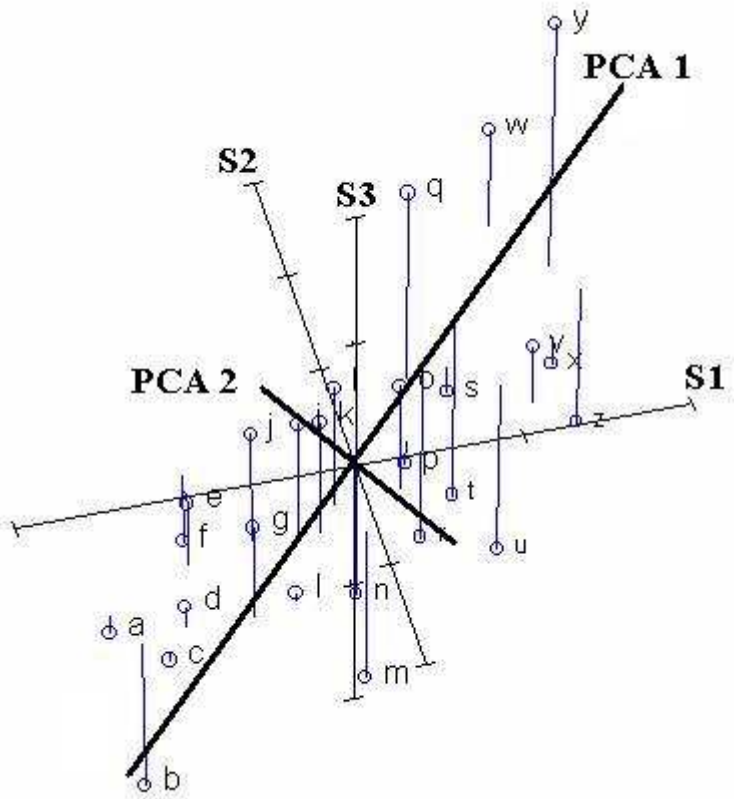
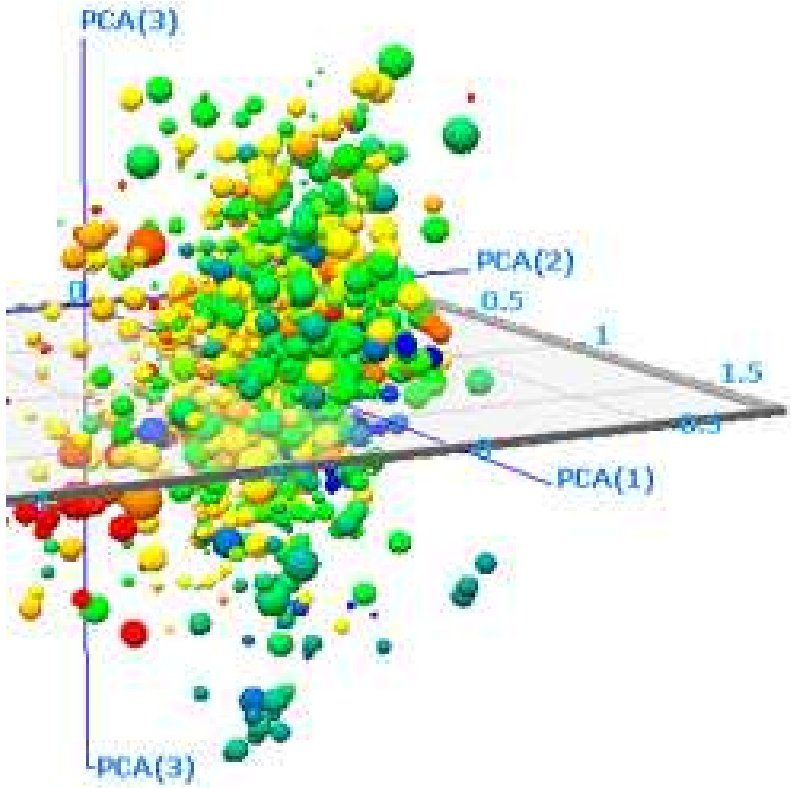
Principal Component Analysis





What is PCA?

- PCA is a variable reduction procedure
- Can be applied to a dataset with a large number of variables if they contain redundancy
- Redundancy occurs when separate variables are measuring the same construct
- The technique reduces the observed variables to artificial variables called **principal components**

Why use PCA?

