**Homework**

**Principal Component Analysis of Olympic performance data**

30th November, 2016

Soütrik BANERJEE

1. Data pre-processing:

There are 11 continuous variables, with 34 observations in the data, and no missing values.

* 1. The variables are centred around their respective means and stored in the matrix oly\_centre.
  2. The standard deviations of the mean-centred variables are given below:

Centred variable Std. Dev

---------------+-------------

c100\_mt\_cen | 0.29

distance\_cen | 0.37

weight\_cen | 1.50

height\_cen | 0.10

c400\_mt\_cen | 1.18

c110\_mt\_cen | 0.61

disc\_cen | 4.50

pole\_vault\_cen | 0.49

javelin\_cen | 6.44

c1500\_mt\_cen | 13.48

score\_cen | 594.58

-----------------------------

* 1. The mean-centred variables are further standardised by scaling, that is dividing the mean-centred variables by their respective standard deviations in order to obtain the respective standardised scores and stored in the matrix oly\_renorm.

1. Principal Components Analysis (PCA) representation of the individuals:
   1. (1 / n)\*(t(oly\_renorm) %\*% oly\_renorm) computes variance-covariance matrix on the standardised variables , which is the same obtained by ((n - 1) / n)\*cov(oly\_renorm), which computes the corrected variance-covariance matrix on the standardised variables using the correction factor (n - 1) / n. Both these procedures compute the corrected correlation matrix as the variables are standardised.
   2. Summary of PCA:

pca\_sum

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

Standard deviation 2.4151253 1.4208932 0.84801191 0.81760692 0.60537162 0.54195434

Proportion of Variance 0.5463257 0.1891016 0.06735598 0.06261255 0.03432546 0.02751045

Cumulative Proportion 0.5463257 0.7354273 0.80278327 0.86539582 0.89972128 0.92723173

Comp.7 Comp.8 Comp.9 Comp.10 Comp.11

Standard deviation 0.52653227 0.46611297 0.44587469 0.283807477 0.0553157295

Proportion of Variance 0.02596703 0.02034954 0.01862078 0.007544317 0.0002865956

Cumulative Proportion 0.95319876 0.97354830 0.99216909 0.999713404 1.0000000000

The 11 standard deviations or eigenvalues (λi) and (cumulative) proportions of the total variance explained are obtained in descending order. The first two principal components explain ~74%.

pca\_sum$loadings

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11

c100\_mt 0.325 -0.202 -0.329 0.163 0.279 0.115 0.513 0.155 0.570

distance -0.330 0.195 -0.470 -0.609 0.379 0.252 0.139 0.118

weight -0.296 -0.396 0.123 0.175 0.258 -0.288 0.390 0.622 0.138

height -0.248 -0.812 -0.432 -0.151 -0.151 0.137 0.153

c400\_mt 0.269 -0.425 -0.241 -0.278 0.588 -0.368 -0.341 -0.102

c110\_mt 0.339 -0.129 -0.193 0.384 -0.367 0.138 -0.416 -0.495 0.308 -0.120

disc -0.280 -0.418 0.433 -0.437 -0.188 -0.554 0.125

pole\_vault -0.356 0.154 0.384 0.655 -0.467 0.195

javelin -0.270 -0.300 -0.164 0.542 -0.464 0.369 -0.110 -0.302 0.172 -0.108 0.144

c1500\_mt -0.545 0.270 -0.545 -0.349 0.156 -0.304 0.138 0.222 -0.129

score -0.406 -0.906

The loadings give the 11 principal components (eigenvectors) for the 11 item events, column wise displayed.

head(pca\_sum$scores)

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

[1,] -2.485055 -0.5935943 -2.37584499 -0.28200498 0.93408795 -1.10093950 -0.41355623

[2,] -2.686145 0.5922282 0.90180117 0.06079728 0.02775373 -0.07244461 -0.05618751

