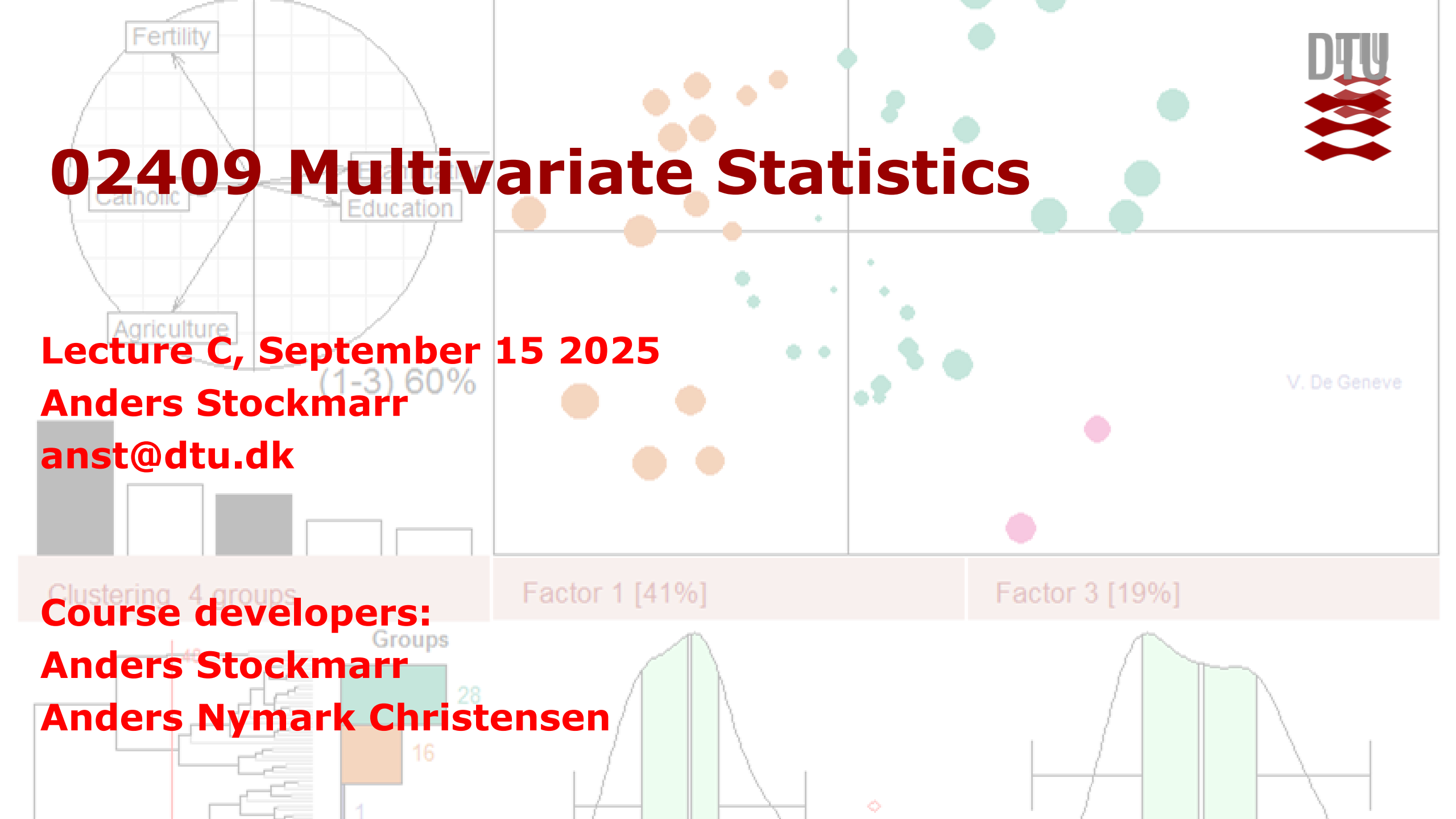



# 02409 Multivariate Statistics

**Lecture C, September 15 2025**

**Anders Stockmarr**

**anst@dtu.dk**

**Course developers:**

**Anders Stockmarr**

**Anders Nymark Christensen**

# Agenda

- Some examples
- A brief recap on estimation
- Partial correlation - examples
- Multiple correlation - examples
- PCA

# Example - Exercise 2.4

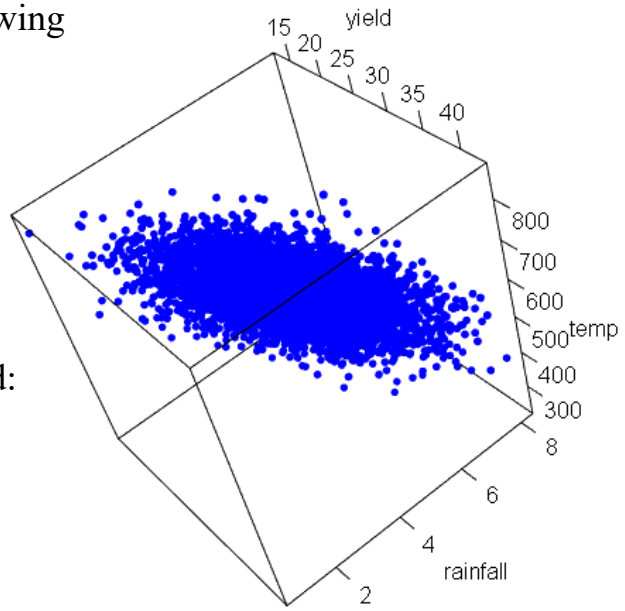
In an investigation on the relationship between weather and crops the following variables were measured:

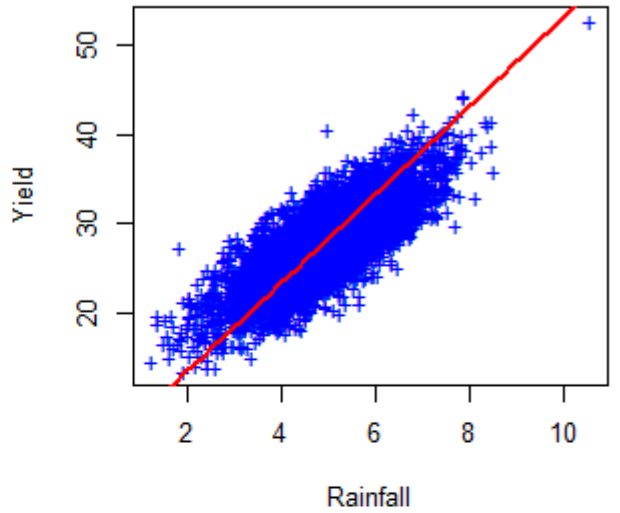
- $X_1$: Yield of hay in cwt per acre
- $X_2$: Spring rainfall in inches
- $X_3$:Accumulated temperature above 42˚F in the spring

Based on observations from 20 years the following estimates were obtained:

- $\hat{\mu}_1 = 28.02$ $\hat{\sigma}_1 = 4.42$ $\hat{\rho}_{12} = 0.80$
- $\hat{\mu}_2 = 4.91$ $\hat{\sigma}_2 = 1.10$ $\hat{\rho}_{13} = -0.40$
- $\hat{\mu}_3 = 594$ $\hat{\sigma}_3 = 85$ $\hat{\rho}_{23} = -0.56$

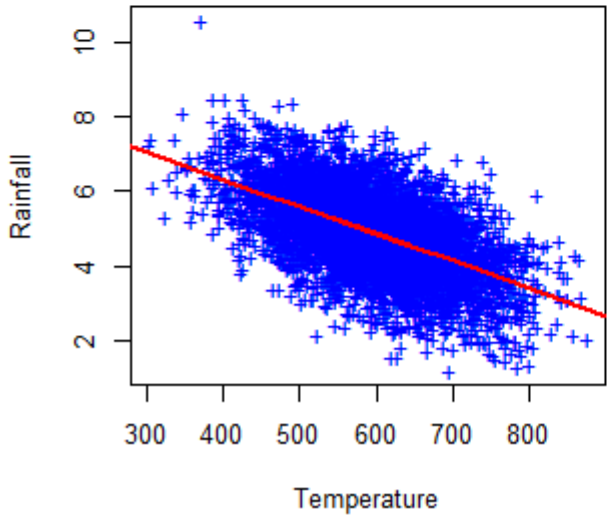
*Compute the partial correlation between yield and temperature given the rainfall.*

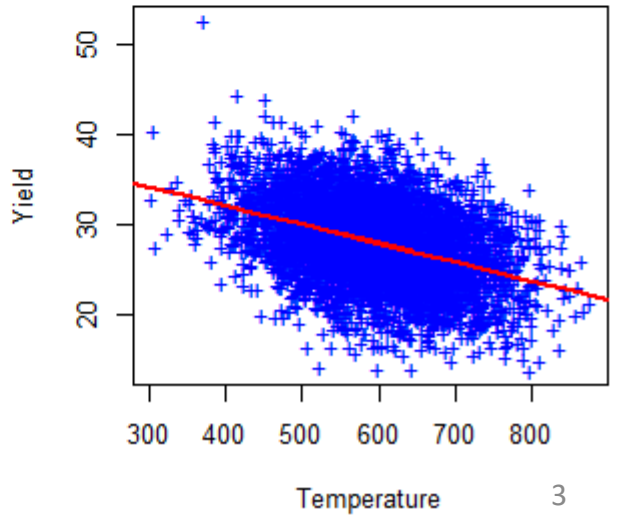


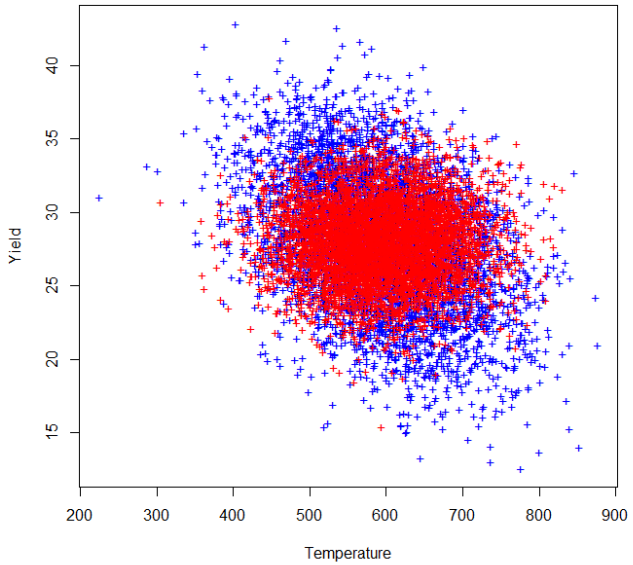

Another way to consider conditional distributions, is to fit a regression line, subtract it, and look at the residuals. I.e., remove the information that is associated with a specific value of the other variable.

