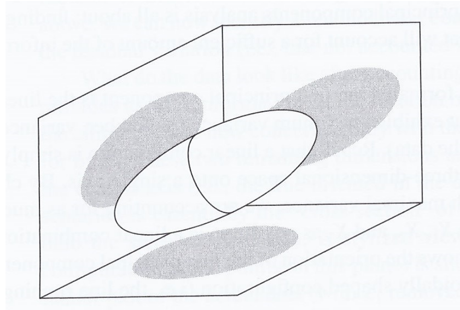
Principal Components Analysis (PCA)

Principal Component Analysis

- Explain the variance-covariance structure of a set of variables through a few *linear* combinations of the variables
- Dimension reduction
  - $m(<p)$ principal components replace the initial p variables if it is possible to account for most of the information in the original data.
- Pattern recognition in the relationship among the variables

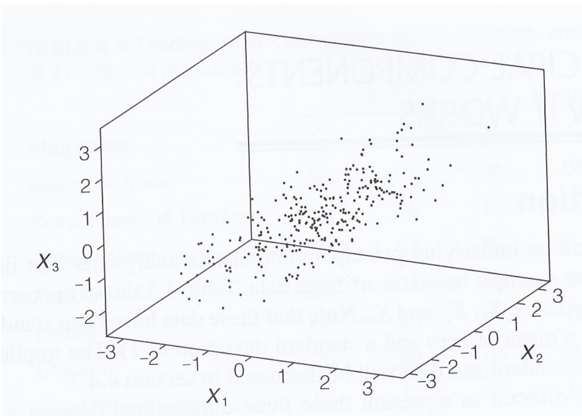
Intuition



- 3 variables: $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x}_2, \boldsymbol{x_3})$
- 3-dimensional data on plots
  - 3-d scatter plot
  - 2-d scatter plot matrix
- Positive relationship among three variables

  ➔ Express most of the information in $(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})$ using one variable

  ➔ Find the linear combination $\mathbf{y} = a_1\boldsymbol{x_1} + a_2\boldsymbol{x_2} + a_3\boldsymbol{x_3}$ with the largest variance



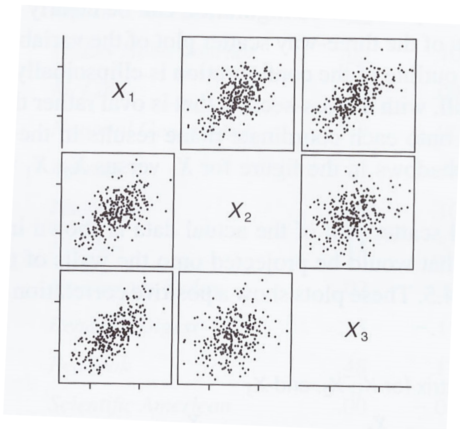
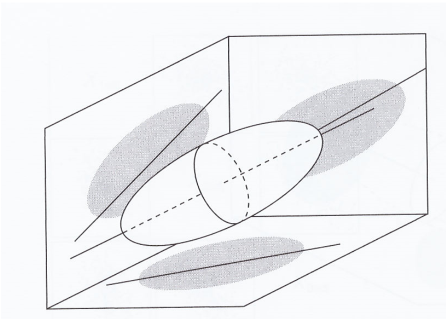


TABLE 4.5 Correlation matrix for X_1, X_2, and X_3

	X_1	X_2	X_3
X_1	1.000	0.562	0.704
X_2	0.562	1.000	0.304
X_3	0.704	0.304	1.000
	$\text{var}(X_1) = 1.00$	$\text{var}(X_2) = 1.00$	$\text{var}(X_3) = 1.00$

