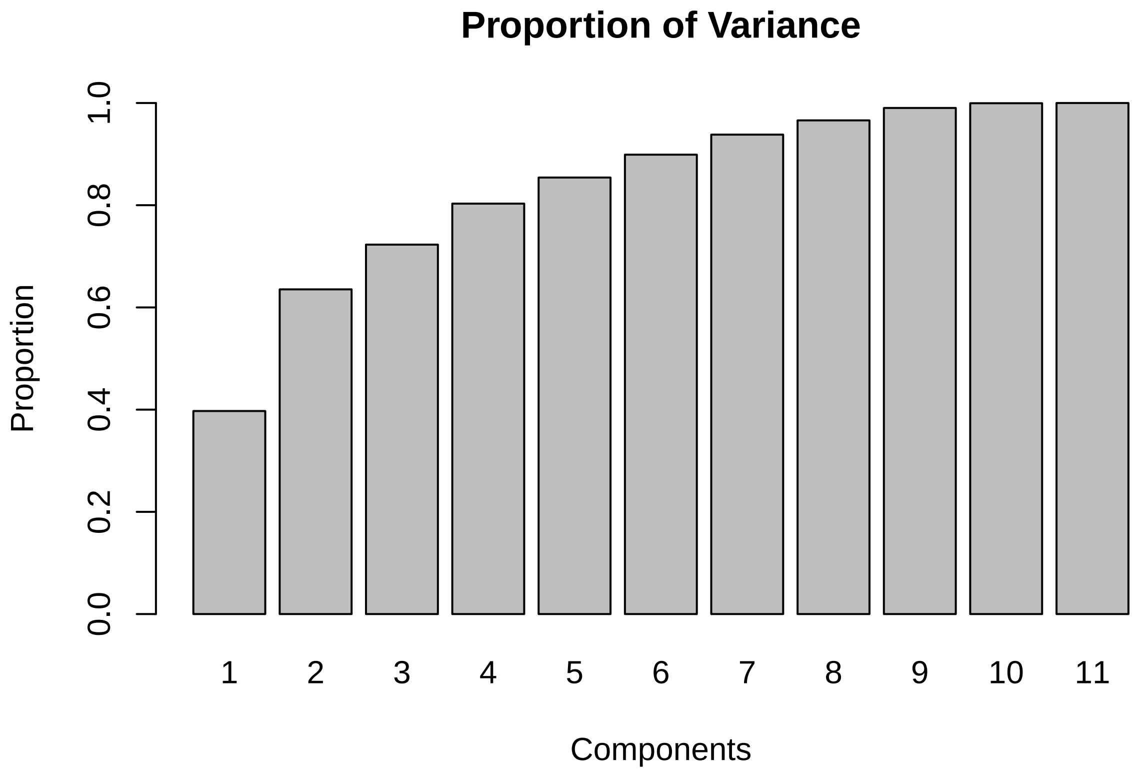


9. How many components do the graphs suggest is sufficient to explain most of the variation in the data? (Hint: consider the first method for choosing the number of PCs described in the slides.)

Two components have high variance before the kink in the scree plot, suggesting 2 should be sufficient, or perhaps 3 to include the kink location itself.

10. Create a cumulative variance plot.

```r
cumplot(decathlon.pca.corr)
```



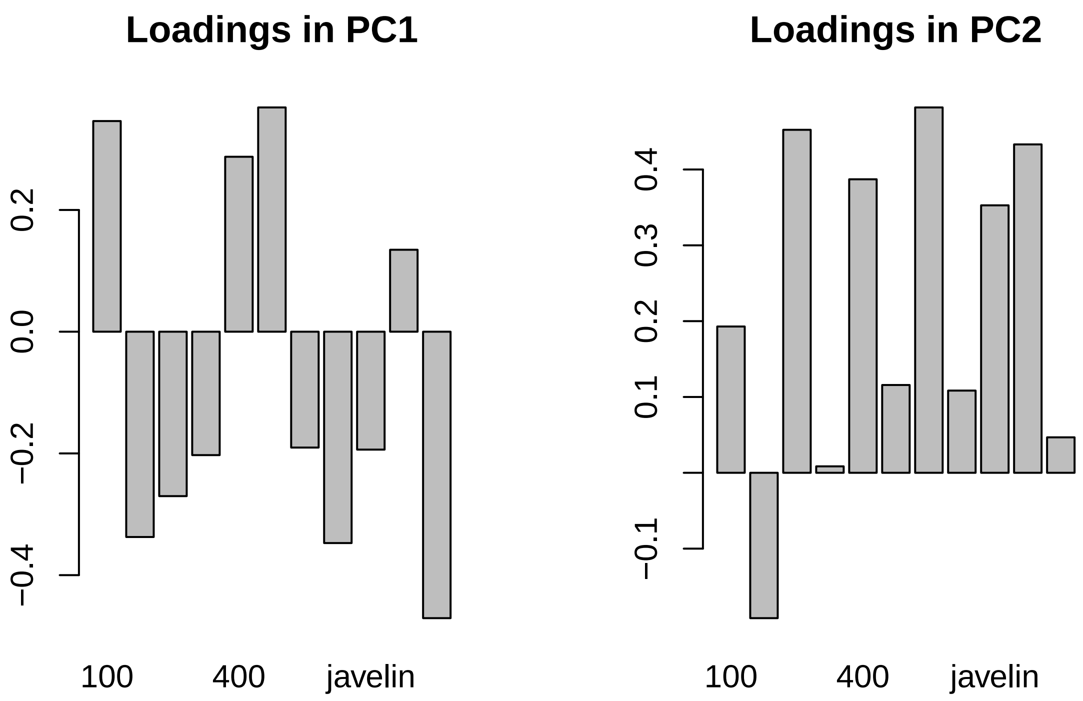
```
##     Component Proportion
##  1:    Comp.1  0.3972695
##  2:    Comp.2  0.6353338
##  3:    Comp.3  0.7228882
##  4:    Comp.4  0.8031322
##  5:    Comp.5  0.8541653
##  6:    Comp.6  0.8989206
##  7:    Comp.7  0.9381192
##  8:    Comp.8  0.9660116
##  9:    Comp.9  0.9902798
## 10:   Comp.10  0.9995670
## 11:   Comp.11  1.0000000
```

