Multivariate Analysis (Slides 4)

- In today's class we examine examples of principal components analysis.
- We shall consider a difficulty in applying the analysis and consider a method for resolving this.

Recap of PCA

- Principal Components Analysis (PCA), as presented here, requires an eigenvalue analysis of the covariance matrix in order to find the linear combinations of the data variables with greatest variance.
- These linear combinations are called Principal Components (PCs).
- The PCs are constructed so that they are uncorrelated with each other.
- The eigenvector corresponding to the largest eigenvalue of the covariance matrix is called the 1^{st} PC.
- The eigenvector corresponding to the 2^{nd} largest eigenvalue of the covariance matrix is called the 2^{nd} PC, *etc*.
- As such, PC's have decreasing variances.
- The proportion of the variation in the data variables that is explained by a PC is equal to that component's associated eigenvalue divided by the sum of all eigenvalues.

Eigenvalue Interpretation

- Each eigenvalue of a covariance matrix can be interpreted as the variance of a linear combination of the variables with weights given by the eigenvector.
- Let **a** be a unit eigenvector corresponding to eigenvalue λ and recall that:

$$\mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \mathbf{a}^T\lambda\mathbf{a} = \lambda.$$

- **Fact:** The sum of the eigenvalues of a covariance matrix is equal to the sum of the variances of the variables in the data, *i.e.*, the trace of $\boldsymbol{\Sigma}$.
- Proof: Omitted
- Hence, the value of a particular eigenvalue divided by the total sum of the eigenvalues is the *proportion of the variance explained* by the associated principal component.

Uncorrelated Principal Components

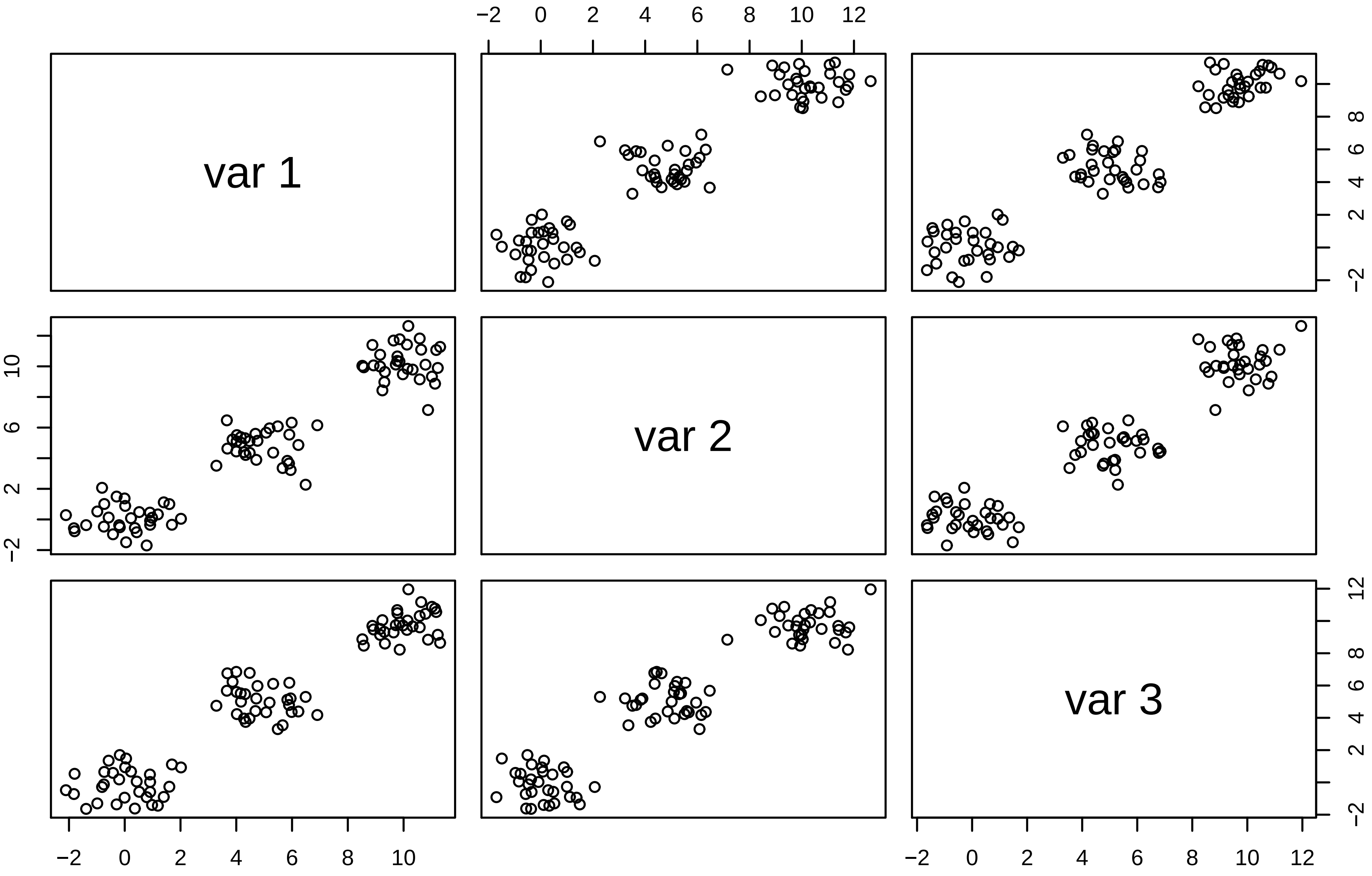
- In order to show that principal components must be uncorrelated consider the following.
- Let **a** and **b** be eigenvectors corresponding to eigenvalues λ_a and λ_b, respectively. Then:

$$\mathrm{Cov}(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{b} = \mathbf{a}^T\lambda_b\mathbf{b} = \lambda_b\mathbf{a}^T\mathbf{b}$$

$$\mathrm{Cov}(\mathbf{b}^T\mathbf{X}, \mathbf{a}^T\mathbf{X}) = \mathbf{b}^T\boldsymbol{\Sigma}\mathbf{a} = \mathbf{b}^T\lambda_a\mathbf{a} = \lambda_a\mathbf{b}^T\mathbf{a}$$

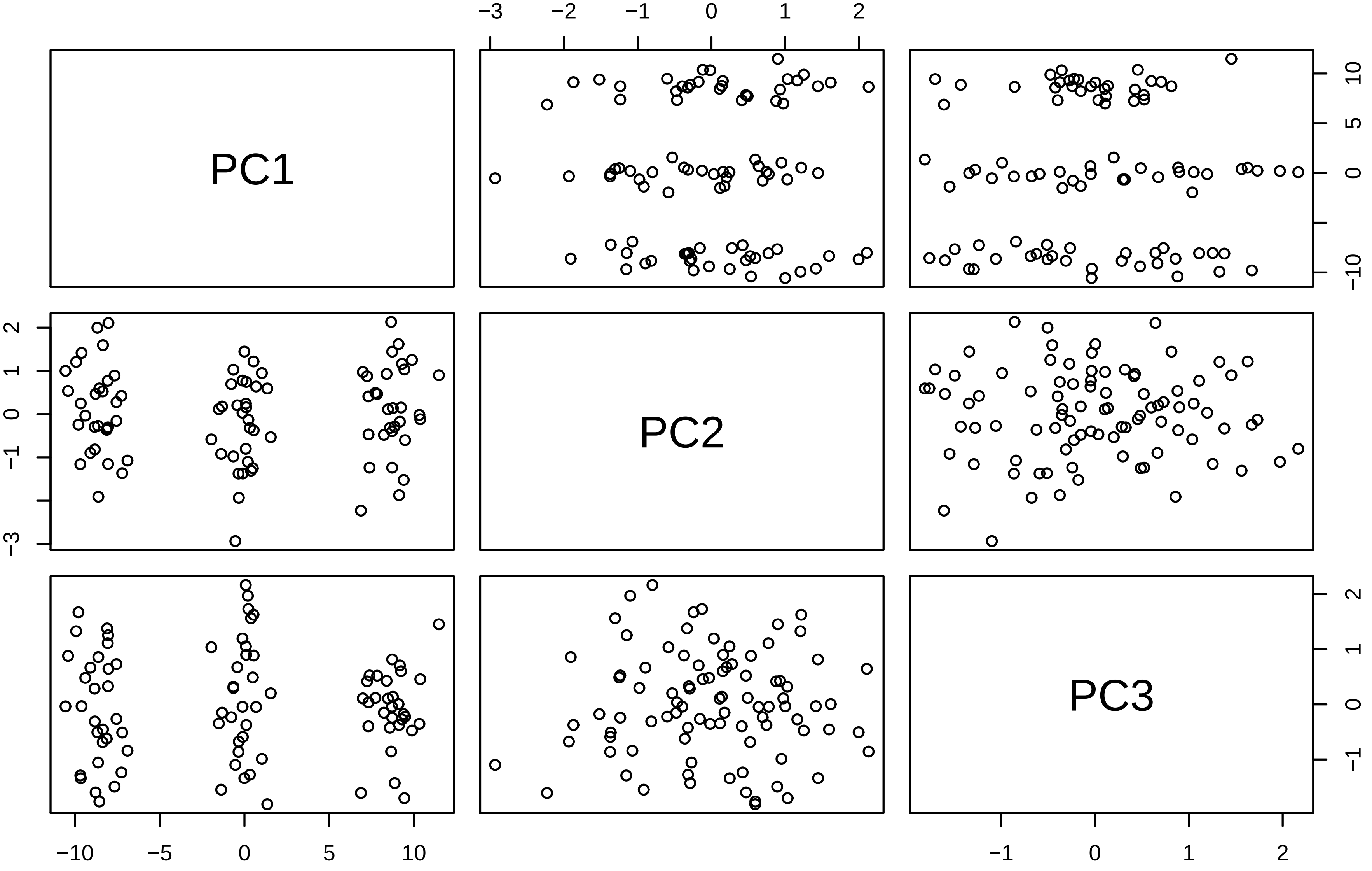
- Hence $\lambda_a = \lambda_b$, or $\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a} = 0$.
- In the former the eigenvectors are the same.
- In the latter, the Covariance is 0.