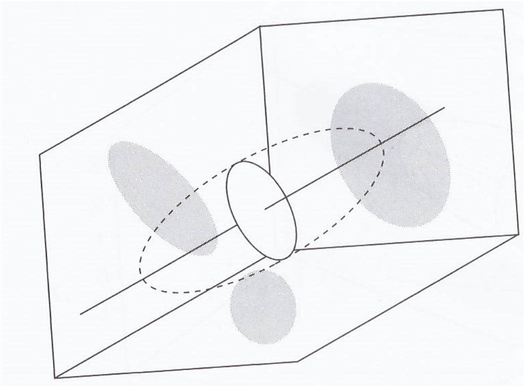
Intuition :The 1st PC

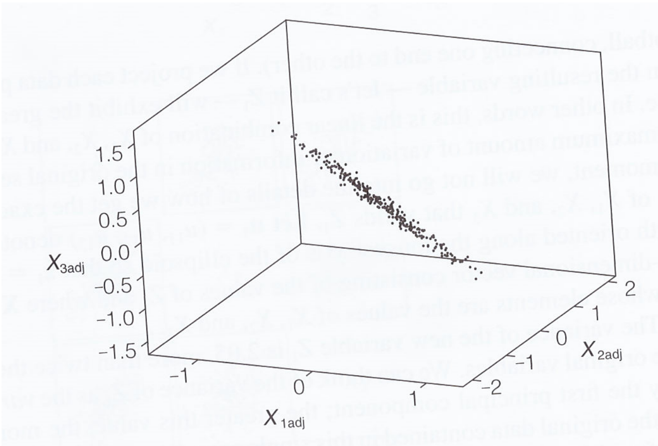


- 1st principal component(PC): the longest axis of the ellipsoid

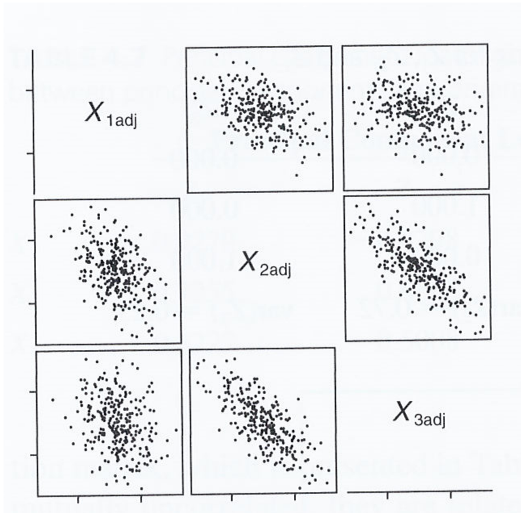
  unit vector in the direction of the 1st PC: $\boldsymbol{a}_1' = (a_{11}, a_{12}, a_{13})$
- Projection of each data point on the 1st PC➔ a new variable $\boldsymbol{y}_1 = \boldsymbol{a}_1'\boldsymbol{X}$
- $var(\boldsymbol{y}_1) = 2.05$: the larger variance means the more information it contains
- The other two dimensions are expressed as a plane orthogonal to the 1st PC (a cross section of the football)

Intuition : The 2nd PC





- What are the data after removing the information in $\boldsymbol{y_1}$?
- Project the data points on the plane orthogonal to the 1st PC
- Scatter plot matrix after removing the information in $\boldsymbol{y_1}$



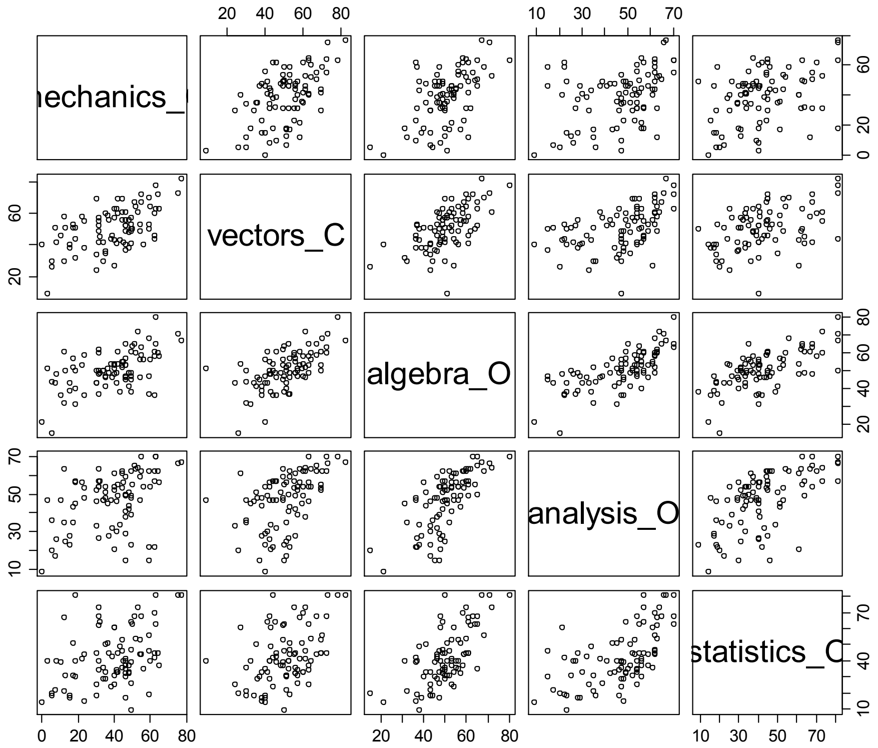
- Find the unit vector of the direction having the largest variance in the left 3d scatter plot: $\boldsymbol{a_2'} = (a_{21}, a_{22}, a_{23})$
- $\boldsymbol{y_2} = \boldsymbol{a_2'}X$