- Can be used to simplify a complicated dataset
- Can be used to understand the structure of a complicated dataset
- Can also be used to interpret a simple dataset

The Work Questionnaire

1. My supervisor treats me with consideration
2. My supervisor consults me concerning important decisions that affect my work
3. My supervisor gives me recognition when I do a good job
4. My supervisor gives me the support I need to do my job well
5. My pay is fair
6. My pay is appropriate, given the amount of responsibility that comes with my job
7. My pay is comparable to the pay earned by other employees whose jobs are similar to mine

Rate 1-7: Strongly agree to strongly disagree

The Work Questionnaire

1. **My supervisor treats me with consideration**
2. **My supervisor consults me concerning important decisions that affect my work**
3. **My supervisor gives me recognition when I do a good job**
4. **My supervisor gives me the support I need to do my job well**
5. My pay is fair
6. My pay is appropriate, given the amount of responsibility that comes with my job
7. My pay is comparable to the pay earned by other employees whose jobs are similar to mine

Rate 1-7: Strongly agree to strongly disagree

The Work Questionnaire

1. My supervisor treats me with consideration
2. My supervisor consults me concerning important decisions that affect my work
3. My supervisor gives me recognition when I do a good job
4. My supervisor gives me the support I need to do my job well
5. **My pay is fair**
6. **My pay is appropriate, given the amount of responsibility that comes with my job**
7. **My pay is comparable to the pay earned by other employees whose jobs are similar to mine**

Rate 1-7: Strongly agree to strongly disagree

Correlations among Seven Job Satisfaction Items

			Correlations				
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1.00						
2	0.75	1.00					
3	0.83	0.82	1.00				
4	0.68	0.92	0.88	1.00			
5	0.03	0.01	0.04	0.01	1.00		
6	0.03	0.02	0.05	0.07	0.89	1.00	
7	0.03	0.06	0.00	0.03	0.91	0.76	1.00

Correlations among Seven Job Satisfaction Items

			Correlations				
Variable	**1**	**2**	**3**	**4**	5	6	7
1	**1.00**						
2	**0.75**	**1.00**					
3	**0.83**	**0.82**	**1.00**				
4	**0.68**	**0.92**	**0.88**	**1.00**			
5	0.03	0.01	0.04	0.01	1.00		
6	0.03	0.02	0.05	0.07	0.89	1.00	
7	0.03	0.06	0.00	0.03	0.91	0.76	1.00

Variables 1-4: Questions relating to your supervisor

Correlations among Seven Job Satisfaction Items

			Correlations				
Variable	1	2	3	4	**5**	**6**	**7**
1	1.00						
2	0.75	1.00					
3	0.83	0.82	1.00				
4	0.68	0.92	0.88	1.00			
5	0.03	0.01	0.04	0.01	**1.00**		
6	0.03	0.02	0.05	0.07	**0.89**	**1.00**	
7	0.03	0.06	0.00	0.03	**0.91**	**0.76**	**1.00**

Variables 5-7: Questions relating to your pay

What is a Principal Component?

- Principal Component is defined as a linear combination of optimally-weighted observed variables
- There are as many Principal Components as there are variables
- Principal Components account for the total variation in the dataset
- The first principal component will account for the most variation, the last will account for the least.

1st Principal Component

- The first principal component will capture the maximal account of total variance in the observed variables

- Under typical conditions, the first component will be correlated with some or many of the observed variables

2nd Principal Component

- Account for a maximal account of variance in the data set that was not accounted for by the first component
- Must be uncorrelated with the first component
- This process continues until the remaining principal components capture very little of the variation

PCA formula

$$c_1 = b_{11}(X_1) + b_{12}(X_2) + \ldots b_{1p}(X_p)$$

where

- c_1 = the subject's score on principal component 1
- b_{1p} = the regression coefficient (or weight) for observed variable *p*, as used in creating principal component 1
- x_p = the subject's score on observed variable *p*

1st Principal Component (Supervisor)

- $c_1 = .44(x_1) + .40(x_2) + .47(x_3) + .32(x_4) + .02(x_5) + .01(x_6) + .03(x_7)$

2nd Principal Component (Pay)

- $c_2 = .01(x_1) + .04(x_2) + .02(x_3) + .02(x_4) + .48(x_5) + .31(x_6) + .39(x_7)$

Weighting of the Principal Components

- Weights applied by using a special type of equation called an **eigenequation**.