**Pictures are simulated data, based on the estimates above.**



# Example - Exercise 2.4




Red dots are simulations from the conditional distributions given a rainfall of 4.91; **far less variable and with less correlation**; but that is because the rain doesn't vary….

**Partial covariance:**

$$\Sigma = \begin{pmatrix} 4.42^2 & 0.8 \cdot 4.42 \cdot 1.1 & -0.40 \cdot 4.42 \cdot 85 \\ * & 1.1^2 & -0.56 \cdot 1.1 \cdot 85 \\ * & * & 85^2 \end{pmatrix}$$

$$\Sigma_{yt,yt} = \begin{pmatrix} 4.42^2 & -0.40 \cdot 4.42 \cdot 85 \\ * & 85^2 \end{pmatrix}$$

$$\Sigma_{yt,r} = \begin{pmatrix} 0.8 \cdot 4.42 \cdot 1.1 \\ -0.56 \cdot 1.1 \cdot 85 \end{pmatrix}, \Sigma_{r,yt} = \Sigma_{yt,r}^T$$

$$\Sigma_{\mathrm{r,r}} = 1.1^2$$

$$\Sigma_{yt,yt|r} = \Sigma_{yt,yt} - \Sigma_{yt,r}\Sigma_{\mathrm{r,r}}^{-1}\Sigma_{r,yt}$$

$$= \begin{pmatrix} 4.42^2 & -0.40 \cdot 4.42 \cdot 85 \\ * & 85^2 \end{pmatrix} -$$

$$\begin{pmatrix} 0.8 \cdot 4.42 \cdot 1.1 \\ -0.56 \cdot 1.1 \cdot 85 \end{pmatrix} \left( \frac{1}{1.1^2} \right) (0.8 \cdot 4.42 \cdot 1.1 \quad -0.56 \cdot 1.1 \cdot 85)$$

$$= \begin{pmatrix} 7.03 & 18.03 \\ * & 4959.24 \end{pmatrix}$$

Partial correlation:

$$\frac{18.03}{\sqrt{7.03 \cdot 4959.24}} = 0.097$$

Compare with the initial correlation: -0.40!

# Estimation of parameters I

The i'th observation

$$\boldsymbol{X}_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{bmatrix}$$

The mean

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

The empirical variance- covariance or dispersion matrix

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T - \frac{n}{n-1}\bar{\boldsymbol{X}}\,\bar{\boldsymbol{X}}^T.$$

Observations collected in data matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T \\ \vdots \\ \boldsymbol{X}_n^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

# Estimation of parameters II

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{1}{n-1}\sum_{i=1}^{n} X_i X_i^T - \frac{n}{n-1}\bar{X}\bar{X}^T.$$

**Theorem 1.32**

Let the situation be as stated above. Then the $100(1-\alpha)\%$ confidence ellipsoid for the unknown mean $\boldsymbol{\mu}$ is

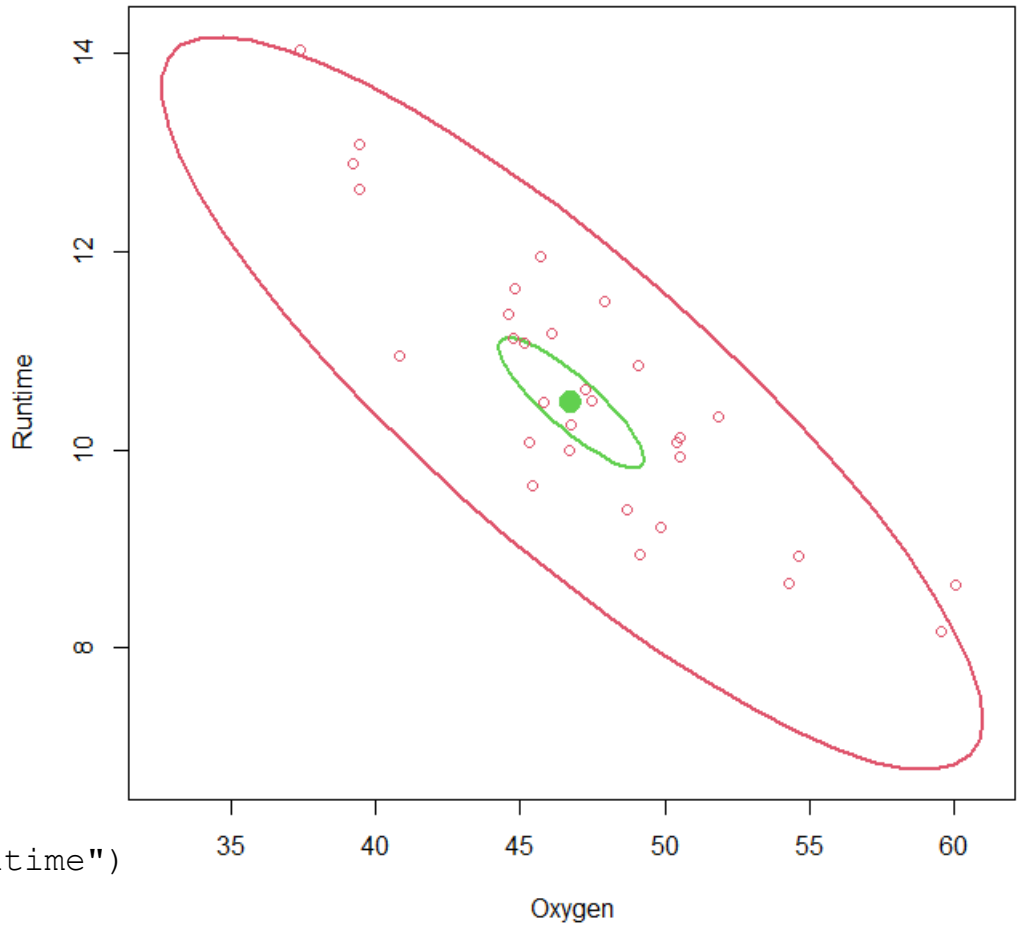
$$\{\boldsymbol{\mu} | (\mu - \bar{\mathbf{x}})^T \mathbf{s}^{-1} (\mu - \bar{\mathbf{x}}) \leq \frac{p(n-1)}{(n-p)n} F(p, n-p)_{1-\alpha}\}$$

and the $100(1-\alpha)\%$ prediction ellipsoid for a coming observation $\mathbf{x}$ is

$$\{\mathbf{x} | (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{s}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \leq \frac{p(n-1)(n+1)}{(n-p)n} F(p, n-p)_{1-\alpha}\}$$

# An example from fitness data

```
Fit<-read.table("Data/sundhed.txt")[,3:4]
names(Fit)<-c("Oxygen","Runtime")

n <- length(Fit$Oxygen)
P <-2
shape <- var(Fit)
center <- sapply(Fit,mean,1)

conf95 <-
  sqrt(p * (n-1) * qf(0.95, p, n-p)/(n*(n-p)))
pred95 <- sqrt(n+1)*rconf95

library(car)
pred.elip95 <-
car::ellipse(center, shape, pred95,draw = FALSE)
conf.elip95 <-
car::ellipse(center, shape, conf95,draw = FALSE)

plot(pred.elip95, type='l', col = 2,lwd=2,xlab="Oxygen",ylab="Runtime")
lines(conf.elip95, type='l', col = 3,lwd=2)
points(Fit$Oxygen, Fit$Runtime,col=2)
points(center[1],center[2],cex=2,pch=19,col=3)
```



# Partial correlation coefficient recap

**The partial correlations between some variables given other variables are simply correlations in the conditional distribution of the 'some' variables given the 'other' variables.** It follows that

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}$$

$$\rho_{ij|kl} = \frac{\rho_{ij|k} - \rho_{il|k} \cdot \rho_{jl|k}}{\sqrt{(1-\rho_{il|k}^2) \cdot (1-\rho_{jl|k}^2)}}$$

Let us look at some examples! ☺

# Exam 2016 - Question 2. What is the partial correlation between $X_2$ and $X_4$, given $X_3$?

Consider a random variable

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}} \qquad \boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

with dispersion (variance-covariance) matrix

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

**Question 2.1.**

The partial correlation $\rho_{24|3}$ is

A: 0 0%

B: $\rho$ 0%

C: $\rho^2$ 0%

D: $1-\rho^2$ 0%

E: $\rho/\sqrt{1-\rho^2}$ 0%

F: Don't know 0%

# Partial correlation coefficient
# Exam 2016 – problem 2

Consider a random variable

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}} \qquad \boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

with dispersion (variance-covariance) matrix

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

**Question 2.1.**

The partial correlation $\rho_{24|3}$ is

*Note that the variances are 1; covariances and correlations are the same…*

$$\rho_{24|3} = \frac{\rho_{24} - \rho_{23}\rho_{43}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{43}^2)}}$$

$$\frac{\rho^2 - \rho \cdot \rho}{\sqrt{(1-\rho^2)(1-\rho^2)}} = 0$$

**Option A**

# Example - Cement strength

| Correlation Matrix | | | | | |
|---|---|---|---|---|---|
| | C3S | C3A | BLAINE | Strgth3 | Strgth28 |
| C3S | 1.000 | . | . | . | . |
| C3A | -0.309 | 1.000 | . | . | . |
| BLAINE | 0.091 | 0.192 | 1.000 | . | . |
| Strgth3 | 0.158 | 0.120 | 0.745 | 1.000 | . |
| Strgth28 | 0.344 | -0.166 | 0.320 | 0.464 | 1.000 |

| Partial Correlation Matrix | | | | | |
|---|---|---|---|---|---|
| | C3S | C3A | BLAINE | Strgth3 | Strgth28 |
| C3S | 1.0000 | | | | |
| C3A | -.3340 | 1.0000 | | | |
| BLAINE | | | | | |
| Strgth3 | 0.1358 | -.0352 | | 1.0000 | |
| Strgth28 | 0.3337 | -.2446 | | 0.3570 | 1.0000 |

# Partial correlation coefficient III

**Theorem 1.37**

Let $R = R_{ij|m+1\ldots p}$ be the empirical partial correlation coefficient between $Z_i$ and $Z_j$ conditioned on (or: for given) $Z_{m+1,\ldots,Z_p}$. It is assumed to be computed from the unbiased estimates of the variance-covariance matrix and from $n$ observations. Then

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} \sim \mathrm{t}(n-2-(p-m)),$$

if $\rho_{ij|m+1,\ldots,p} = 0$.