Example : Open/closed book

- Test score for mechanics, vectors, algebra, analysis, statistics

```
> head(data,10)
   mechanics_C vectors_C algebra_O analysis_O statistics_O
1           77        82        67         67           81
2           63        78        80         70           81
3           75        73        71         66           81
4           55        72        63         70           68
5           63        63        65         70           63
6           53        61        72         64           73
7           51        67        65         65           68
8           59        70        68         62           56
9           62        60        58         62           70
10          64        72        60         62           45
> str(data)
'data.frame': 88 obs. of  5 variables:
 $ mechanics_C : int  77 63 75 55 63 53 51 59 62 64 ...
 $ vectors_C   : int  82 78 73 72 63 61 67 70 60 72 ...
 $ algebra_O   : int  67 80 71 63 65 72 65 68 58 60 ...
 $ analysis_O  : int  67 70 66 70 70 64 65 62 62 62 ...
 $ statistics_O: int  81 81 81 68 63 73 68 56 70 45 ...

> cor(data)
             mechanics_C vectors_C algebra_O analysis_O statistics_O
mechanics_C    1.0000000 0.5534052 0.5467511  0.4093920    0.3890993
vectors_C      0.5534052 1.0000000 0.6096447  0.4850813    0.4364487
algebra_O      0.5467511 0.6096447 1.0000000  0.7108059    0.6647357
analysis_O     0.4093920 0.4850813 0.7108059  1.0000000    0.6071743
statistics_O   0.3890993 0.4364487 0.6647357  0.6071743    1.0000000
```



How to Find the Principal Components?

- $S \in R^{p \times p}$: sample covariance matrix of $\boldsymbol{x}$
- Find $\mathbf{y} = \boldsymbol{a}'X$ with the maximum variance

$$var(\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}$$

If X is standardized, $\boldsymbol{S}$ is the correlation matrix $\boldsymbol{R}$, $var(\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{R}\boldsymbol{a}$

- Maximize $\boldsymbol{a}'\boldsymbol{R}\boldsymbol{a}$ subject to $\boldsymbol{a}'\boldsymbol{a} = 1$
  - Use Lagrange multiplier method

$$L = \boldsymbol{a}'\boldsymbol{R}\boldsymbol{a} - \lambda(\boldsymbol{a}'\boldsymbol{a} - 1)$$
$$\frac{\partial L}{\partial \boldsymbol{a}} = 2\boldsymbol{R}\boldsymbol{a} - 2\lambda\boldsymbol{a} = 0$$
$$\boldsymbol{R}\boldsymbol{a} = \lambda\boldsymbol{a}$$

➔**Eigenvalue problem (λ: eigenvalue of $\boldsymbol{R}$, a: eigenvector of $\boldsymbol{R}$)**

➔If R is full rank, there exist p of real number eigenvalues.

➔If R is positive definite, all the eigenvalues are positive.

- Eigenvalues: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, the corresponding eigenvectors: $\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p}$
- 1st PC: $\boldsymbol{y_1} = \boldsymbol{e_1'}X$
- 2nd PC: $\boldsymbol{y_2} = \boldsymbol{e_2'}X$

…

If $S = \{s_{ik}\}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e_1}), (\lambda_2, \boldsymbol{e_2}), \ldots, (\lambda_p, \boldsymbol{e_p})$, the ith sample principal component is given by

$$y_i = \boldsymbol{e_i'}\boldsymbol{x} = e_{i1}x_1 + e_{i2}x_2 + \cdots + e_{ip}x_p, \qquad i = 1,2,\ldots,p$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and $\boldsymbol{x}$ is any observation on the variables $X_1, X_2, \ldots, X_p$.

Properties of the Principal Components

- Sample variance$(y_k) = \lambda_k, \quad k = 1, 2, \dots, p$
- Sample covariance$(y_i, y_k) = 0\ , \ i \neq k$
- Total sample variance$= \sum_{i=1}^{p} s_{ii} = \lambda_1 + \lambda_2 + \ \dots + \lambda_p$
- Correlation$(y_i, x_k) = r_{y_i, x_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{s_{kk}}}, \quad i, k = 1,2, \dots, p$
- Proportion of (standardized) sample variance due to ith principal component $= \frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_p} (= \frac{\lambda_i}{p})$