- The weights produced by these eigenequations are optimal weights in the sense that, for a given set of data, no other set of weights could produce a better representation of the data

- A set of components that are the most successful in accounting for variance in the dataset

Variance

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{(n-1)}}$$

- Standard Deviation

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{(n-1)}$$

- Variance

Covariance

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)\left(X_i - \bar{X} \right)}{(n-1)}$$

- Variance

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)}{(n-1)}$$

- Covariance

Covariance Matrix

$$C^{mn} = \left(c_{i,j}, c_{i,j}, = \text{cov}\left(Dim_i, Dim_j\right)\right),$$

For a dataset with three dimensions

$$C = \begin{pmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{pmatrix}$$

Eigenanalysis

- Eigenanalysis is a mathematical operation on a square, symmetric matrix.
- A square matrix has the same number of rows as columns.
- A symmetric matrix is the same if you switch rows and columns.
- The answer to an eigenanalysis consists of a series of eigenvalues and eigenvectors.

Eigenvectors and Eigenvalues

- The eigenvectors of a square matrix are the non-zero vectors that, after being multiplied by the matrix, remain proportional to the original vector (i.e., change only in magnitude, not in direction).

- For each eigenvector, the corresponding eigenvalue is the factor by which the eigenvector changes when multiplied by the matrix

- This can be expressed as:

$$Av = \lambda v$$

if A is a square matrix then a non-zero vector v is an eigenvector of A if there is a scalar λ

if there is a scalar λ it is said to be the eigenvalue of A corresponding to v.

Eigenvectors and Eigenvalues

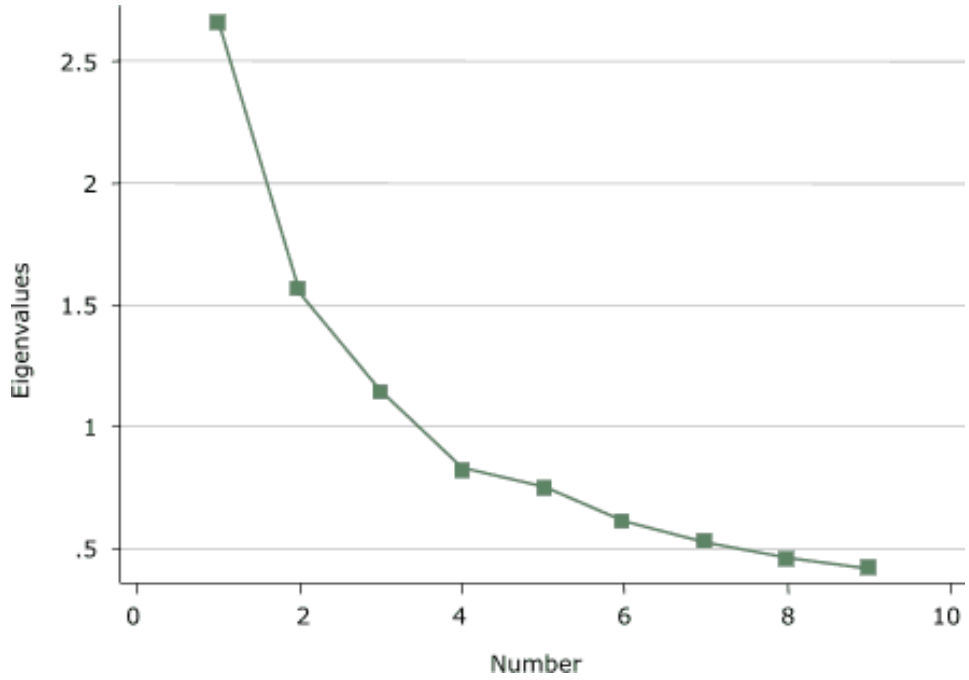
- Each eigenvalue has an eigenvector
- There are as many eigenvectors and eigenvalues as there are rows in the initial matrix
- The eigenvalue is a measure of the strength of an axis and the amount of variation along that axis
- Eigenvalues are usually ranked from the greatest to the least
- The first eigenvalue is often called the "dominant" or "leading" eigenvalue – Relates back to Principal Components

Eigenanalysis

- It is possible to perform a eigenanalyses analytically (that is, get exact results) only for very small matrices (e.g. three rows and columns).