[3,] -1.986176 0.6251049 0.01707781 0.92440178 0.14301338 0.20707708 0.66363659

[4,] -2.446042 -0.4337844 0.54290579 -0.48187924 -0.88945961 -0.32905882 -0.84393909

[5,] -1.909035 2.2701508 0.41827822 0.11949649 0.35572711 0.32685165 0.60294770

[6,] -2.345180 1.5175227 -0.19003199 -1.71180671 -0.22535795 -0.89407601 0.01008317

Comp.8 Comp.9 Comp.10 Comp.11

[1,] -0.25510975 0.24539019 0.21831153 -0.01991287

[2,] -0.16741110 0.15176077 0.16845417 -0.02094016

[3,] -0.01611195 0.15183492 -0.15241116 -0.03202174

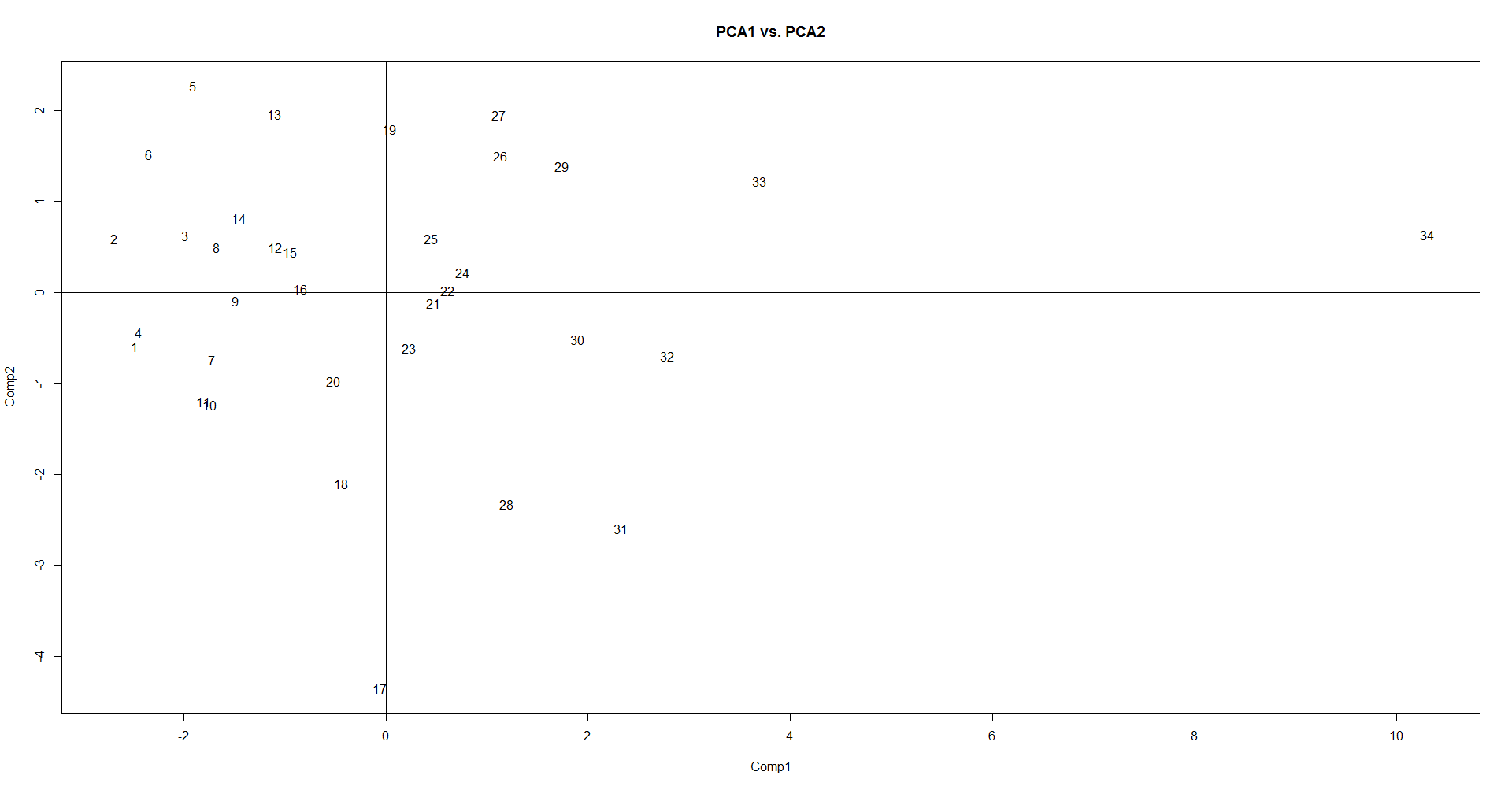
[4,] -0.33756003 -0.64863117 0.01312498 -0.03413176