**Example 1.39**

Let us investigate whether the value of $r_{24|3}$ is significantly different from 0. We find with $r_{24|3} = R$:

$$\begin{aligned}\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} &= \frac{-0.035}{\sqrt{1-0.035^2}}\cdot\sqrt{51-2-(5-4)} \\ &= -0.243 = \mathrm{t}(48)_{40\%}.\end{aligned}$$

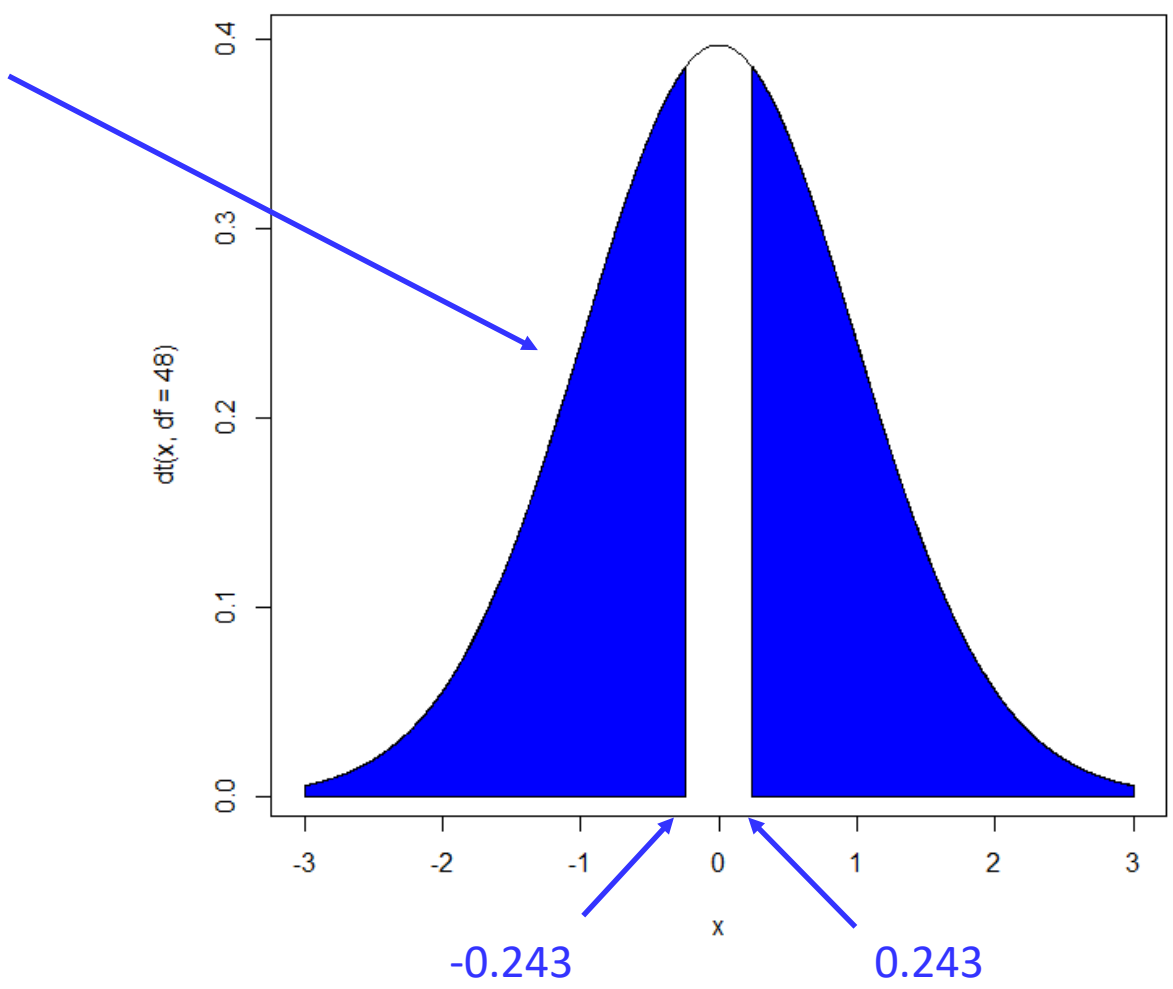
A hypothesis that $\rho_{24|3}$ is 0 will therefore be accepted using a test at level $\alpha$ for $\alpha < 80\%$. (Note: this is by nature a two-sided test.)

Note that from the example text one can deduce (nearly) that
$2 \cdot p(t_{48} < -0.243) = 0.81$;
the p-value is 0.81.

# Test-values, example 1.37

```
p<-2*pt(-.243, df=48)
P
[1] 0.8090408
```

There is therefore no evidence in the data
that supports that C3A and Strength3 are correlated,
once the BLAINE level is known.



# Multiple correlation coefficient - example

We now define the multiple correlation coefficient between $Y_i$ , $i = 1, \ldots, m$ and $\boldsymbol{X}$ as the maximal correlation between $Y_i$ and a linear combination of $\boldsymbol{X}$'s elements. It is denoted $\rho_{y_i|x}$.

**Theorem 1.42**

We consider the situation above. Let $\sigma_i$ be the $i$'th column in $\boldsymbol{\Sigma}_{xy}$, i.e. $\sigma_i^T$ is the $i$'th row in $\boldsymbol{\Sigma}_{yx}$. Further, let $\sigma_{ii}$ denote the i'th diagonal element, i.e. the variance of $Y_i$

Then

$$\rho_{y_i|x} = \frac{\sqrt{\sigma_i^T \boldsymbol{\Sigma}_{xx}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \boldsymbol{\Sigma}_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \boldsymbol{\Sigma}_i}{\sigma_{ii} \det \boldsymbol{\Sigma}_{xx}} = \frac{\mathrm{V}(Y_i|X)}{\mathrm{V}(Y_i)},$$

# Exam 2016 problem 2.2. What is the multiple correlation of $X_1$, given $X_2$ and $X_3$?

Consider a random variable

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

with dispersion (variance-covariance) matrix

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

A: 0 0%

B: $\rho$ 0%

C: $\rho^2$ 0%

D: $1 - \rho^2$ 0%

E: $\rho / \sqrt{1 - \rho^2}$ 0%

F: Don't Know 0%

# Exam 2016 – Problem 2.2

Consider a random variable

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{ii} & \boldsymbol{\sigma}_i^T \\ \boldsymbol{\sigma}_i & \boldsymbol{\Sigma}_{xx} \end{bmatrix},$$

$$1 - \rho^2_{Y_i|X} = \frac{\det\boldsymbol{\Sigma}_i}{\sigma_{ii}\det\boldsymbol{\Sigma}_{xx}} = \frac{\mathrm{V}(Y_i|\boldsymbol{X})}{\mathrm{V}(Y_i)},$$

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

with dispersion (variance-covariance) matrix

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

## Question 2.2.

The squared multiple correlation $\rho^2_{1|23}$ is

1 ☐ 0

2 ☐ $\rho$

3 ☐ $\rho^2$

4 ☐ $1-\rho^2$

5 ☐ $\rho/\sqrt{1+\rho^2}$

6 ☐ Don't know.

$$\rho^2_{1|23} = 1 - \frac{\det\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}}{1 \cdot \det\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}} =$$

$$1 - \frac{1\cdot 1\cdot 1 + \rho\cdot\rho\cdot\rho^2 + \rho^2\cdot\rho\cdot\rho - \rho^2\cdot 1\cdot\rho^2 - \rho\cdot\rho\cdot 1 - 1\cdot\rho\cdot\rho}{1-\rho^2} =$$

$$1 - \frac{1 + \rho^4 + \rho^4 - \rho^4 - \rho^2 - \rho^2}{1-\rho^2} =$$

$$1 - \frac{\rho^4 - 2\rho^2 + 1}{1-\rho^2} = 1 - \frac{(1-\rho^2)(1-\rho^2)}{1-\rho^2} = \rho^2$$

Alternative:

$$\rho^2_{1|23} = \Sigma_{1,23}\Sigma^{-1}_{23,23}\Sigma_{23,1} = \begin{pmatrix}\rho & \rho^2\end{pmatrix}\begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}^{-1}\begin{pmatrix}\rho \\ \rho^2\end{pmatrix}$$

$$= \frac{\rho^2}{1-\rho^2}\begin{pmatrix}1 & \rho\end{pmatrix}\begin{pmatrix}1 & -\rho \\ -\rho & 1\end{pmatrix}\begin{pmatrix}1 \\ \rho\end{pmatrix} = \frac{\rho^2}{1-\rho^2}\begin{pmatrix}1-\rho^2 & 0\end{pmatrix}\begin{pmatrix}1 \\ \rho\end{pmatrix} = \rho^2$$

**Option C**

# Example - Cement strength

| Correlation Matrix | | | | | |
|---|---|---|---|---|---|
| | C3S | C3A | BLAINE | Strgth3 | Strgth28 |
| C3S | 1.000 | . | . | . | . |
| C3A | -0.309 | 1.000 | . | . | . |
| BLAINE | 0.091 | 0.192 | 1.000 | . | . |
| Strgth3 | 0.158 | 0.120 | 0.745 | 1.000 | . |
| Strgth28 | 0.344 | -0.166 | 0.320 | 0.464 | 1.000 |

| Partial Correlation Matrix | | | | | |
|---|---|---|---|---|---|
| | C3S | C3A | BLAINE | Strgth3 | Strgth28 |
| C3S | 1.0000 | | | | |
| C3A | -.3340 | 1.0000 | | | |
| BLAINE | | | | | |
| Strgth3 | 0.1358 | -.0352 | | 1.0000 | |
| Strgth28 | 0.3337 | -.2446 | | 0.3570 | 1.0000 |

# Multiple correlation coefficient II

**Theorem 1.45**

Let $R = \hat{\rho}_{y_i|x}$ be the empirical multiple correlation coefficient between $Y_i$ and $\mathbf{X} = (Z_{m+1}, \ldots, Z_p)$ based upon $n$ observations. Then

$$\frac{R^2}{1-R^2} \cdot \frac{n-(p-m)-1}{p-m} \sim \mathrm{F}(p-m, n-(p-m)-1),$$

if $\rho_{y_i|x} = \rho_{y_i|z_{m+1},\ldots,p} = 0$.

**Example 1.44**

To get an impression of to which degree the content of $C_3A$ and $C_3S$ in example 1.39 can explain the variation in e.g. 3-day strength we can compute the multiple correlation coefficient between strength day 3 and ($C_3S$, and $C_3A$). We find

$$1-\hat{\rho}^2_{4|12} = \frac{\det\begin{bmatrix} 1 & 0.158 & 0.120 \\ 0.158 & 1 & -0.309 \\ 0.120 & -0.309 & 1 \end{bmatrix}}{1 \cdot \det\begin{bmatrix} 1 & -0.309 \\ -0.309 & 1 \end{bmatrix}}$$

where the indices of the variables correspond to those used in example 1.35. We find

$$\hat{\rho}^2_{4|12} = 1 - 0.9435 = 0.0565.$$

The data therefore indicate that only about 6% of the variation in the strength of the cement (from samples which have been collected the way these data have been collected) can be explained by variations in $C_3S$- and $C_3A$- content alone.