- For large matrices, eigenanalysis requires an iterative approach which eventually "closes in" on the answer (in most cases).

Scree Plot

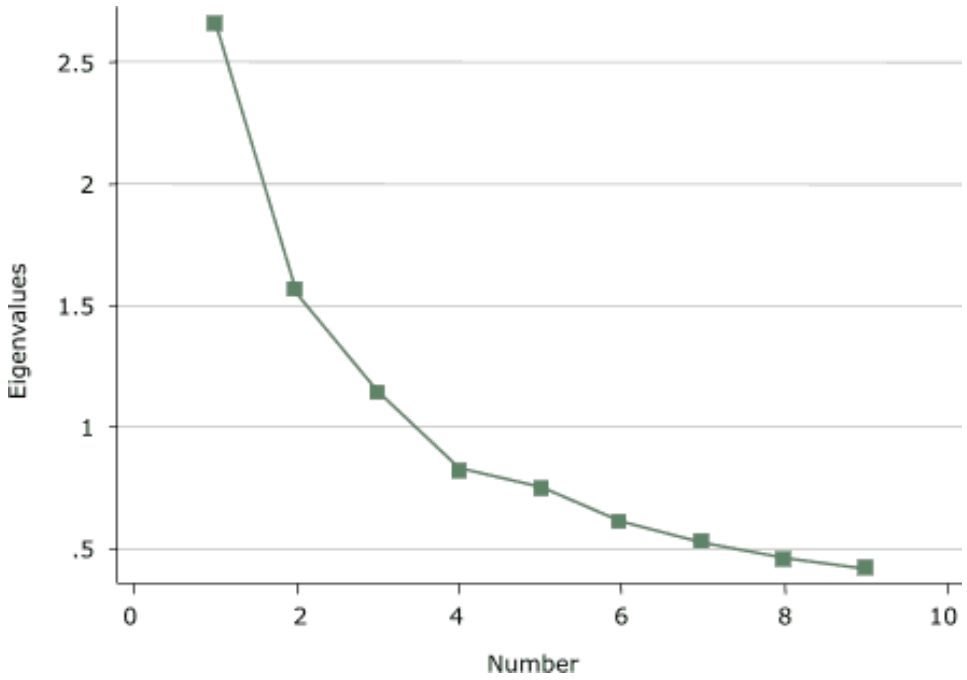


Shows the decreasing rate at which variance is explained by additional principal components

Scree Plot

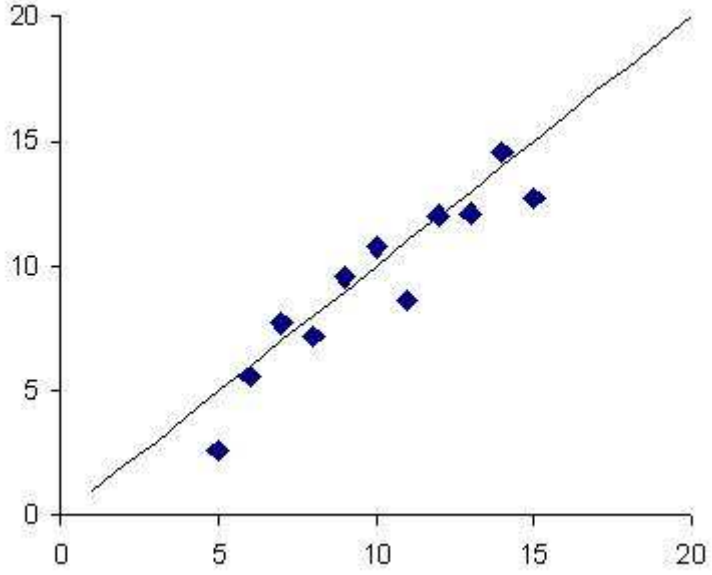
- ‘Scree’ refers to the loose rubble that lies at the base of the cliff
- Idealised cut-off point for principal components