Example : Open/closed book

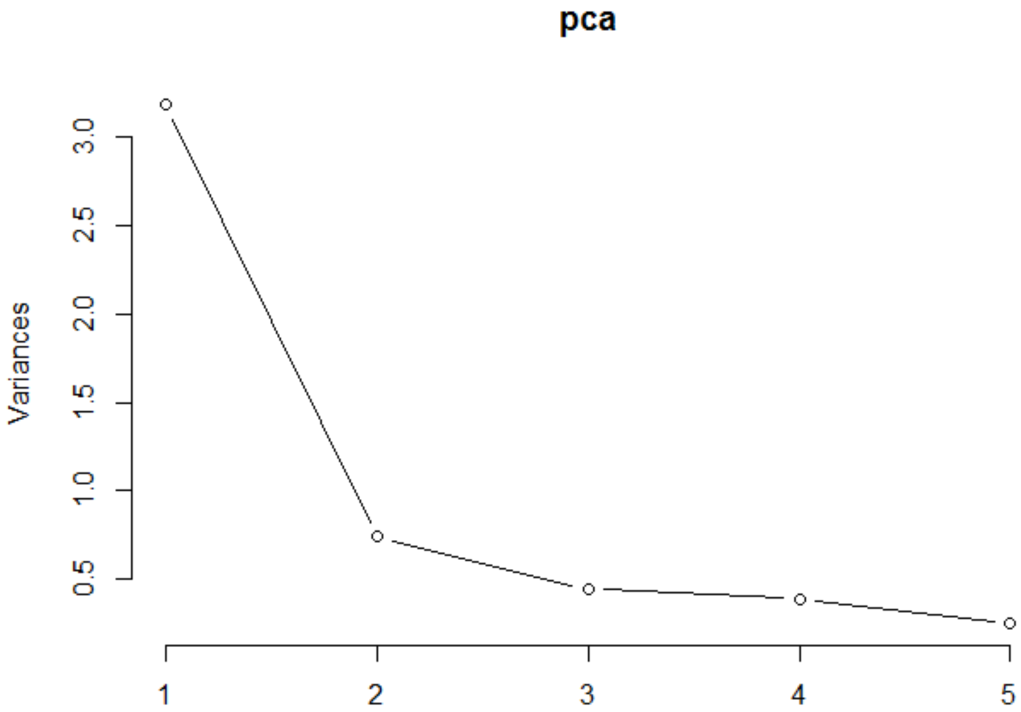
```
> pca=prcomp(data, scale=T)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.7835 0.8600 0.66706 0.62281 0.49658
Proportion of Variance 0.6362 0.1479 0.08899 0.07758 0.04932
Cumulative Proportion  0.6362 0.7841 0.87310 0.95068 1.00000
```

Number of Principal Components

- No definitive answers
- Consider
  - the amount of total variance explained
  - the relative sizes of the eigenvalues
  - the subject-matter interpretations of the components
- Rule of thumb
  - Use a *scree plot*
    - ✓ eigenvalues ordered from largest to smallest
    - ✓ Look for an elbow in the scree plot
    - ✓ The point at which the remaining eigenvalues are relatively small and all about the same size
  - Use the point at which the proportion of the variance explained by the principal components is between 70% and 90%
  - Exclude PC whose eigenvalues are less than the average $\sum_{i=1}^{p} \lambda_i / p$
    - ✓ If $\boldsymbol{R}$ is used for calculating PC, exclude PC whose eigenvalues are less than 1

Example : Open/closed book

```
> plot(pca,type='l')
```



Calculating Principal Components Scores

- The m principal components scores for individual i with original $p{\times}1$ vector of variable values $\boldsymbol{x}_i$ are obtained as

$$
\begin{gathered}
y_{i1} = \boldsymbol{a}_1^T \boldsymbol{x}_i \\
y_{i2} = \boldsymbol{a}_2^T \boldsymbol{x}_i \\
\vdots \\
y_{im} = \boldsymbol{a}_m^T \boldsymbol{x}_i
\end{gathered}
$$