**Example 1.47**

Consider the situation in example 1.44. We now want to examine if it can be assumed that the multiple correlation between $X_4$ and $(X_1, X_2)$ is 0. (Note that $p = 3$ and $m = 1$.) We find the statistic

$$\frac{R^2}{1-R^2}\frac{51-(3-1)-1}{3-1} = \frac{0.0565}{0.9435} \cdot \frac{48}{2} = 1.44.$$

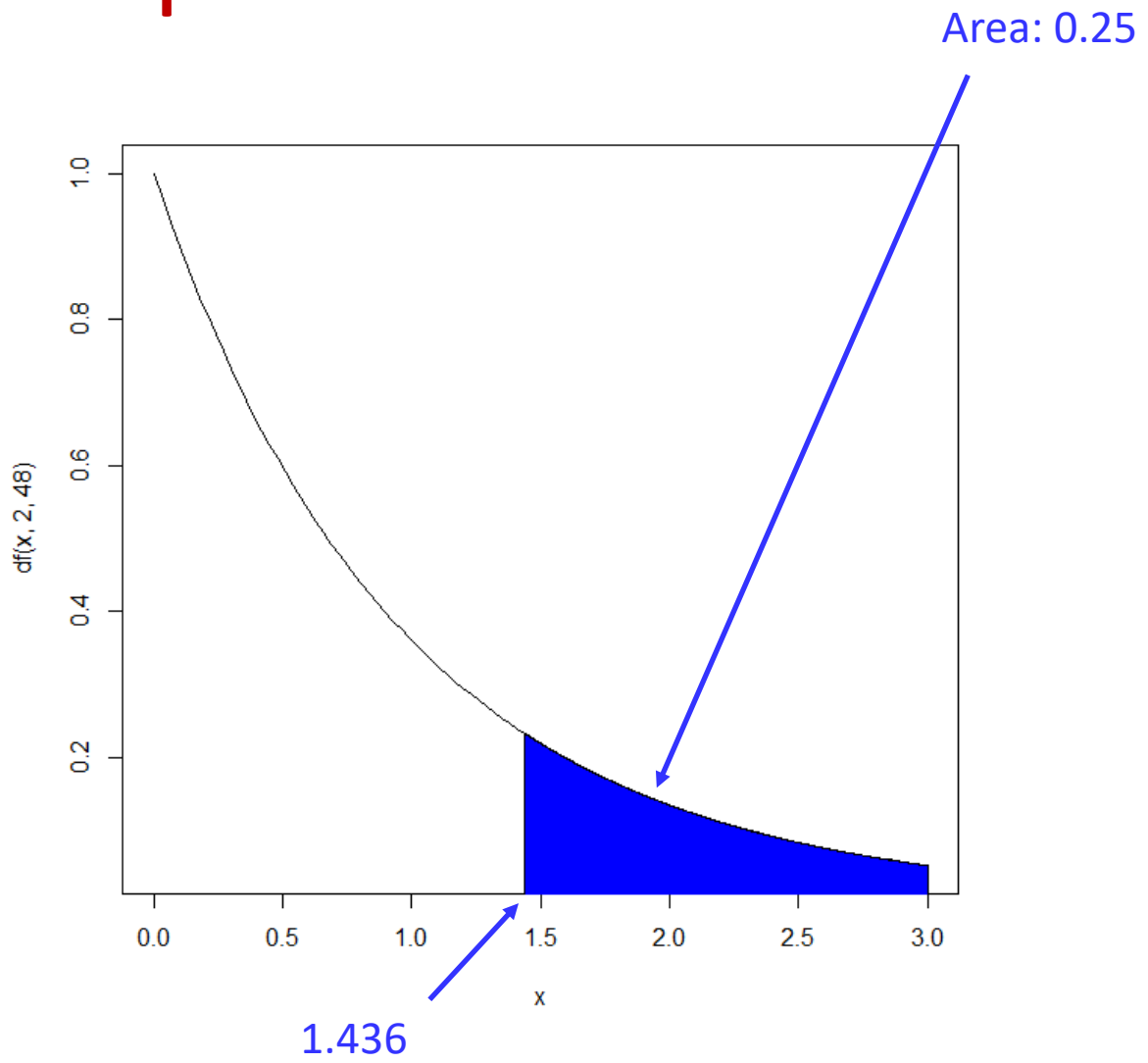
Since

$$\mathrm{F}(2, 48)_{0.90} = 2.42,$$

we will at least accept a hypothesis that $\rho_{4|12} = 0$ for any level $\alpha < 10\%$. With the available data it cannot be rejected that $\rho_{4|12} = 0$. This does not mean that it is not different from 0 (which it probably is), only that we cannot be sure using the available data because the true (but unknown) value of $\rho_{4|12}$ is probably rather small.

# Test-values, example 1.47

```
R2<-1-det(matrix(c(1,0.158,0.12,0.158,1,
          -0.309,0.12,-0.309,1),nrow=3))/
      det(matrix(c(1,-0.309,-0.309,1),nrow=2))
R2

# 1.436485

(R2/(1-R2))*48/2

# 1.436


1-pf((R2/(1-R2))*48/2,2,48)

# 0.2477989
```



# PCA: Principle Components Analysis

# Example: The Wine data

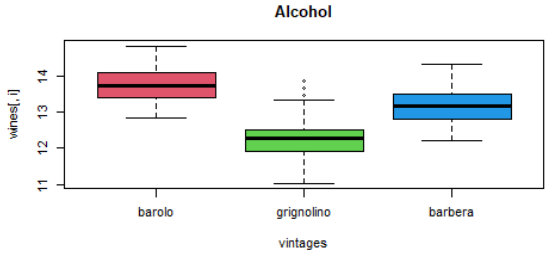
```
load("Data/Winedata.Rdata")
```

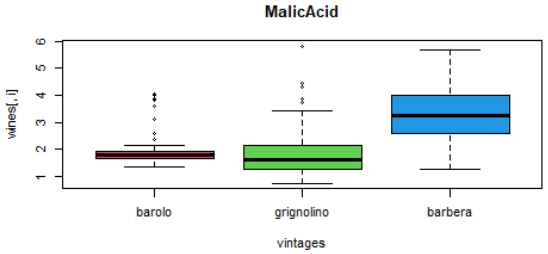
- A dataset consisting of n=178 Italian wines. Of these, 59 are Barolo wines, 71 are Grignolino wines, and 48 are Barbera wines. Two sub-elements: `wines` and `vintages` (Barolo, Grignolino or Barbera).

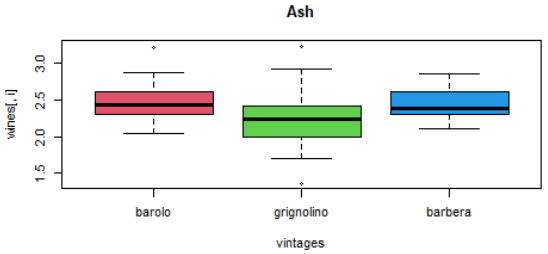
- 13 characteristics of the wines:

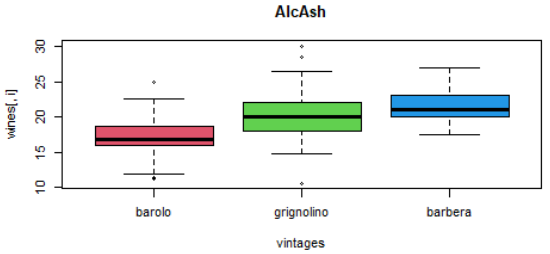
1) Alcohol
2) MalicAcid
3) Ash
4) Alkalinity of ash: AlcAsh
5) Magnesium: Mg
6) Total Phenols: Phenols
7) Flavanoids: Flavor
8) Nonflavanoid phenols: NonFlavPhenols
9) Proanthocyanins: Proa
10) Color intensity: Color
11) Color hue: Hue
12) OD280/OD315 protein measurement: OD
13) Proline (amino acid): Proline