Example: 3D Data



PCA: 3D Data

- The first principal component picks up the group structure.

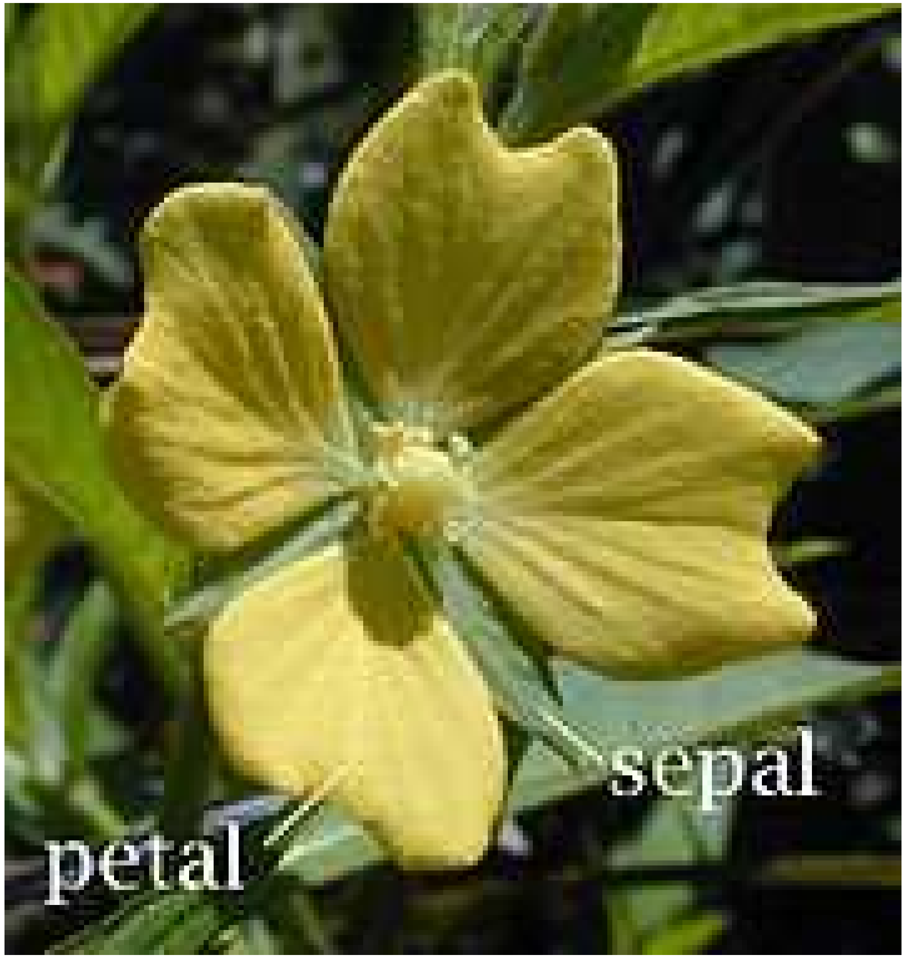


Fisher's Iris Data

- This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.
- The species are Iris setosa, versicolor, and virginica.