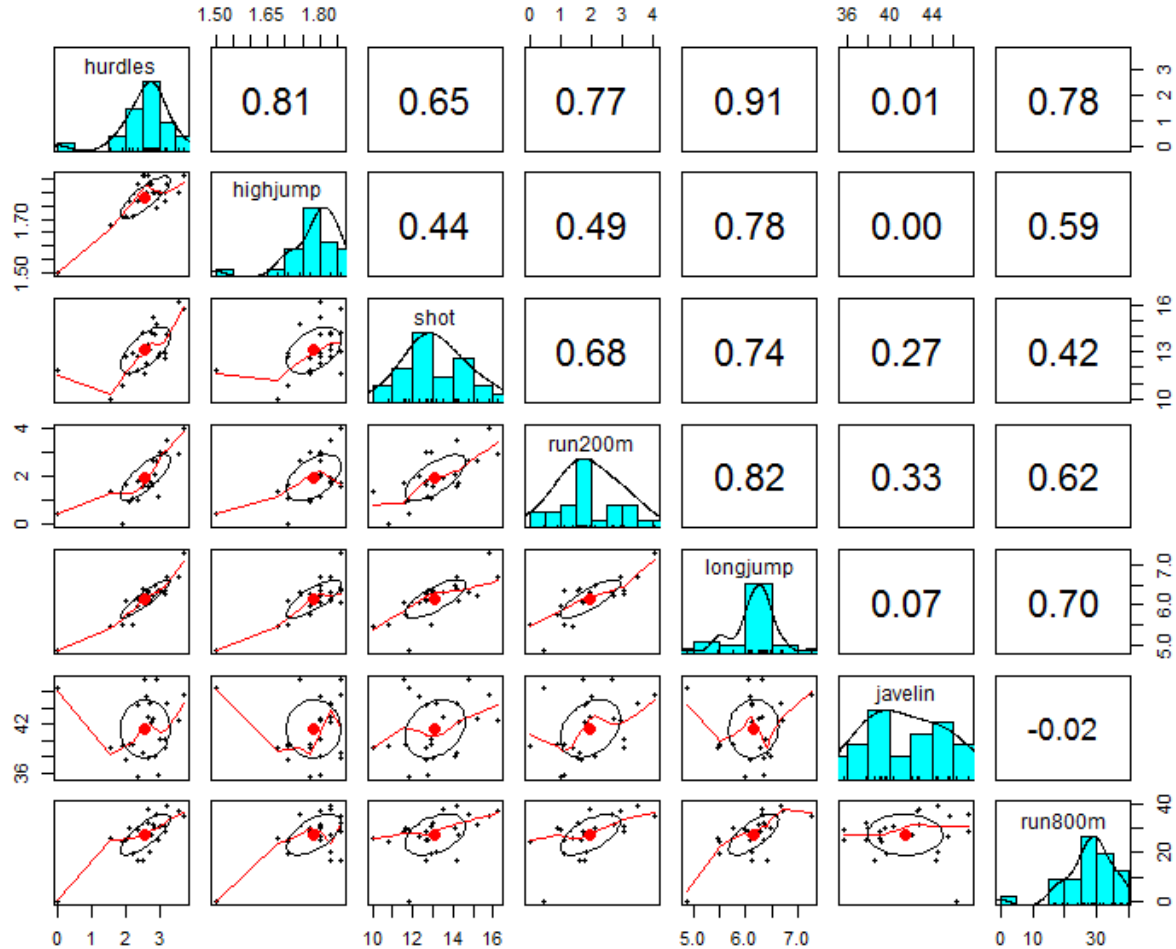
Example: Olympic Heptathlon Results

- The results for 25 competitors in all seven disciplines in the 1988 Olympics
- 100m hurdles, shot put, high jump, 200m, long jump, javelin, 800m
- Explore the structure of the data and assess how the derived principal components scores related to the scores assigned by the official scoring system

```
> heptathlon$hurdles <- with(heptathlon, max(hurdles)-hurdles)
> heptathlon$run200m <- with(heptathlon, max(run200m)-run200m)
> heptathlon$run800m <- with(heptathlon, max(run800m)-run800m)
> score <- which(colnames(heptathlon) == "score")
>
> round(cor(heptathlon[,-score]), 2)
         hurdles highjump shot run200m longjump javelin run800m
hurdles     1.00     0.81 0.65    0.77     0.91    0.01    0.78
highjump    0.81     1.00 0.44    0.49     0.78    0.00    0.59
shot        0.65     0.44 1.00    0.68     0.74    0.27    0.42
run200m     0.77     0.49 0.68    1.00     0.82    0.33    0.62
longjump    0.91     0.78 0.74    0.82     1.00    0.07    0.70
javelin     0.01     0.00 0.27    0.33     0.07    1.00   -0.02
run800m     0.78     0.59 0.42    0.62     0.70   -0.02    1.00
```

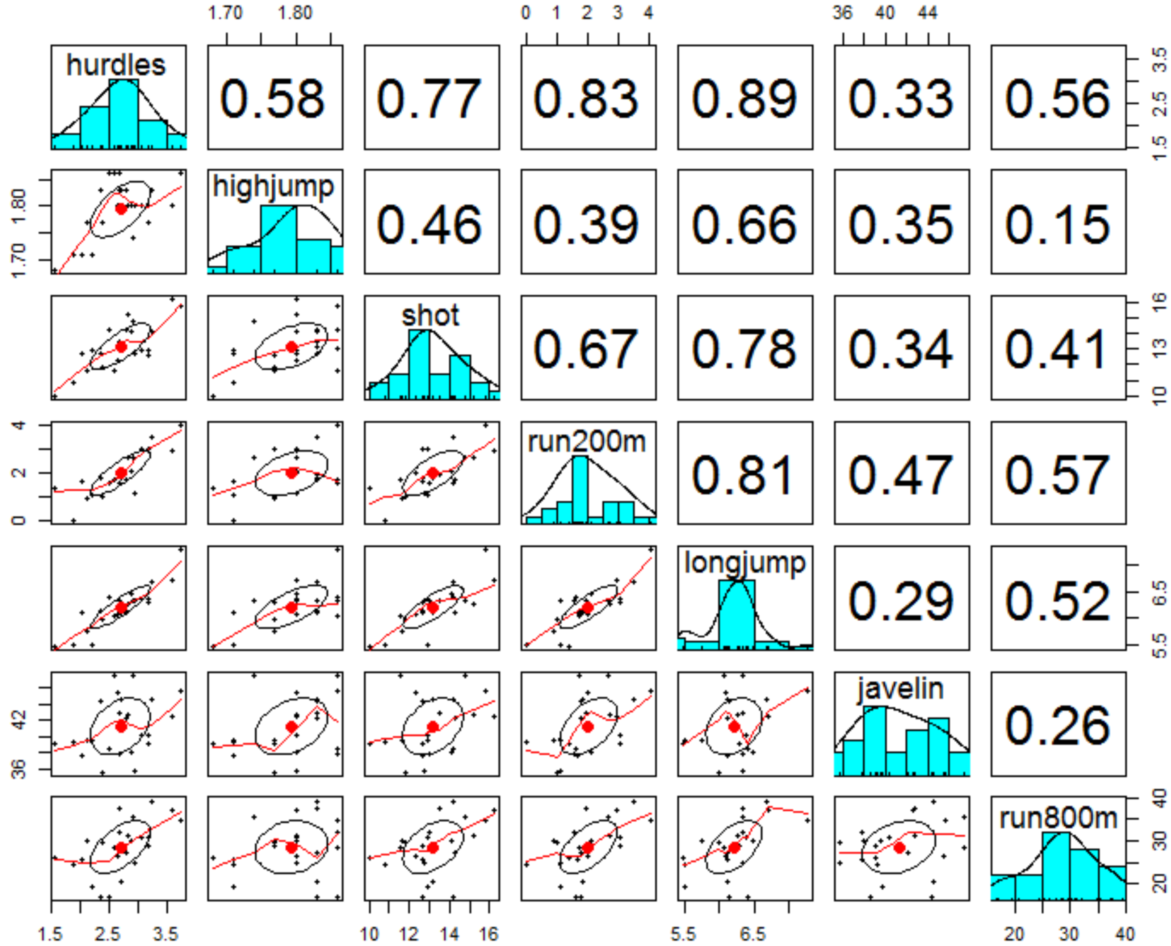
```r
library(psych)
pairs.panels(heptathlon[,-score])
```