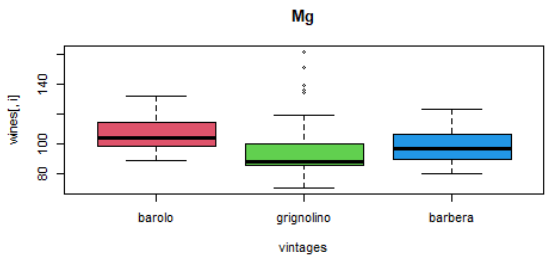
# Variation of Vintages

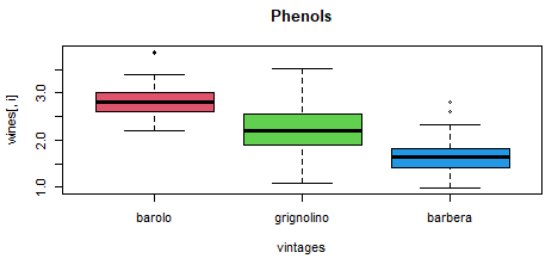


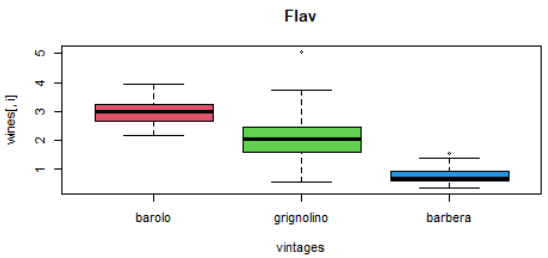



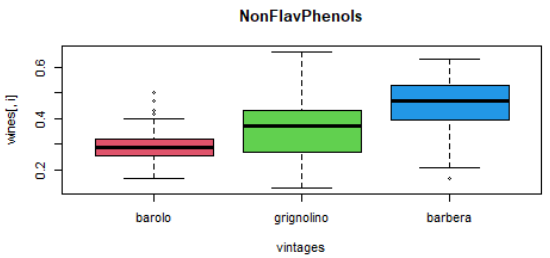



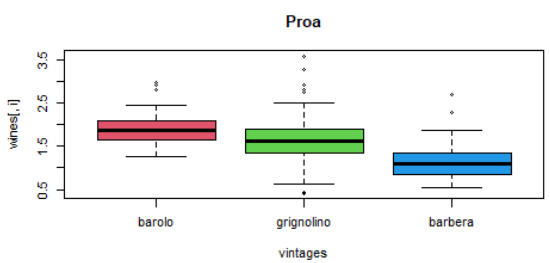



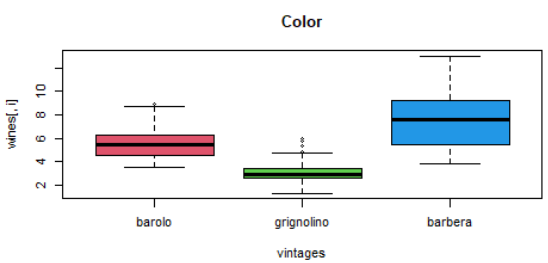



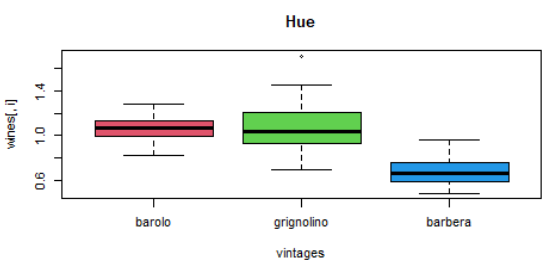



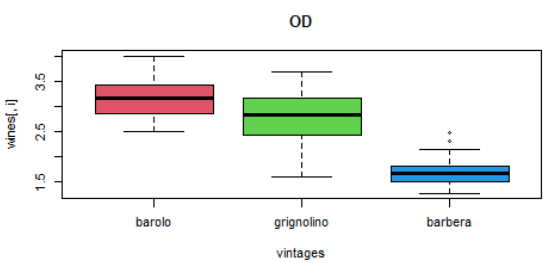



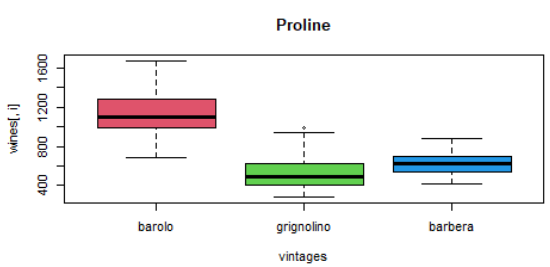



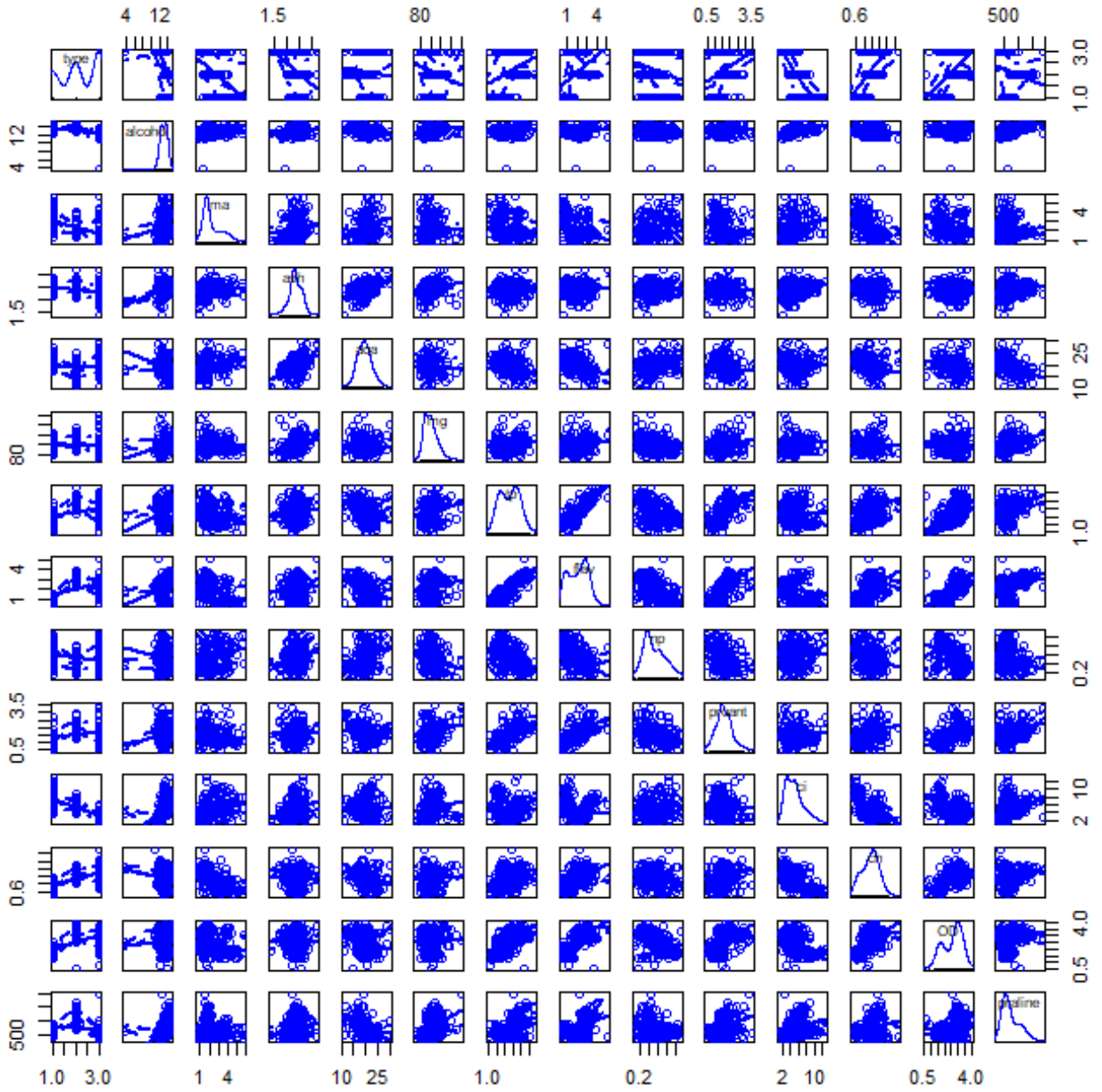



# The Wine Data



# Separation of Wines

- What causes wines to be different?
- With the 13 characteristics, we can distinguish wines through differences in the charateristics. But are all 13 characteristics necessary? Some may be redundant.
- If we can identify scales (linear combinations of the characteristics) where the characteristics vary the most, we can also (hope to) find a scale that differentiates optimally between the wines.

# Separation of Wines

- The variance covariance matrix of the wine characteristics:

```
round(var(wines),digits=2)
              Alcohol MalicAcid   Ash  AlcAsh      Mg Phenols  Flav NonFlavPhenols  Proa  Color   Hue   OD  Proline
Alcohol          0.66      0.09  0.05   -0.84    3.14    0.15  0.19          -0.02  0.06   1.03 -0.01 0.04   164.57
MalicAcid        0.09      1.25  0.05    1.08   -0.87   -0.23 -0.46           0.04 -0.14   0.64 -0.14 -0.29  -67.55
Ash              0.05      0.05  0.08    0.41    1.12    0.02  0.03           0.01  0.00   0.16  0.00 0.00   19.32
AlcAsh          -0.84      1.08  0.41   11.15   -3.97   -0.67 -1.17           0.15 -0.38   0.15 -0.21 -0.66 -463.36
Mg               3.14     -0.87  1.12   -3.97  203.99    1.92  2.79          -0.46  1.93   6.62  0.18 0.67 1769.16
Phenols          0.15     -0.23  0.02   -0.67    1.92    0.39  0.54          -0.04  0.22  -0.08  0.06 0.31   98.17
Flav             0.19     -0.46  0.03   -1.17    2.79    0.54  1.00          -0.07  0.37  -0.40  0.12 0.56  155.45
NonFlavPhenols  -0.02      0.04  0.01    0.15   -0.46   -0.04 -0.07           0.02 -0.03   0.04 -0.01 -0.04  -12.20
Proa             0.06     -0.14  0.00   -0.38    1.93    0.22  0.37          -0.03  0.33  -0.03  0.04 0.21   59.55
Color            1.03      0.64  0.16    0.15    6.62   -0.08 -0.40           0.04 -0.03   5.37 -0.28 -0.71  230.77
Hue             -0.01     -0.14  0.00   -0.21    0.18    0.06  0.12          -0.01  0.04  -0.28  0.05 0.09   17.00
OD               0.04     -0.29  0.00   -0.66    0.67    0.31  0.56          -0.04  0.21  -0.71  0.09 0.50   69.93
Proline        164.57    -67.55 19.32 -463.36 1769.16   98.17 155.45         -12.20 59.55 230.77 17.00 69.93 99166.72
```

# Separation of Wines

- To avoid scaling problems, we must scale the data to the same scale. Variances: Look in correlations instead-

- The `scale` function in R subtracts the mean and divide by the sd:

$$X^{scaled} = \frac{X - mean(X)}{sd(X)}$$

```
round(var(scale(wines)),digits=2)
```

| | Alcohol | MalicAcid | Ash | AlcAsh | Mg | Phenols | Flav | NonFlavPhenols | Proa | Color | Hue | OD | Proline |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alcohol | **1.00** | 0.09 | 0.21 | -0.31 | 0.27 | 0.29 | 0.24 | -0.16 | 0.14 | 0.55 | -0.07 | 0.07 | 0.64 |
| MalicAcid | 0.09 | **1.00** | 0.16 | 0.29 | -0.05 | -0.34 | -0.41 | 0.29 | -0.22 | 0.25 | -0.56 | -0.37 | -0.19 |
| Ash | 0.21 | 0.16 | **1.00** | 0.44 | 0.29 | 0.13 | 0.12 | 0.19 | 0.01 | 0.26 | -0.07 | 0.00 | 0.22 |
| AlcAsh | -0.31 | 0.29 | 0.44 | **1.00** | -0.08 | -0.32 | -0.35 | 0.36 | -0.20 | 0.02 | -0.27 | -0.28 | -0.44 |
| Mg | 0.27 | -0.05 | 0.29 | -0.08 | **1.00** | 0.21 | 0.20 | -0.26 | 0.24 | 0.20 | 0.06 | 0.07 | 0.39 |
| Phenols | 0.29 | -0.34 | 0.13 | -0.32 | 0.21 | **1.00** | 0.86 | -0.45 | 0.61 | -0.06 | 0.43 | 0.70 | 0.50 |
| Flav | 0.24 | -0.41 | 0.12 | -0.35 | 0.20 | 0.86 | **1.00** | -0.54 | 0.65 | -0.17 | 0.54 | 0.79 | 0.49 |
| NonFlavPhenols | -0.16 | 0.29 | 0.19 | 0.36 | -0.26 | -0.45 | -0.54 | **1.00** | -0.37 | 0.14 | -0.26 | -0.50 | -0.31 |
| Proa | 0.14 | -0.22 | 0.01 | -0.20 | 0.24 | 0.61 | 0.65 | -0.37 | **1.00** | -0.03 | 0.30 | 0.52 | 0.33 |
| Color | 0.55 | 0.25 | 0.26 | 0.02 | 0.20 | -0.06 | -0.17 | 0.14 | -0.03 | **1.00** | -0.52 | -0.43 | 0.32 |
| Hue | -0.07 | -0.56 | -0.07 | -0.27 | 0.06 | 0.43 | 0.54 | -0.26 | 0.30 | -0.52 | **1.00** | 0.57 | 0.24 |
| OD | 0.07 | -0.37 | 0.00 | -0.28 | 0.07 | 0.70 | 0.79 | -0.50 | 0.52 | -0.43 | 0.57 | **1.00** | 0.31 |
| Proline | 0.64 | -0.19 | 0.22 | -0.44 | 0.39 | 0.50 | 0.49 | -0.31 | 0.33 | 0.32 | 0.24 | 0.31 | **1.00** |

# Separation of Wines

- Let us consider the correlation matrix:

```
Sigma<-var(scale(wines))
```

The sum of the standardized variances:

```
sum(diag(Sigma))
[1] 13
```

Of course – there are 13 variables.