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
...............................................................
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
...............................................................
101          6.3         3.3          6.0         2.5  virginica
102          5.8         2.7          5.1         1.9  virginica
103          7.1         3.0          5.9         2.1  virginica
```

Fisher’s Iris Data



View of a Flower’s Petal and Sepal

Fisher's Iris Data

- PCA can be used to reduce the dimension of the iris data set. PCA was applied to the data with results as shown below.

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	2.056	0.4926	0.2797	0.15439
Proportion of Variance	0.925	0.0531	0.0171	0.00521
Cumulative Proportion	0.925	0.9777	0.9948	1.00000

Principal components

	PC1	PC2	PC3	PC4
Sepal.Length	0.36	-0.66	0.58	0.32
Sepal.Width	-0.08	-0.73	-0.60	-0.32
Petal.Length	0.86	0.17	-0.08	-0.48
Petal.Width	0.36	0.08	-0.55	0.75

How do we interpret PCA output?

- First examine the ‘Importance of components’ section.
- The first row in the table shows the standard deviation of each PC (NB the standard deviation is the square root of the eigenvalue associated with that component).
- The second row gives the proportion of the variation in the data accounted for by each PC. Note that this is equal to: $\frac{\mathrm{sd}^2}{\sum \mathrm{sd}^2}$
- The third row gives the cumulative proportion of the variance accounted for by the PCs.
- An indication of the number of PCs required to adequately summarize the data can be inferred by examining the proportion of the variation explained by the PCs. In this case, the first PC accounts for 93% of the variation in the data.
- The second principal component only accounts for an additional 5% of the variation. This means that the multivariate structure of the data is essentially just one-dimensional.

How do we interpret PCA output?

- Consider now the ‘new’ variables (PCs) which capture the variation in the data. We focus on interpreting PC1.
- The column labeled PC1 is the eigenvector of the data covariance matrix associated with the largest eigenvalue.
- It’s elements are the coefficients or *loadings* of each original variable on the first PC.
- It matters if the loadings have opposite signs, but not which is positive and which is negative. The magnitudes of the loadings are also important.
- In PC1, Sepal.length, Petal.Length and Petal.Width all have large positive loadings, whereas Sepal.Width has a negative loading which is close to zero.
- Thus we could interpret the first PC as an overall measure of the area of the flower visible from above (sepal width not important).
- PC2 then appears to contrast flowers who have relatively large sepals and small petals, against those with small sepals and large petals.

A Problem

- In the Iris data example the flower measurements were all made in centimeters. Would the results be different if sepal dimensions were recorded in mm and petal dimensions in cm, for example?

- Unfortunately, yes. As PCA seeks to maximize variance it can be sensitive to scale differences across variables.

- Including a variable such as sepal.length in mm (where the range of values will be from 43 to 79) will have a much bigger impact on the variance of the linear combination/PC than the variable petal.length does if measured in cm (where the range of values will be from 1.0 to 6.9), regardless of the pattern of covariation between variables.

- Standardizing ensures that the data are expressed in comparable units.

A Solution

- We want to use units of measurement for each variable that are comparable.
- One possibility is to make each variable have variance equal to one.
- To do this, divide the values for each variable by the sample standard deviation.
- We could then do a principal components analysis on the covariance matrix for the transformed data.
- **Fact:** The covariance matrix of a set of variables with variance one, is a correlation matrix.

Fisher's Iris Data re-done

- Dividing each variable by the sample standard deviation and performing PCA on the standardized data yields:

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.71	0.956	0.3831	0.14393
Proportion of Variance	0.73	0.229	0.0367	0.00518
Cumulative Proportion	0.73	0.958	0.9948	1.00000

Principal components

	PC1	PC2	PC3	PC4
Sepal.Length	0.52	-0.38	0.72	0.26
Sepal.Width	-0.27	-0.92	-0.24	-0.12
Petal.Length	0.58	-0.02	-0.14	-0.80
Petal.Width	0.56	-0.07	-0.63	0.52

2003 Heptathlon Results