[5,] -0.02003676 0.13984925 -0.39195437 -0.04016003

[6,] 0.38946068 0.05404685 -0.14960103 0.02339809

The scores give the linear contributions of each standardised observational unit or row (= athlete) to the respective principal components, which are column wise displayed here for the first 6 athletes only (from the total of 34 athletes). In another way, the score value for an observational unit, e.g., the PC1 score for an athlete, is the distance from the origin, along the direction vector (eigenvector) of the PC1, up to the point where that observational unit projects onto the scaled eigenvector.

* 1. Plot of PC1 vs. PC2 scores for the 34 numbered athletes



1. PCA representation of the variables:
   1. cor(oly\_renorm[, 1], pca\_oly$scores) represents the correlation between 100 metres race event and all the 11 principal component scores. Particularly, the correlation between 100 metres race and PC1 scores is quite high, ~0.8 here. This means that this event performance is strongly correlated to the average performance of the athletes as the PC1 determines the average inertia, or the size variable. It also shows how this event correlates with other principal component scores.

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8

[1,] 0.7975913 -0.2906213 -0.2831333 0.1349345 0.1713266 0.06299968 0.2739705 0.07313511

Comp.9 Comp.10 Comp.11

[1,] 0.2578688 0.02766282 -0.004866813

* 1. The norm of a vector is the distance (here Euclidean), which is the scalar product of the elements of the vector.
  2. cor(oly\_renorm, pca\_oly$scores[, 1]) represents the correlation between the PC1 scores and all the 11 item events. One can observe here that PC1 (dimension 1) distinguishes those having positive [‘right half’] scores (athletes good in sprint, hurdles and medium range running events) from those having negative [‘left half’] scores (athletes good in decathlon events – e.g., high jump, shotput, javelin, discus, pole vault, long distance marathon running, and the official performance named as ‘score’ here).

[,1]