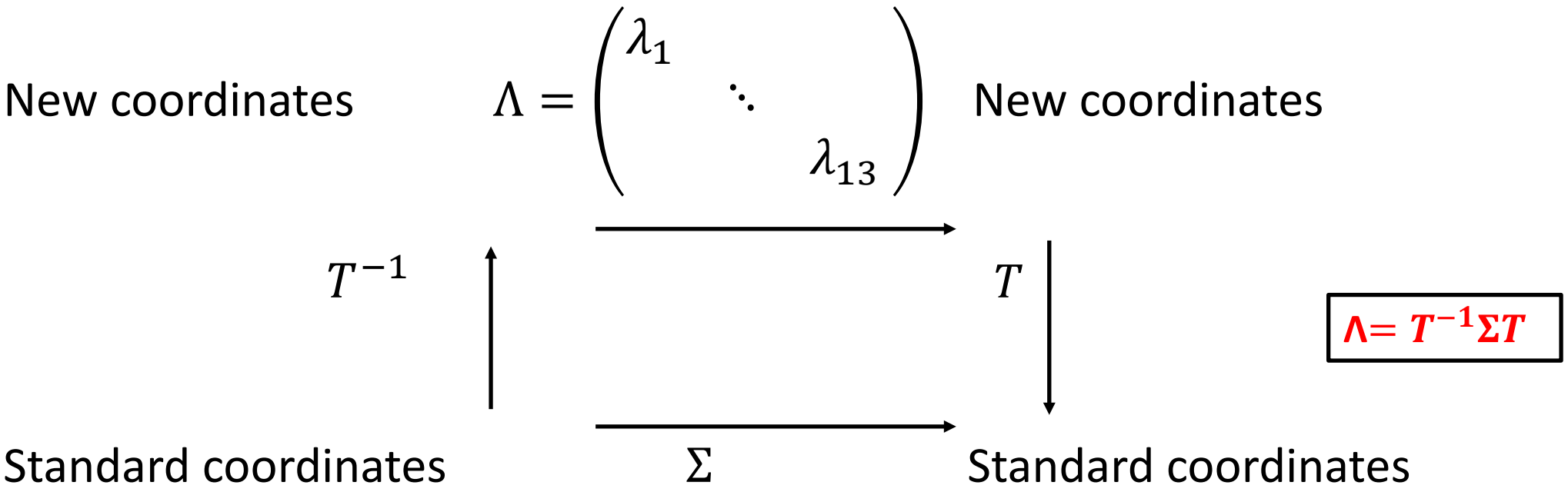
**Original question:** In which direction (scale) do wines data vary the most?

Let us start by representing the data in a set of coordinates where no correlation is present, to get an overview not disturbed by correlations.

# Separation of Wines

- No correlations means that $\Sigma$ is represented by a diagonal matrix in these directions;
- in other words that the new coordinates $T$ consists of eigenvectors for $\Sigma$; solutions to the equation

$$\Sigma v = \lambda v$$




$$\boldsymbol{\Lambda = T^{-1}\Sigma T}$$

# Separation of Wines

- Eigenvectors in R:

```
T<-eigen(Sigma)$vectors
```

The inverse of T is equal to the matrix transpose t(T):

```
Lambda<-t(T)%*%Sigma%*%T
round(Lambda, digits=2)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
 [1,] 4.71  0.0  0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00  0.00   0.0
 [2,] 0.00  2.5  0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00  0.00   0.0
 [3,] 0.00  0.0  1.45 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00  0.00   0.0
 [4,] 0.00  0.0  0.00 0.92 0.00 0.00 0.00 0.00 0.00  0.00  0.00  0.00   0.0
 [5,] 0.00  0.0  0.00 0.00 0.85 0.00 0.00 0.00 0.00  0.00  0.00  0.00   0.0
 [6,] 0.00  0.0  0.00 0.00 0.00 0.64 0.00 0.00 0.00  0.00  0.00  0.00   0.0
 [7,] 0.00  0.0  0.00 0.00 0.00 0.00 0.55 0.00 0.00  0.00  0.00  0.00   0.0
 [8,] 0.00  0.0  0.00 0.00 0.00 0.00 0.00 0.35 0.00  0.00  0.00  0.00   0.0
 [9,] 0.00  0.0  0.00 0.00 0.00 0.00 0.00 0.00 0.29  0.00  0.00  0.00   0.0
[10,] 0.00  0.0  0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.25  0.00  0.00   0.0
[11,] 0.00  0.0  0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.23  0.00   0.0
[12,] 0.00  0.0  0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00  0.17   0.0
[13,] 0.00  0.0  0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.00  0.00  0.00   0.1
```

# Separation of Wines

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] | [,13] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [1,] | **4.71** | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [2,] | 0.00 | **2.5** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [3,] | 0.00 | 0.0 | **1.45** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [4,] | 0.00 | 0.0 | 0.00 | **0.92** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [5,] | 0.00 | 0.0 | 0.00 | 0.00 | **0.85** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [6,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | **0.64** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [7,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | **0.55** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [8,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.35** | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [9,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.29** | 0.00 | 0.00 | 0.00 | 0.0 |
| [10,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.25** | 0.00 | 0.00 | 0.0 |
| [11,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.23** | 0.00 | 0.0 |
| [12,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.17** | 0.0 |
| [13,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.1** |

- It turns out that this matrix **provides us with the answer to our problem**:
- Any (normed) linear combination of the 13 standardized characteristics will also be a (normed) linear combination of the 13 uncorrelated eigenvalues;
- A little consideration shows that because of this, the variance of any normed linear combination can never exceed the maximum variance of the eigenvectors – **4.71**.
- The solution is thus **the first eigenvector**: `T[,1]`.

# Separation of Wines

- The combination of the scaled data that varies the most:

```
round(T[,1],digits=2)
 [1] -0.14  0.25  0.00  0.24 -0.14 -0.39 -0.42  0.30 -0.31  0.09 -0.30
[12] -0.38 -0.29
```

- Thus the most varying combination of the scaled data is

$$-0.14 * \widetilde{Alcohol} + 0.25 * \widetilde{MalicAcid} + 0 * \widetilde{Ash} + 0.24 * \widetilde{AlcAsh} - 0.14 * \widetilde{Mg} - 0.39 * \widetilde{\text{Phenols}} - 0.42 * \widetilde{PhenolsFlav} + 0.30 * \widetilde{NonFlavPhenols} - 0.31 * \widetilde{Proa} + 0.09 * \widetilde{Color} - 0.30 * \widetilde{Hue} - 0.38 * \widetilde{OD} - 0.29 * \widetilde{Proline}$$

Where the $\sim$ versions are the scaled variables, with the mean subtracted and divided by the standard deviation.

- This is the scale that we want to look at, when we seek to maximize the separation of wines.

# Separation of Wines

- The combination of the original data that varies the most:

```
round(T[,1]*sqrt(diag(var(wines))),digits=2)
 [1] -0.12  0.27  0.00  0.80 -2.03 -0.25 -0.42  0.04 -0.18  0.21 -
0.07
[12] -0.27 -90.30
```

- Thus the most varying combination of the original data is

$$-0.12 * \boldsymbol{Alcohol} + 0.27 * \boldsymbol{MalicAcid} + 0 * \boldsymbol{Ash} + 0.8 * \boldsymbol{AlcAsh} - 2.03 * \boldsymbol{Mg} - 0.25 * \text{Phenols} - 0.42 * \boldsymbol{PhenolsFlav} + 0.04 * \boldsymbol{NonFlavPhenols} - 0.18 * \boldsymbol{Proa} + 0.21 * \boldsymbol{Color} - 0.07 * \boldsymbol{Hue} - 0.27 * \boldsymbol{OD} - 90.30 * \boldsymbol{Proline}$$

# Separation of Wines

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] | [,13] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [1,] | **4.71** | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [2,] | 0.00 | **2.5** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [3,] | 0.00 | 0.0 | **1.45** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [4,] | 0.00 | 0.0 | 0.00 | **0.92** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [5,] | 0.00 | 0.0 | 0.00 | 0.00 | **0.85** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [6,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | **0.64** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [7,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | **0.55** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [8,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.35** | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 |
| [9,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.29** | 0.00 | 0.00 | 0.00 | 0.0 |
| [10,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.25** | 0.00 | 0.00 | 0.0 |
| [11,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.23** | 0.00 | 0.0 |
| [12,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.17** | 0.0 |
| [13,] | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.1** |

- **Further conclusions:**

- The scale, uncorrelated with the first eigenvector, that varies the most, is exactly the 2nd eigenvector `T[,2]`, with variance **2.5**. And so on…

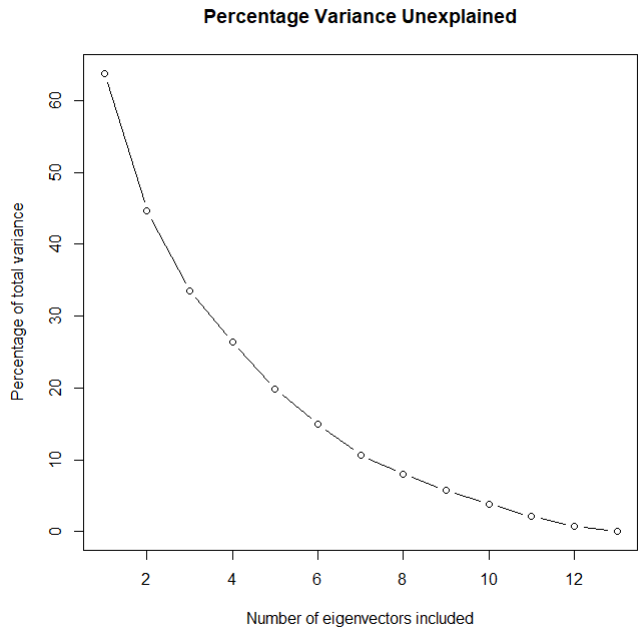
- The total variation after the coordinate shift is unchanged:

```
sum(diag(Lambda))
[1] 13
```

- Also note that the contribution from the 13th eigenvector is only 0.1/13, **0.7%**

# Separation of Wines

```
plot(100*(13-cumsum(diag(Lambda)))/13,type="b",
     main="Percentage Variance Unexplained",
     xlab='Number of eigenvectors included',
     ylab='Percentage of total variance')
```




| Eigenvectors | % variance explained |
|---|---|
| 1 | 36 |
| 2 | 55 |
| 3 | 67 |
| 4 | 74 |
| 5 | 80 |
| 6 | 85 |
| 7 | 89 |
| 8 | 92 |
| 9 | 94 |
| 10 | 96 |
| 11 | 98 |
| 12 | 99 |
| 13 | 100 |

# Principal Component Analysis

- PCA is a method to handle many variables, which are mutually correlated.
- PCA seeks to identify underlying dimensions in a data material, and to estimate the relationship between these.
- PCA may be used as a data reducing method, often in relation to multiple regression.
- PCA reduces the number of explanatory variables to a lesser number of ”principal komponents”, with (we hope) nearly as much of the variation as the inital variables.