```
   Name             100m HighJump ShotPutt  200m LongJump Javelin  800m Points
 1 Dufour          14.08     1.67    13.34 25.24     5.62   39.83 132.8   5723
 2 Klueft          13.18     1.94    14.19 22.98     6.68   49.90 132.1   7001
 3 Butor           13.92     1.79    12.51 24.67     5.86   46.43 136.7   6035
 4 Sazanovich      13.67     1.76    16.81 24.25     6.47   44.93 136.5   6524
 5 Netseporuk      13.91     1.76    13.97 24.96     6.05   50.05 139.8   6154
 6 Barber          13.05     1.91    12.97 23.92     6.61   49.60 133.7   6755
 7 Hollman         14.10     1.85    12.05 24.72     6.06   41.01 135.8   6018
 8 Lewis           13.37     1.64    15.25 24.55     6.19   49.88 139.6   6254
 9 Kesselschlaeger 13.34     1.76    13.77 24.94     6.17   41.57 137.2   6134
10 Strataki        13.93     1.79    13.34 24.69     6.03   44.27 138.1   6077
11 Bacher          14.01     1.76    13.32 24.91     5.99   47.11 129.8   6166
12 Kazanina        14.44     1.73    13.23 25.06     5.91   50.17 132.4   6047
13 Naumenko        14.20     1.79    12.88 24.98     6.00   42.15 135.4   5971
14 Klavina         13.92     1.76    14.24 24.37     6.07   41.17 149.6   5932
15 Skujyte         14.44     1.76    16.35 25.76     5.86   47.57 138.6   6077
16 Chernyavskaya   13.85     1.76    12.76 25.03     5.99   37.83 129.4   5969
17 Prokhorova      13.87     1.82    13.36 23.99     6.46   43.60 128.3   6452
18 Roshchupkina    14.32     1.70    14.53 24.21     5.85   43.22 133.0   6034
```

Example: Heptathlon Data

- An eigenvalue analysis of the heptathlon data gives:

```
Eigenvalues
24.66 15.94  1.29  0.48  0.08  0.02  0.00
Eigenvectors
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,]  0.00  0.02 -0.02 -0.45  0.83 -0.32 -0.08
[2,]  0.00  0.00  0.02  0.06  0.01 -0.29  0.95
[3,]  0.09 -0.09 -0.99  0.01  0.00  0.03  0.03
[4,]  0.01  0.04 -0.01 -0.83 -0.53 -0.20 -0.01
[5,] -0.01 -0.03 -0.03  0.33 -0.18 -0.88 -0.29
[6,]  0.07 -0.99  0.09 -0.05  0.00  0.01  0.00
[7,]  0.99  0.07  0.08  0.02  0.01 -0.01  0.00
```

PCA: Rescaled Heptathlon Data

- In this example race times are recorded in seconds and jumps or throws are recorded in meters.

- Again the results would be different if different units were used.

- If race times are recorded in minutes, we get the following:

```
Eigenvalues
15.95  1.45  0.07  0.01  0.00  0.00  0.00
Eigenvectors
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,]  0.00  0.00 -0.02  0.01  0.04 -0.01  1.00
[2,]  0.00 -0.02  0.19 -0.07  0.98  0.00 -0.04
[3,] -0.10  0.99 -0.01  0.02  0.03  0.00  0.00
[4,]  0.00  0.00 -0.03  0.00  0.01 -1.00 -0.01
[5,] -0.03  0.02  0.98 -0.04 -0.19 -0.03  0.03
[6,] -0.99 -0.10 -0.02  0.00  0.00  0.00  0.00
[7,]  0.00  0.02 -0.05 -1.00 -0.06  0.00  0.01
```

Covariance: Rescaled Heptathlon Data

- The covariance matrix for the transformed data is:

```
           X100m HighJump ShotPutt    X200m LongJump  Javelin    X800m
X100m    0.00005 -0.00021 -0.00035  0.00004 -0.00148 -0.00649 -0.00001
HighJum -0.00021  0.00543 -0.02851 -0.00044  0.01404  0.05244 -0.00116
ShotPut -0.00035 -0.02851  1.59275 -0.00015  0.06286  1.41576  0.03350
X200m    0.00004 -0.00044 -0.00015  0.00011 -0.00231 -0.00992  0.00011
LongJum -0.00148  0.01404  0.06286 -0.00231  0.08024  0.39736 -0.00262
Javeli  -0.00649  0.05244  1.41576 -0.00992  0.39736 15.79961  0.00777
X800m   -0.00001 -0.00116  0.03350  0.00011 -0.00262  0.00777  0.00679
```

- Now the javelin score has the highest variance!

Correlation: Heptathlon Data

- The correlation matrix for the heptathlon data is:

```
         X100m HighJump ShotPutt X200m LongJump Javelin X800m
X100m     1.00    -0.42    -0.04  0.63    -0.77   -0.24 -0.02
HighJump -0.42     1.00    -0.31 -0.59     0.67    0.18 -0.19
ShotPutt -0.04    -0.31     1.00 -0.01     0.18    0.28  0.32
X200m     0.63    -0.59    -0.01  1.00    -0.79   -0.24  0.13
LongJump -0.77     0.67     0.18 -0.79     1.00    0.35 -0.11
Javelin  -0.24     0.18     0.28 -0.24     0.35    1.00  0.02
X800m    -0.02    -0.19     0.32  0.13    -0.11    0.02  1.00
```

- Now all variables have the same variance.

PCA: Heptathlon Data

- The eigenvalue analysis of the correlation matrix is:

```
Eigenvalues
3.10 1.55 0.87 0.66 0.45 0.30 0.08
Eigenvectors
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,]  0.46 -0.09 -0.26  0.17  0.69 -0.29  0.35
[2,] -0.43 -0.30  0.08  0.39  0.53  0.39 -0.37
[3,] -0.01  0.70 -0.14 -0.46  0.39  0.21 -0.29
[4,]  0.50  0.03 -0.10  0.17 -0.16  0.81  0.18
[5,] -0.54  0.09  0.04 -0.16  0.12  0.21  0.78
[6,] -0.24  0.37 -0.67  0.54 -0.22 -0.10  0.00
[7,]  0.09  0.52  0.66  0.51  0.05 -0.11  0.10
```

- Better for basing interpretation on.

PCA: Heptathlon Data

- The principal components are difficult to interpret because, whilst a small running time is good, so is a large throwing or jumping distance.
- Suggestion to replace *time* by -*time* and repeat analysis.

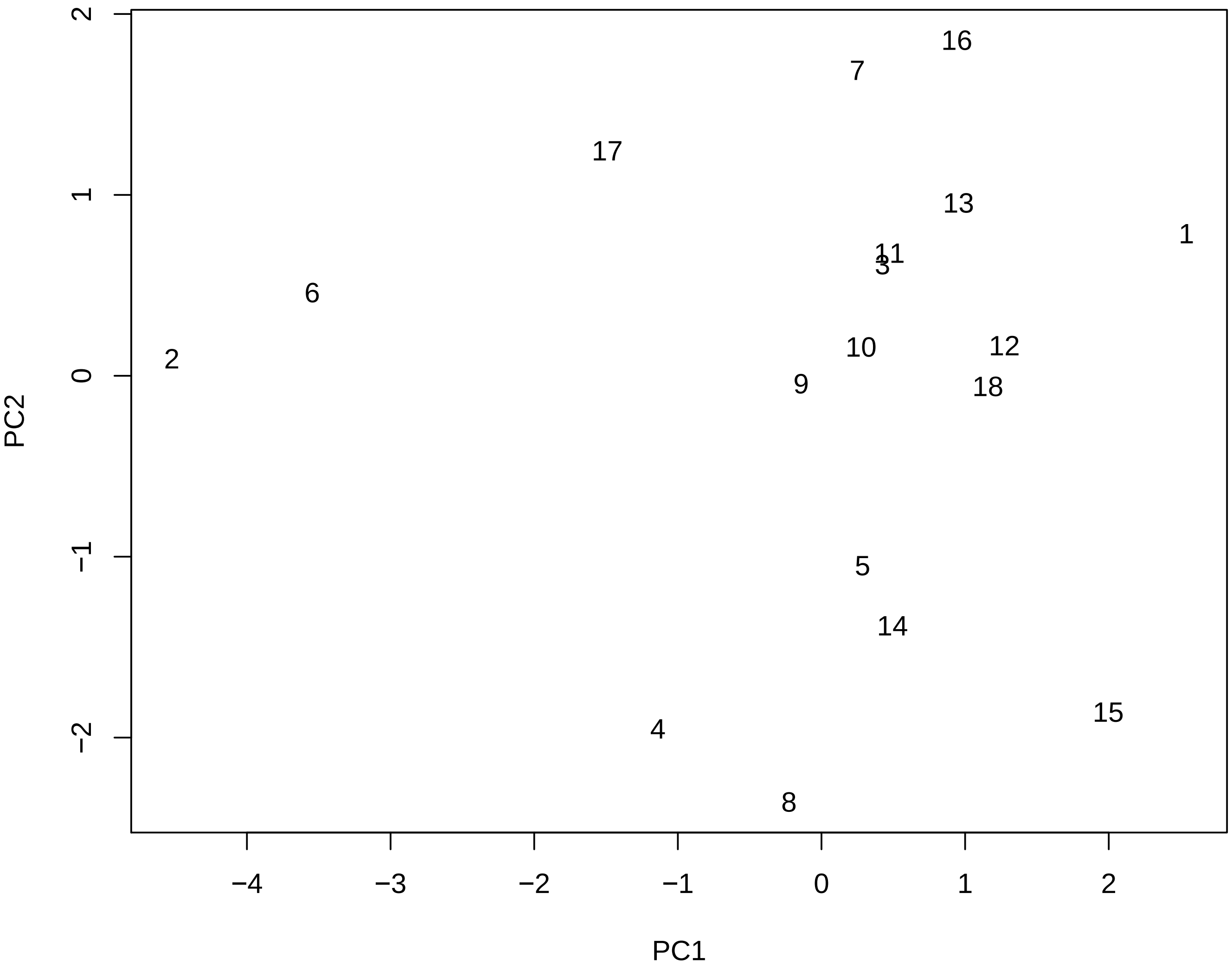
```
Variable     PC1     PC2     PC3     PC4     PC5     PC6     PC7
100m      -0.460  -0.095  -0.264   0.168  -0.692   0.286   0.348
HighJump  -0.431   0.297  -0.076  -0.393   0.527   0.389   0.369
ShotPutt  -0.012  -0.697   0.141   0.462   0.392   0.208   0.288
200m      -0.495   0.026  -0.097   0.168   0.161  -0.810   0.184
LongJump  -0.540  -0.088  -0.037   0.160   0.116   0.210  -0.785
Javelin   -0.238  -0.371   0.673  -0.542  -0.219  -0.103   0.004
800m      -0.091   0.520   0.664   0.506  -0.046   0.111   0.100