- Most pairs of events are positively correlated
- javelin event and the others are less correlated
- There is an outlier (PNG)



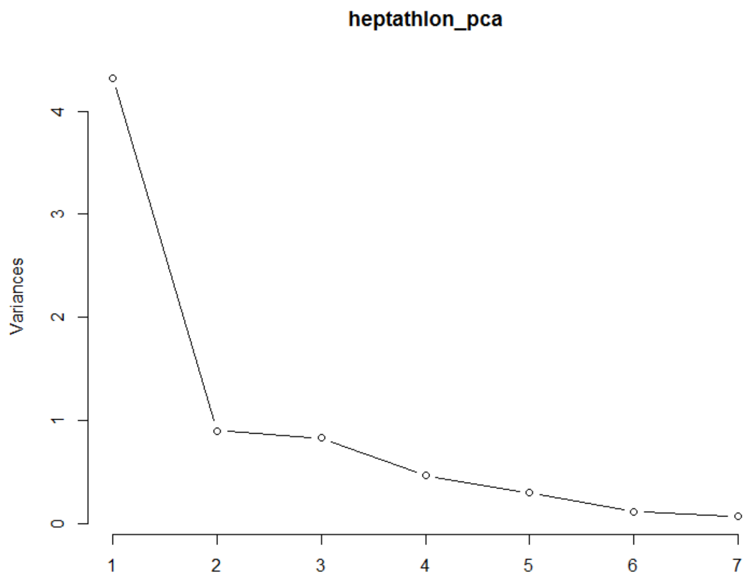
- Scatterplot matrix excluding the outlier

```
> heptathlon2 <- heptathlon[-grep("PNG", rownames(heptathlon)),]
> round(cor(heptathlon2[,-score]), 2)
         hurdles highjump shot run200m longjump javelin run800m
hurdles     1.00     0.58 0.77    0.83     0.89    0.33    0.56
highjump    0.58     1.00 0.46    0.39     0.66    0.35    0.15
shot        0.77     0.46 1.00    0.67     0.78    0.34    0.41
run200m     0.83     0.39 0.67    1.00     0.81    0.47    0.57
longjump    0.89     0.66 0.78    0.81     1.00    0.29    0.52
javelin     0.33     0.35 0.34    0.47     0.29    1.00    0.26
run800m     0.56     0.15 0.41    0.57     0.52    0.26    1.00
> pairs.panels(heptathlon2[,-score])
```