**What PCA does is essentially the contents of the preceding slides!**

- The eigenvectors on the previous slides are **exactly** what is known as the **Principal Components**.

# Identification of Lower- Dimensional Spaces

*Figure 3.10 Data swarm (quasi-planar)*

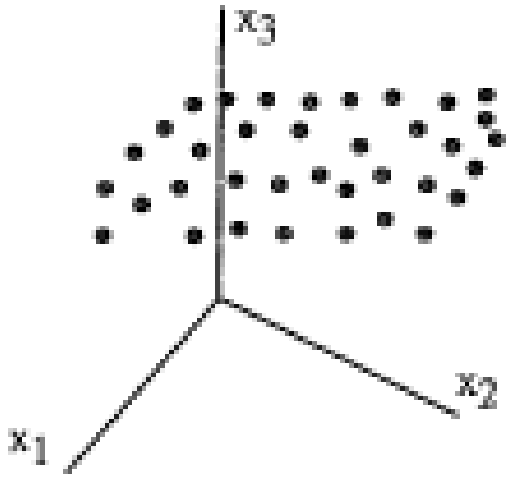
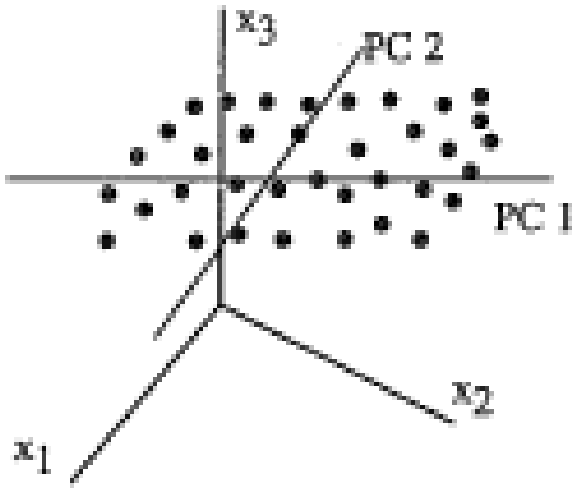



*Figure 3.11 Data swarm with 2 PCs*



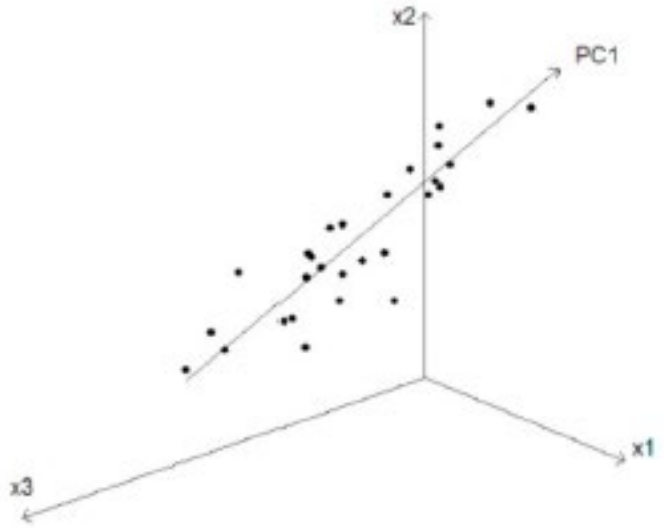
# The First Principal Component – PC1

We look for the direction that **explains as much as possible of the variation in the data**. We assume here 3 variables and n=28:

$$\boldsymbol{p_1} = t_{11}\boldsymbol{X_1}$$

where $\sum_{j=1}^{3} {t_{1j}}^2 = 1$.

- p are the ”scores”;
- t are the ”loadings”.



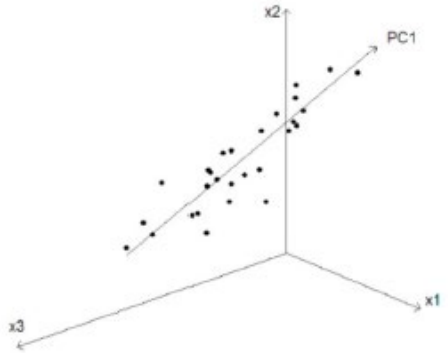
# The First Principal Component – PC1

We look for the direction that **explains as much as possible of the variation in the data**. We assume here 3 variables and n=28:

$$\boldsymbol{p_1} = t_{11}\boldsymbol{X_1} + t_{21}\boldsymbol{X_2} + t_{31}\boldsymbol{X_3},$$

where $\sum_{j=1}^{3} {t_{1j}}^2 = 1$.

- p are the ”scores”;
- t are the ”loadings”.




- Important distinction from the book:

- In the book, **the scores ARE the principal components**, not the direction!
- The turning (coordinate change) matrix $T$ is labeled $P$, and the principal components are labeled $Y$:

$$Y = P^T X$$

Here:

$$Scores = T^T X$$

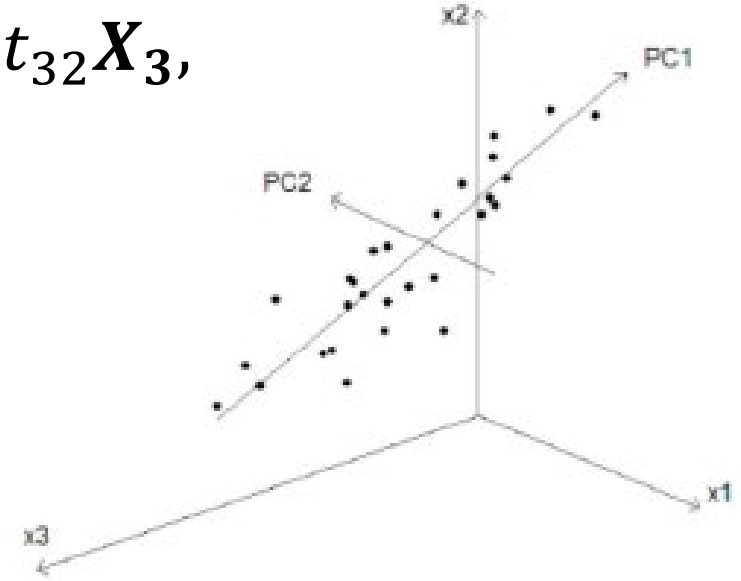
# The Second Principal Component – PC2

We consider the  plane perpendicular to **PC1**, and find the linear combination the explans the 2nd most variation:

$$\boldsymbol{p_2} = t_{12}\boldsymbol{X_1} + t_{22}\boldsymbol{X_2} + t_{32}\boldsymbol{X_3},$$

where $\sum_{j=1}^{3} {t_{2j}}^2 = 1$

PC1 og PC2 and are orthogonal.



# A few theorems:

**Theorem 6.3**

The principal components are uncorrelated and the variance of the $i'$th component is $\lambda_i$ i.e. the $i'$th largest eigenvalue.

**Theorem 6.5**

The total variance i.e. the sum of variance of the original variables is equal to the sum of the variance of the principal components i.e.

$$\sum_i \mathrm{V}(X_i) = \sum_i \mathrm{V}(Y_i)$$

# Important note

- Theorems on the previous slide only works **if the variance matrix $\Sigma$ is essentially known**, ie. if the Principal Components are known.

- Very often, the variance matrix is not known and will be estimated from the data (like the wines data);

**In that case, the 'theorems' one the previous slide only holds approximately, and only for large samples!**

- Reason: When estimated from data, the principal components in terms of directions are themselves stochastic variables; converting observations to scores is done by **multiplying with a RANDOM matrix**, not a fixed matrix.

- If the uncertainty of the the Principal Components may be neglected (large samples), one may assume the theorems to hold even for estimated variances.

# An observation

- If $X$ is multivariate normal with unknown variance, are the scores then multivariate normal?

**NO!**

Reason as on the previous slide.

- However, such an assumption is often not needed: Often PCA is used as an explorative data technique that bypasses statistical modeling (for a start).

# PCA of the Wines Data

```
wines.PC<- PCA(scale(wines))
names(wines.PC)
[1] "scores"        "loadings"      "var"           "totalvar"
[5] "centered.data"
```

```
summary(wines.PC)

PCA model of a mean-centered matrix of 178 by 13
Number of PCs to cover 90 percent of the variance: 8

            Var Cumul. var.
PC 1  36.198848    36.19885
PC 2  19.207490    55.40634
PC 3  11.123631    66.52997
PC 4   7.069030    73.59900
PC 5   6.563294    80.16229
PC 10  1.930019    96.16972
```

- Let's take a look at the components

Remember this:

| Eigenvectors | % variance explained |
|---|---|
| 1 | 36 |
| 2 | 55 |
| 3 | 67 |
| 4 | 74 |
| 5 | 80 |
| 6 | 85 |
| 7 | 89 |
| 8 | 92 |
| 9 | 94 |
| 10 | 96 |
| 11 | 98 |
| 12 | 99 |
| 13 | 100 |

# PCA of the Wines Data – the Loadings

- The **loadings** are the *coordinates of the principal components*:

```
> head(wines.PC$loadings,n=3)
 PC 1        PC 2         PC 3        PC 4         PC 5        PC 6
Alcohol   -0.144329395 -0.4836515 -0.20738262 -0.0178563  0.26566365 -0.2135386
MalicAcid  0.245187580 -0.2249309  0.08901289  0.5368903 -0.03521363 -0.5368138
Ash        0.002051061 -0.3160688  0.62622390 -0.2141756  0.14302547 -0.1544747
                 PC 7        PC 8         PC 9        PC 10        PC 11
Alcohol   -0.05639636 -0.39613926 -0.50861912 -0.21160473  0.22591696
MalicAcid  0.42052391 -0.06582674  0.07528304  0.30907994 -0.07648554
Ash       -0.14917061  0.17026002  0.30769445  0.02712539  0.49869142
                PC 12        PC 13
Alcohol    0.26628645 -0.01496997
MalicAcid -0.12169604 -0.02596375
Ash        0.04962237  0.14121803
> head(T,n=3)
             [,1]       [,2]        [,3]       [,4]        [,5]       [,6]        [,7]
 [,1]       [,2]        [,3]        [,4]        [,5]       [,6]
[1,] -0.144329395 -0.4836515 -0.20738262 -0.0178563  0.26566365 0.2135386
[2,]  0.245187580 -0.2249309  0.08901289  0.5368903 -0.03521363 0.5368138
[3,]  0.002051061 -0.3160688  0.62622390 -0.2141756  0.14302547 0.1544747
             [,7]        [,8]        [,9]       [,10]       [,11]       [,12]
[1,]  0.05639636  0.39613926  0.50861912  0.21160473 -0.22591696 -0.26628645
[2,] -0.42052391  0.06582674 -0.07528304 -0.30907994  0.07648554  0.12169604
[3,]  0.14917061 -0.17026002 -0.30769445 -0.02712539 -0.49869142 -0.04962237
            [,13]
[1,]  0.01496997
[2,]  0.02596375
[3,] -0.14121803
>
```

- The principal components are only identified up to a sign change.