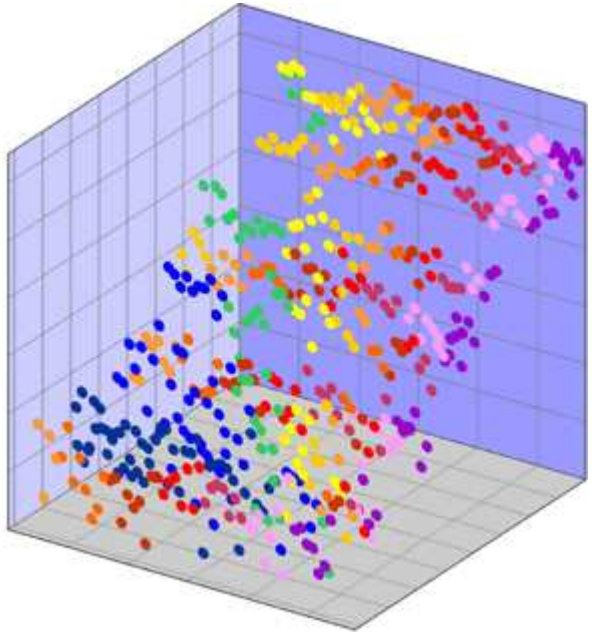
How many principal components?