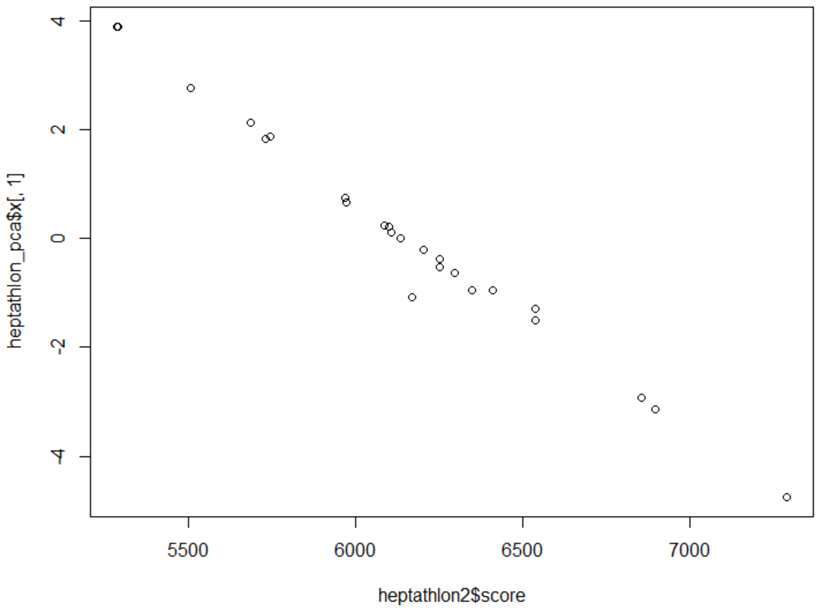
```
> heptathlon_pca <- prcomp(heptathlon2[, -score], scale = TRUE)
> print(heptathlon_pca)
Standard deviations:
[1] 2.08 0.95 0.91 0.68 0.55 0.34 0.26

Rotation:
           PC1    PC2   PC3    PC4    PC5    PC6    PC7
hurdles  -0.45  0.058 -0.17  0.048 -0.199  0.847 -0.070
highjump -0.31 -0.651 -0.21 -0.557  0.071 -0.090  0.332
shot     -0.40 -0.022 -0.15  0.548  0.672 -0.099  0.229
run200m  -0.43  0.185  0.13  0.231 -0.618 -0.333  0.470
longjump -0.45 -0.025 -0.27 -0.015 -0.122 -0.383 -0.749
javelin  -0.24 -0.326  0.88  0.060  0.079  0.072 -0.211
run800m  -0.30  0.657  0.19 -0.574  0.319 -0.052  0.077
> summary(heptathlon_pca)
Importance of components:
                         PC1   PC2   PC3    PC4    PC5    PC6     PC7
Standard deviation     2.079 0.948 0.911 0.6832 0.5462 0.3375 0.26204
Proportion of Variance 0.618 0.128 0.119 0.0667 0.0426 0.0163 0.00981
Cumulative Proportion  0.618 0.746 0.865 0.9313 0.9739 0.9902 1.00000
```