11. How many principal components do you need to explain at least 80% of the variance?

4 PCs explain 80.3% of variance.

12. Plot the loadings in the first 2 PCS.

```
par(mfrow=c(1,2))
barplot(decathlon.pca.corr$loadings[,1],main="Loadings in PC1")
barplot(decathlon.pca.corr$loadings[,2],main="Loadings in PC2")
```



```
par(mfrow=c(1,1))

decathlon.pca.corr$loadings
```

```
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## 100          0.346  0.193  0.315         0.446         0.258  0.663  0.108
## long_jump   -0.337 -0.192  0.152  0.198 -0.407 -0.102  0.753  0.142
## shot_put    -0.270  0.452 -0.115  0.107         0.222 -0.100        -0.422
## high_jump   -0.203         0.770 -0.509               -0.124 -0.153  0.102
## 400          0.287  0.387  0.195        -0.148 -0.318  0.130 -0.149 -0.651
## 110_hurdles  0.368  0.116  0.212  0.377  0.100  0.209  0.272 -0.640  0.207
## discus      -0.190  0.482                       0.607  0.157         0.167
## pole_vault  -0.347  0.108 -0.180 -0.134  0.686 -0.383  0.284 -0.275
## javelin     -0.194  0.353  0.237  0.555 -0.136 -0.443 -0.339         0.306
## 1500         0.135  0.433 -0.282 -0.451 -0.329 -0.281  0.171         0.457
## score       -0.471         0.126
##             Comp.10 Comp.11
## 100          0.104   0.107
## long_jump           -0.136
## shot_put     0.660  -0.138
## high_jump    0.131  -0.190
## 400         -0.345   0.124
## 110_hurdles  0.253   0.141
## discus      -0.526  -0.167
## pole_vault          -0.200
## javelin     -0.121  -0.180
## 1500         0.232   0.181
## score                0.864
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.091  0.091  0.091  0.091  0.091  0.091  0.091  0.091  0.091
## Cumulative Var  0.091  0.182  0.273  0.364  0.455  0.545  0.636  0.727  0.818
##                Comp.10 Comp.11
## SS loadings      1.000   1.000
## Proportion Var   0.091   0.091
## Cumulative Var   0.909   1.000
```

13. Which 5 variables explain most of the variability in the first PC and how are they related to each other?

Score, 110 hurdles, pole vault, 100, and long jump explain most of the variability in the first PC. Score is positively correlated with long jump and pole vault, meaning longer/higher jumps result in higher scores. Score is negatively correlated with 110 hurdles and 100 meaning lower (i.e. faster) times result in higher scores.

14. Which 5 variables explain most of the variability in the second PC and how are they related to each other?

Discus, shot put, 1500, 400 and javelin explain most of the variability in the second PC. They are all positively correlated with each other. It is not surprising discus, shot put and javelin or correlated with each other, or the 1500 and 400, but it is interesting the two sets are correlated with each other in the second mode of variability. Score has a loading of virtually 0, so these events appear independent of one's score.

15. Based on all of the analyses above, which events do you think are most important to an athlete's decathlon score?

The short sprints (100, 110 hurdles) and jumping events (long jump and pole vault) seem most important to one's decathlon score. The throwing events are of secondary importance. Interestingly, the 400 is too long a sprint to be very predictive and is actually more correlated with the 1500, the least important event. The high jump is not very important nor is it very correlated with the other jumping events.