c100\_mt 0.7975913

distance -0.8097141

weight -0.7245773

height -0.6068554

c400\_mt 0.6599689

c110\_mt 0.8321740

disc -0.6852117

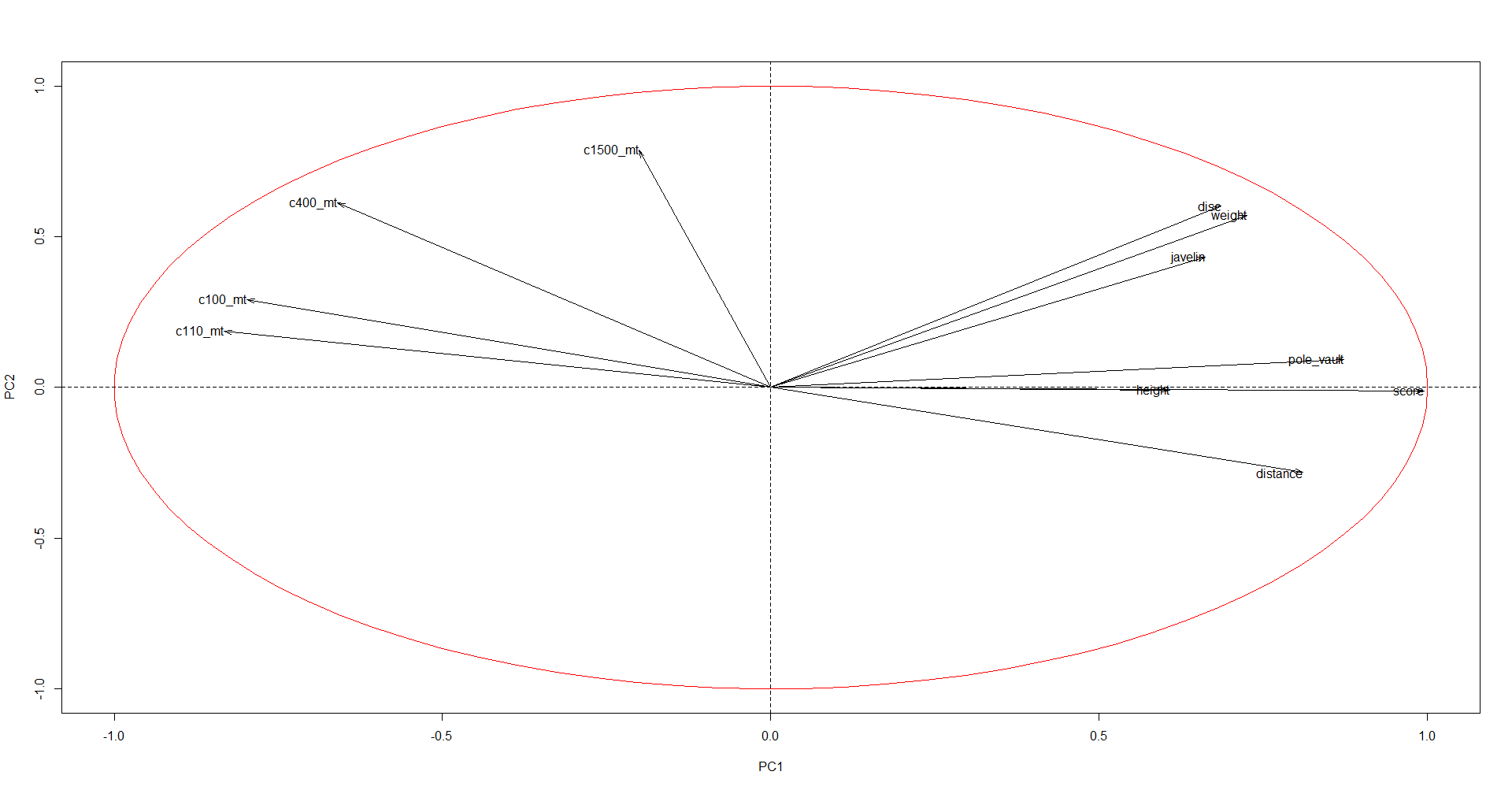
pole\_vault -0.8723055

javelin -0.6610545

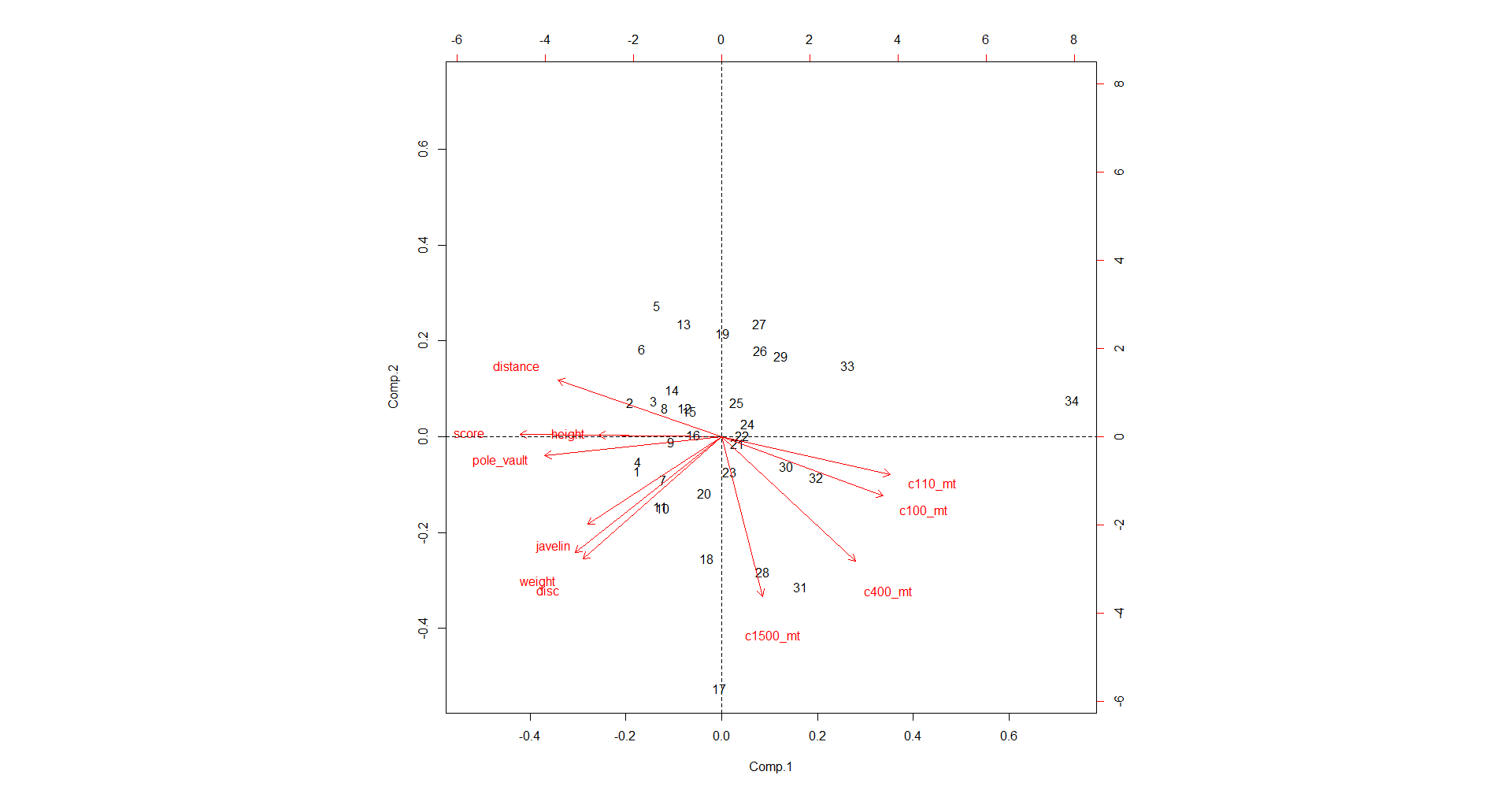
c1500\_mt 0.2010652

score -0.9942753

* 1. Again, PC2 (dimension 2) distinguishes those who are long distance runners and high jumpers (positive [‘upper half’] scores) from others (negative [‘lower half’] scores) [output not shown]. PC2, and so forth are called the shape or comparison variables.
  2. Correlation circle:

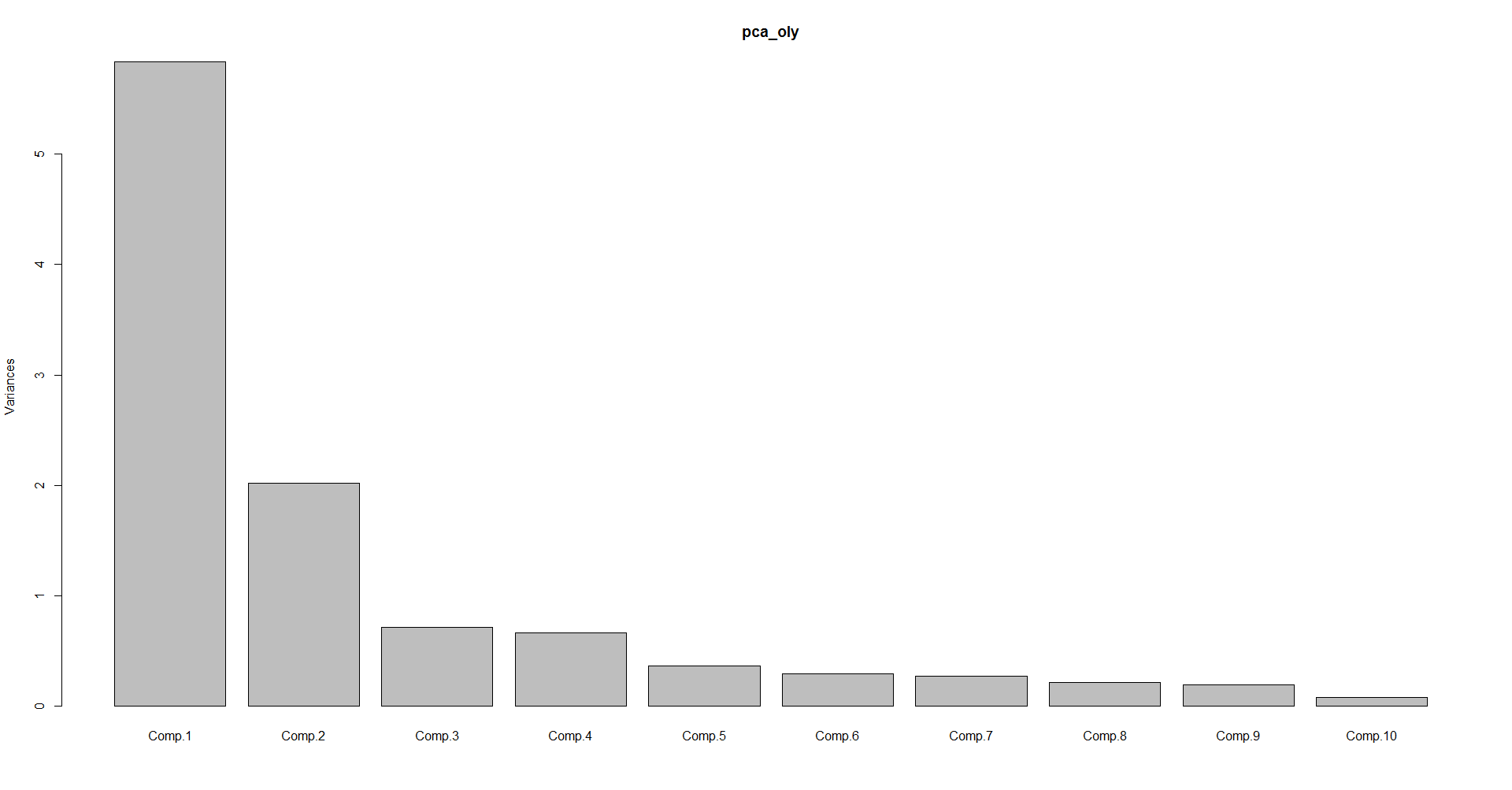


1. Generally:
   1. Biplot –



* 1. Screeplot

plot(pca\_oly) or screeplot(pca\_oly) produces a screeplot, which shows the squared eigenvalues (= squared standard deviations, λi) in descending order.



* 1. If the original values are not standardised (= mean-centred and then scaled by standard deviation), then the variables with greater variability will “pull” the results (eigenvalues) towards itself, thereby falsely biasing the result heavily. In this case, the variable “score” will cause this impact. Because PCA tries to capture the total variance in the set of variables, it requires that the input variables have similar scales of measurement.

Discussion

In the current study, the variable “score” correlates very strongly with the PC1 score. This variable, which is a summary official score of the ‘decathlon, marathon’ events, picks up alone maximal variability to explain the overall performance for these events. This variable should be excluded in the current analysis, if we are to differentiate between the ‘sprint running’ events with that of ‘decathlon, marathon’ events.