- Scree drop off
- Kaiser Criterion
- % Variance

Co-ordinate systems





2D graphs

3D graphs

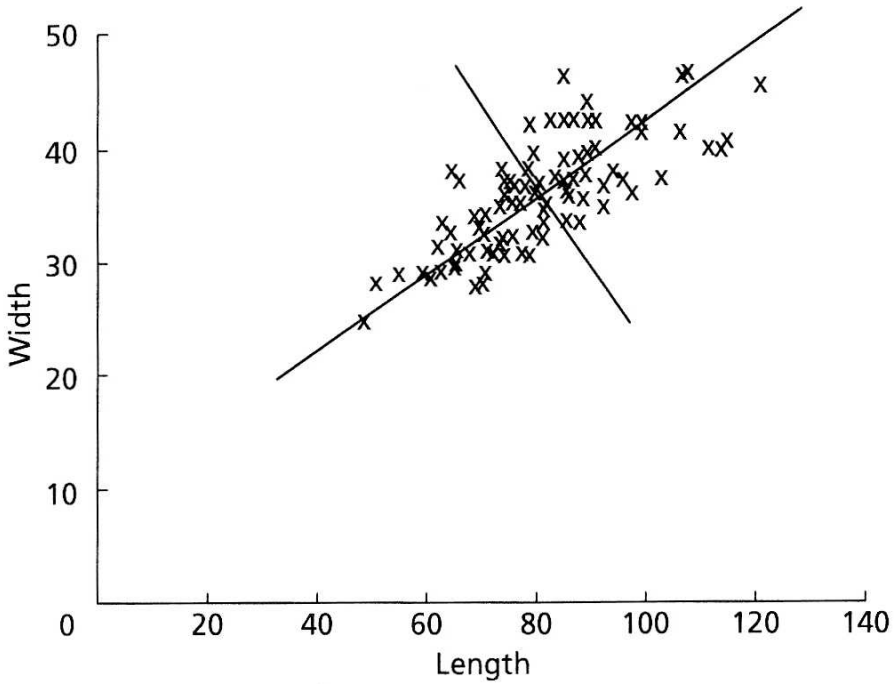
Multidimensional Co-ordinate systems

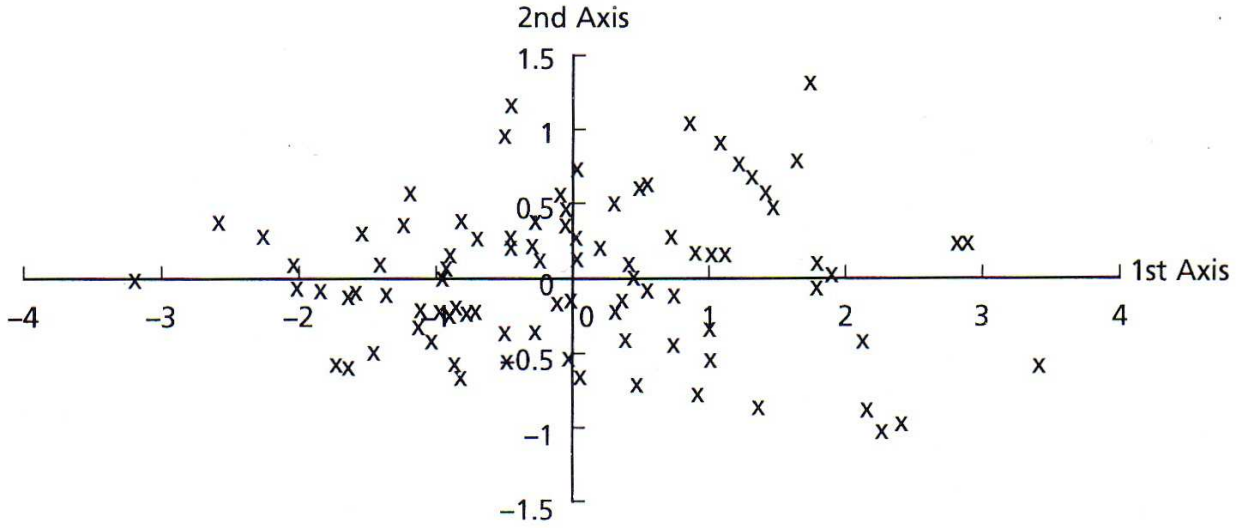
- Can not be visualised

BUT

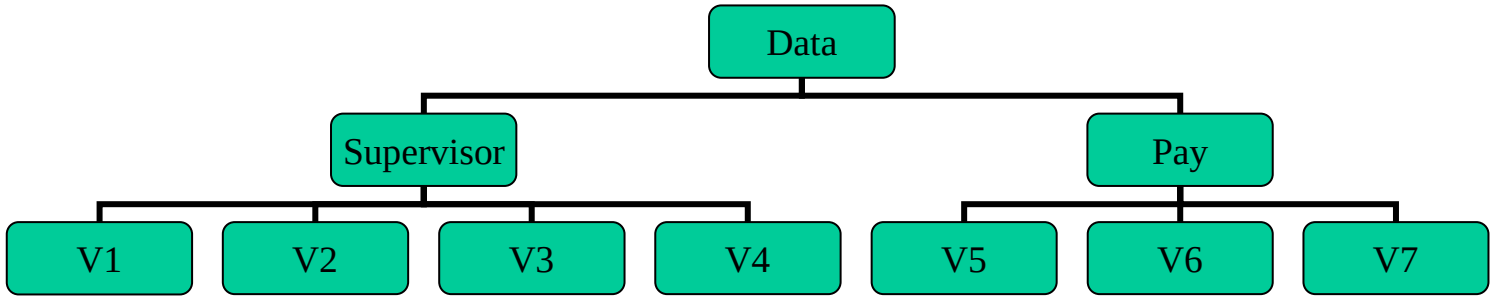
- Can be transformed into a different co-ordinate system
- Can have a preferred co-ordinate system **defined by the data itself**

Rotation of Axes





PCA is not Factor Analysis



- Factor Analysis (FA) relies on the assumption of an underlying causal structure
- FA assumes that co-variation in the observed variables is due to the presence of latent variables
- PCA makes no such assumptions

PCA for data interpretation

- The Mercer & Hall data set
- Measures two variables: grain and straw yield.
- Measured quantities are the outcomes of processes which we can not directly observe:
  (1) plant growth;
  (2) partition of plant growth between grain and straw
- PCA can be used to gain insight into these

Standardisation of Values

- PCA can be standardized or not. In the first case the variables are scaled to zero mean and unit variance, thus giving them equal importance mathematically.

- In the second, the original units are used; thus variables with more variance are more important. In the present case, although the two variables are computed in the same units of measure, they are of intrinsically different magnitude (there is much more straw than grain)

- To reveal the relation between the variables they must be standardised before computing the components

prcomp() function

- The prcomp function computes the PCA

- The variables are first scaled, resulting in standardized PCA

```
mhw <-
    read.csv('http://rcourse.iop.kcl.ac.uk/2011/4thu/session2/mhw.
    csv', header=T, sep=",")
pc <- prcomp(mhw[, c("grain", "straw")], scale = T)
summary(pc)
biplot(pc)
```

How to interpret the biplot

- **Orientation**: The direction of the vector, with respect to the PC space. The more a vector, which represents an original variable, is parallel to a PC axis, the more it contributes to that PC.

- **Length**: Length in the space defined by the displayed PCs; the longer the vector, the more variability of this variable is represented by the two displayed PCs.