Eigenv    3.099   1.5482  0.8667  0.6570  0.4533  0.2999  0.0754
Propor     0.433   0.221   0.124   0.094   0.065   0.043   0.011
Cumula     0.443   0.664   0.788   0.882   0.946   0.989   1.000
```

- What does this tell us?

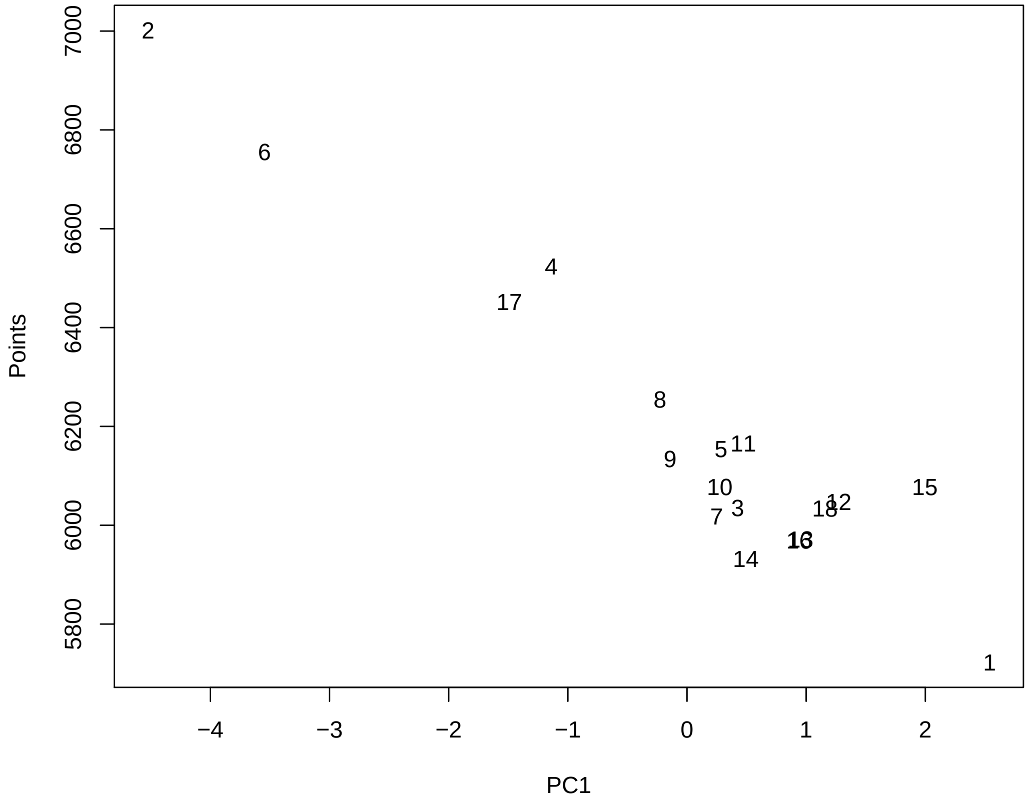
Plot of Principal Components

- We can plot the first two principal components to look for structures in the data.



PCA: Heptathlon Data

- The first principal component records some measure of overall ability.
- Note that the good athletes are on the left-hand side of the plot and the points tend to drop as we move right.

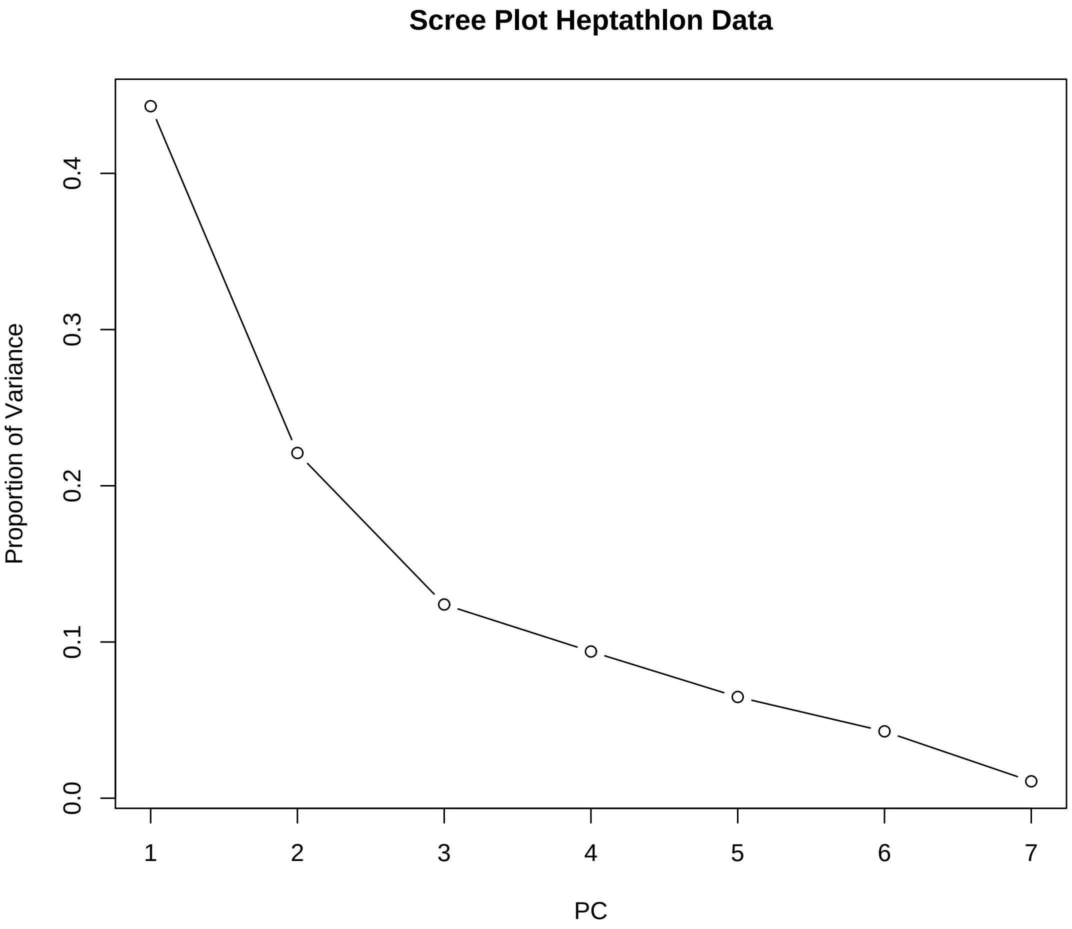


- How can we interpret the second principal component?

PCA: Heptathlon Data

- How to choose the appropriate number of principal components?
- No correct answer but there are rules of thumb:
  - Keep adding until a fixed proportion of variance is included.
  - Find a kink in a *Scree* plot (variance explained against principal component number). This means that the marginal additional variance explained is reducing as a function of principal component, *i.e.*, the added benefit of including an additional principal component may no longer be worth the extra cost of model complexity (remember we are seeking dimension reduction).

PCA: Heptathlon Data



What would be the appropriate number of principal components to include here?