- The first two PC explains 74.6% of the total variation.
- The scree plot shows how the first two PC dominate.

```
> cor(heptathlon2$score, heptathlon_pca$x[,1])
[1] -0.99
> plot(heptathlon2$score, heptathlon_pca$x[,1])
```




```
> heptathlon_pca$x
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Joyner-Kersee (USA)	-4.758	-0.140	-0.006	0.2934	-0.362	-0.271	-0.476
John (GDR)	-3.148	0.949	-0.244	0.5492	0.754	0.378	-0.052
Behmer (GDR)	-2.926	0.695	0.622	-0.5547	-0.190	-0.258	0.111
Sablovskaite (URS)	-1.288	0.179	0.251	0.6372	0.604	-0.216	0.531
Choubenkova (URS)	-1.503	0.962	1.781	0.7840	0.590	0.080	-0.301
Schulz (GDR)	-0.958	0.351	0.413	-1.1135	0.715	-0.254	0.038
Fleming (AUS)	-0.953	0.500	-0.265	-0.1402	-0.866	0.037	0.230
Greiner (USA)	-0.633	0.376	-1.140	0.1426	0.208	-0.142	-0.064
Lajbnerova (CZE)	-0.382	-0.712	-0.068	0.0872	0.677	0.250	0.356
Bouraga (URS)	-0.522	0.777	-0.481	0.2837	-1.188	0.399	0.197
Wijnsma (HOL)	-0.218	-0.234	-1.154	-1.2601	0.375	-0.203	0.175
Dimitrova (BUL)	-1.076	0.516	-0.312	-0.1270	-0.920	0.267	0.211
Scheider (SWI)	0.003	-1.447	1.583	-1.2544	-0.205	0.176	-0.039
Braun (FRG)	0.109	-1.636	0.470	0.3626	-0.147	0.261	-0.013
Ruotsalainen (FIN)	0.209	-0.689	1.152	-0.1129	-0.315	0.184	-0.141
Yuping (CHN)	0.233	-1.960	-1.541	0.5983	0.175	-0.502	0.050
Hagger (GB)	0.660	-0.088	-1.797	-0.1824	-0.051	0.551	-0.464
Brown (USA)	0.757	-2.043	0.452	0.4769	-0.382	-0.266	-0.111
Mulliner (GB)	1.881	0.915	-0.359	0.7996	-0.069	-0.733	-0.313
Hautenauve (BEL)	1.828	0.726	-1.049	-0.7118	0.141	0.069	-0.075
Kytola (FIN)	2.118	0.399	0.190	-0.7884	0.418	-0.034	0.121
Geremias (BRA)	2.771	0.035	0.170	1.3856	0.285	0.381	0.346
Hui-Ing (TAI)	3.901	1.202	0.944	-0.0024	-0.671	-0.528	0.094
Jeong-Mi (KOR)	3.897	0.367	0.391	-0.1523	0.425	0.373	-0.411

The correlation between the first principal component score and the official score is -0.99

Biplot

- Display the data points and the relationship among the original variables in a scatter plot of PCs

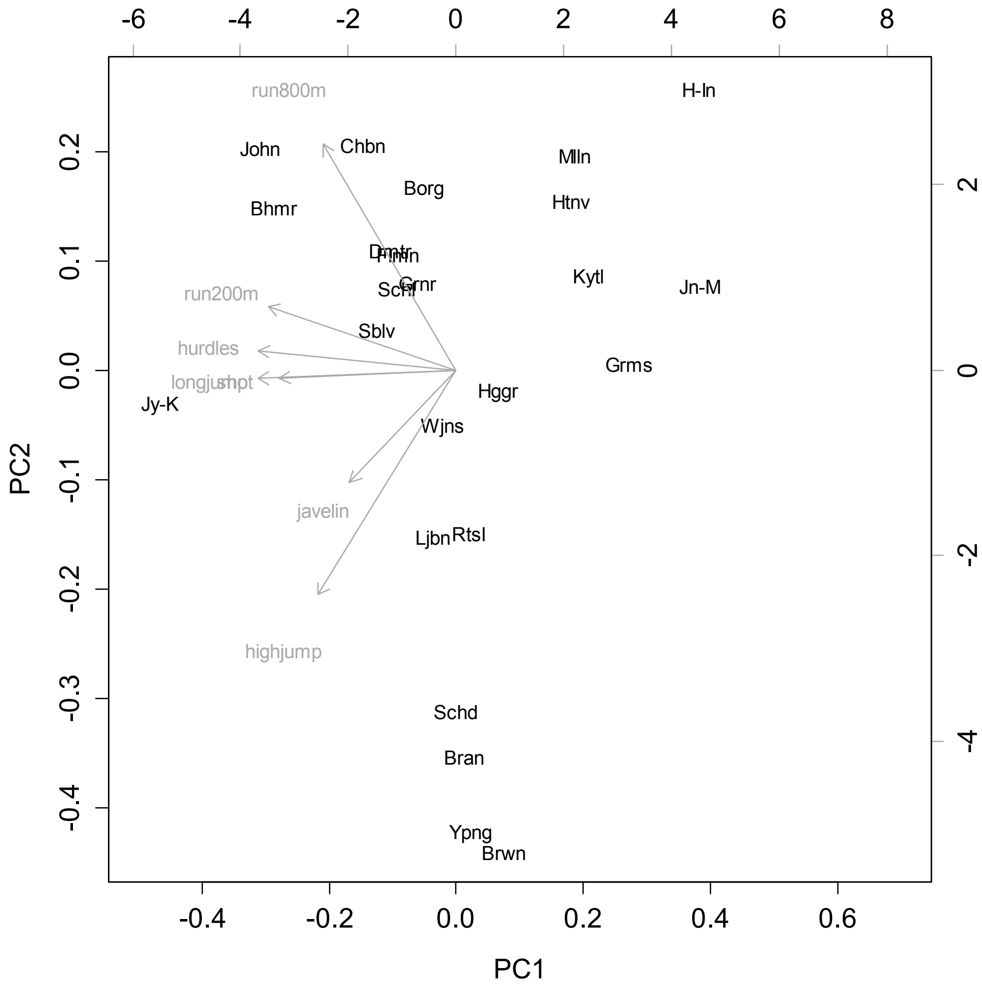
$$\frac{\sqrt{n}}{\sqrt{\lambda_i}} X\boldsymbol{e_i} \ vs. \frac{\sqrt{n}}{\sqrt{\lambda_j}} X\boldsymbol{e_j}$$

- Numbers in the plot
    - observation number
    - scatter plot of the first two principal components
- Red vectors

$$\frac{1}{\sqrt{n}}(\sqrt{\lambda_i}\boldsymbol{e_i}, \sqrt{j}\boldsymbol{e_j})$$

    - The correlations between the original variables and the principal components (PC loadings)
    - The length of the vector: variance of the original variable (all the same if correlation matrix is used)
    - The direction of the vector: if a variable is parallel to a PC, the variable has a big influence on the PC

- Joyner-Kersee has good records for hurdle, longjump, shot, run200
- run200m, hurdles, longjump, and shot are highly correlated
- Javelin and highjump are highly correlated
- PC1 largely separates the competitors by their overall scores
- PC2 indicates which are their best events




```
> tmp <- heptathlon[, -score]
> rownames(tmp) <- abbreviate(gsub(" \\(.*", "", rownames(tmp)))
> biplot(prcomp(tmp, scale = TRUE), col = c("black", "darkgray"), xlim = + c(-0.5, 0.7), cex = 0.7)
```

Graphing the Principal Components

- Use reduced information by principal components
- Check the normal assumption
  - Q-Q plot of each principal component
  - Scatter plots for pairs of the first few principal components

- Identify suspect observations (outliers)
  - Boxplot of each principal component
  - Scatter plots for pairs of the first few principal components