- **Angles**: Angles between vectors of different variables show their correlation in the space spanned by the two displayed PC's:

  Small angles represent high positive correlation
  Right angles represent lack of correlation
  Opposite angles represent high negative correlation

Accounting for Variability

- First PC, accounting for about 85% of the variance in both grain and straw yield, represents the overall yield level
- Second PC, accounting for about 15% of the variance, represents the grain/straw ratio, i.e., plant morphology independent of yield.
- The great majority of the overall variability in this field is due to variable yield. However, there is still a fair amount of variation in plant morphology that does not depend on overall plant size
- Since this is one variety of wheat (i.e., no genetic variation) and one management, we conclude that local enviromental factors affect not only yield but also plant morphology

Another example of PCA

```
ph <- read.delim('http://rcourse.iop.kcl.ac.uk/S11/pheno.csv',
    sep=',', row.names=1 )
ph<-ph[apply(ph,1,function(x)!any(is.na(x))),]
pc2 <- prcomp(ph, scale.=T)
plot(pc2)
biplot(pc2)

heatmap(as.matrix(ph))

heatmap(scale(ph)
```

Visualising PCA

```
pc2
str(pc2)
loadings(pc2)
varimax(pc2$rotation[,1:2])
pc2$rotation
heatmap(pc2$rot)
heatmap(pc2$x)
```

Visualising PCA

```
# plot pairs of loading columns (=eigenvectors)
plot(PC1 ~ PC2 , data=pc2$r)
plot(PC2 ~ PC3 , data=pc2$r)

# another convenient way
with(data.frame(pc2$r), plot(PC1,PC2))
with(data.frame(pc2$r), text(PC1,PC2,rownames(pc2$r),pos=3, cex=0.6))
```

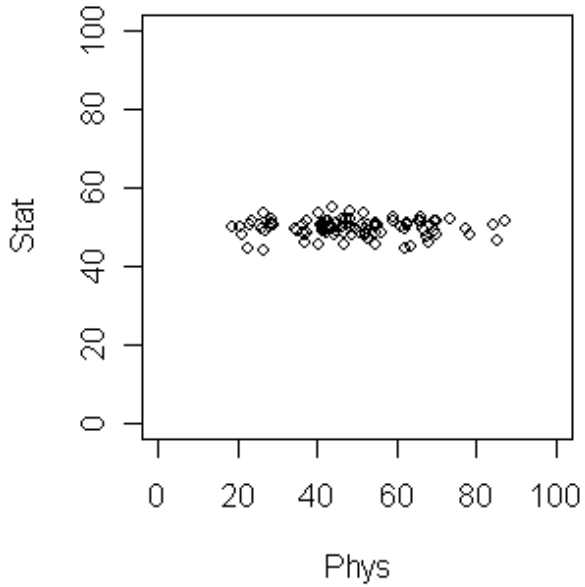
Problem 1

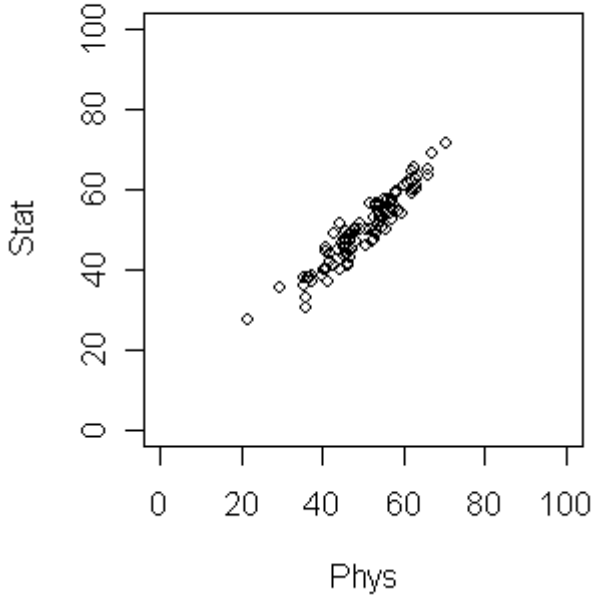
Dataset showing incidence of crimes and percentage of urban population in 1973 America

```
library(stats)
?USArrests
```

What is the relationship between the two?

Problem 2





Which grade, if any, is a better discriminating factor?

```
PS <- read.table('http://rcourse.iop.kcl.ac.uk/2011/4thu/session2/PS.txt', header=T)
```