# PCA of the Wines Data – the Scores

- The **scores** are the *new coordinates* of the (scaled) wines data *relative to the principal components*:

```
> head(wines.PC$scores,n=3)
          PC 1       PC 2       PC 3       PC 4       PC 5       PC 6
[1,] -3.307421 -1.4394023 -0.1652728 -0.2150246 -0.6910933 -0.2232504
[2,] -2.203250  0.3324551 -2.0207571 -0.2905387  0.2569299 -0.9245123
[3,] -2.509661 -1.0282507  0.9800541  0.7228632  0.2503270  0.5477310
           PC 7        PC 8       PC 9      PC 10     PC 11         PC 12
[1,] 0.59474883  0.06495586 -0.6396384 -1.0180840 0.4502932 -0.5392891439
[2,] 0.05362434  1.02153432  0.3079780 -0.1592521 0.1422560 -0.3871456499
[3,] 0.42301218 -0.34324787  1.1745213 -0.1130420 0.2858665 -0.0005819316
            PC 13
[1,]  0.066052305
[2,] -0.003626273
[3,] -0.021655423

> head(scale(wines)%*%T,n=3)
          [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
[1,] -3.307421 -1.4394023 -0.1652728 -0.2150246 -0.6910933  0.2232504
[2,] -2.203250  0.3324551 -2.0207571 -0.2905387  0.2569299  0.9245123
[3,] -2.509661 -1.0282507  0.9800541  0.7228632  0.2503270 -0.5477310
           [,7]        [,8]       [,9]      [,10]      [,11]        [,12]
[1,] -0.59474883 -0.06495586  0.6396384 1.0180840 -0.4502932 0.5392891439
[2,] -0.05362434 -1.02153432 -0.3079780 0.1592521 -0.1422560 0.3871456499
[3,] -0.42301218  0.34324787 -1.1745213 0.1130420 -0.2858665 0.0005819316
            [,13]
[1,] -0.066052305
[2,]  0.003626273
[3,]  0.021655423
>
```

- Note the same sign changes

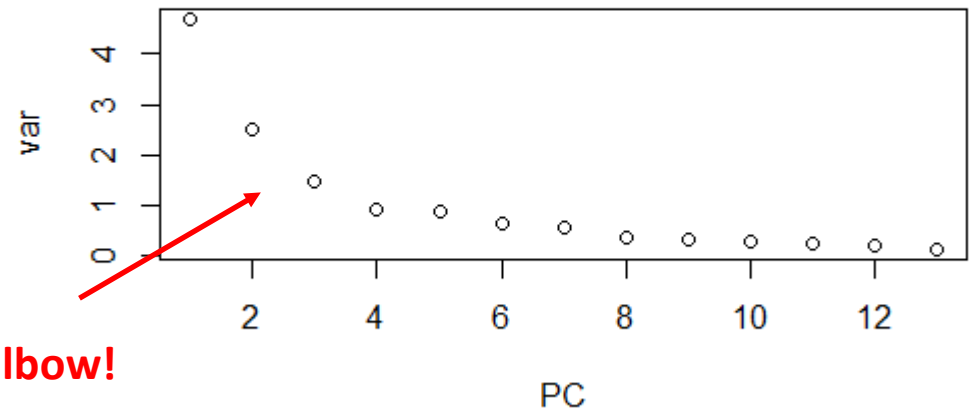
# PCA of the Wines data
# var, totalvar and centered.data

```
wines.PC$var
     PC 1      PC 2      PC 3      PC 4      PC 5      PC 6      PC 7      PC 8
4.7058503 2.4969737 1.4460720 0.9189739 0.8532282 0.6416570 0.5510283 0.3484974
     PC 9     PC 10     PC 11     PC 12     PC 13
0.2888799 0.2509025 0.2257886 0.1687702 0.1033779

wines.PC$totalvar
[1] 13
wines.PC$centered.data
[1] TRUE
```

We recognize the eigenvalues of the matrix $\Lambda$, and the sum of these. Lastly, an indicator that we have ‘done the right thing’ (in this case).

# Selecting the Number of Principal Components – the Scree Plot

```
plot(1:13,wines.PC$var,xlab="PC",ylab="var")
```




- Rules of thumb:

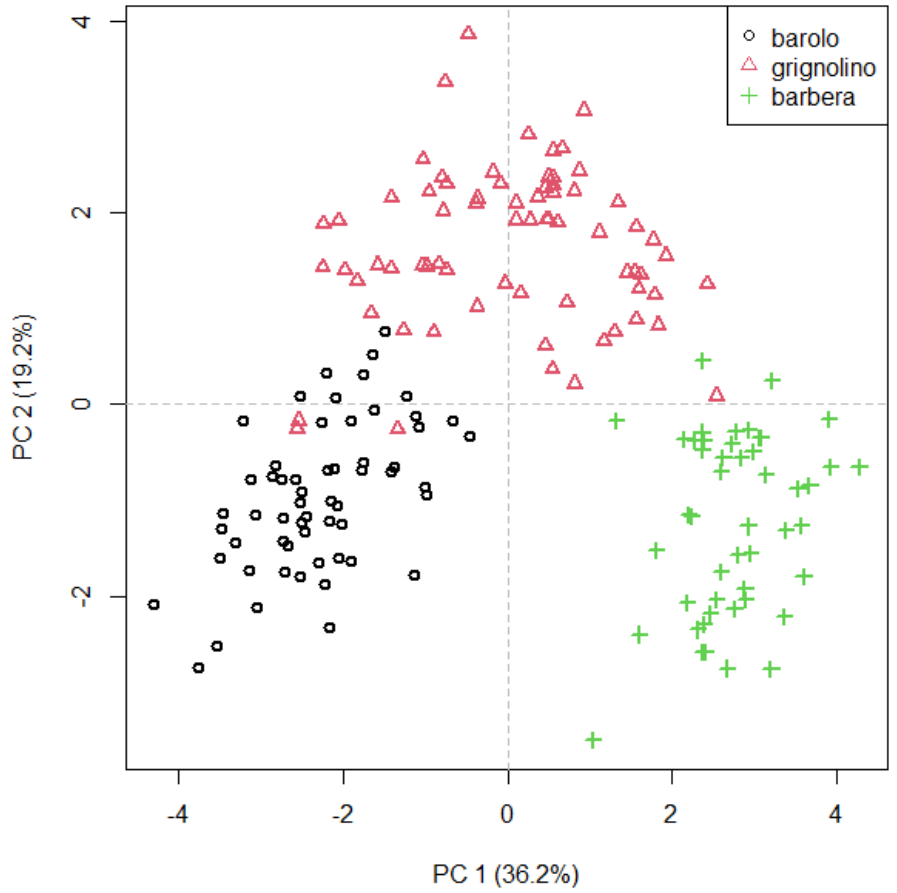
i. You select a number of principal components where the 'elbow' of the graph is.

ii. You usually (but not always) only select principal components with a variance greater than 1 – if the value is lower, the PC explain less than one average ordinary observation.

iii. You fix the amount of total variation that you need explained – t.ex. 80%.

# The Score plot

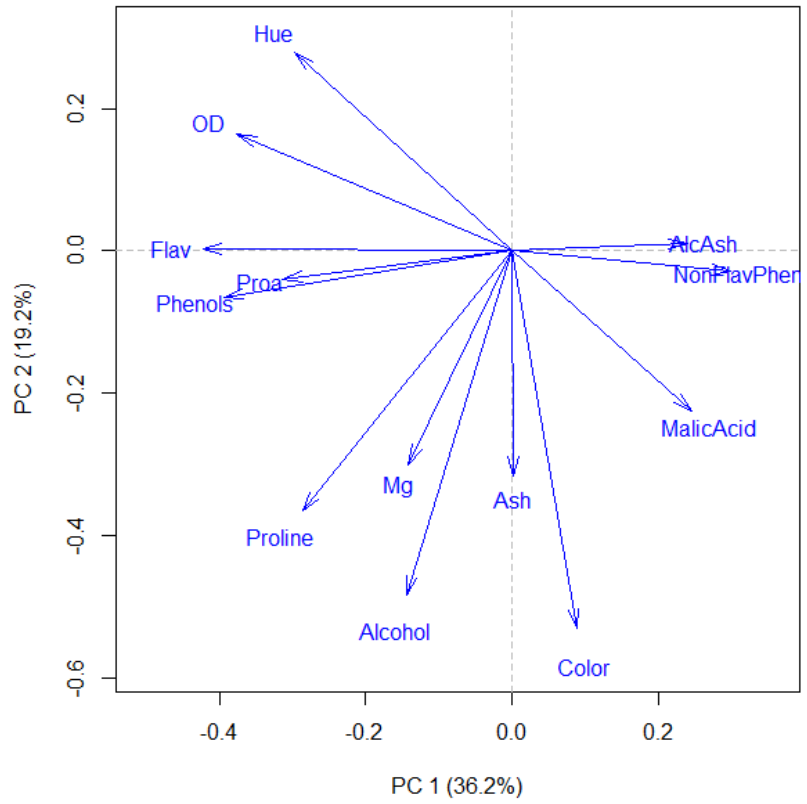
```
scoreplot(wines.PC, col = vintages, pch= as.numeric(vintages), lwd=2)
legend("topright",levels(vintages), col=1:3,pch=1:3)
```




**Apparently, we can more or less separate the vintages from just TWO dimensions!**

# The Loading plot

```
loadingplot(wines.PC, show.names= TRUE)
```




The **Loading plot** shows the points

$$\begin{bmatrix} Corr(X_i, PC1) \\ Corr(X_i, PC2) \end{bmatrix}$$

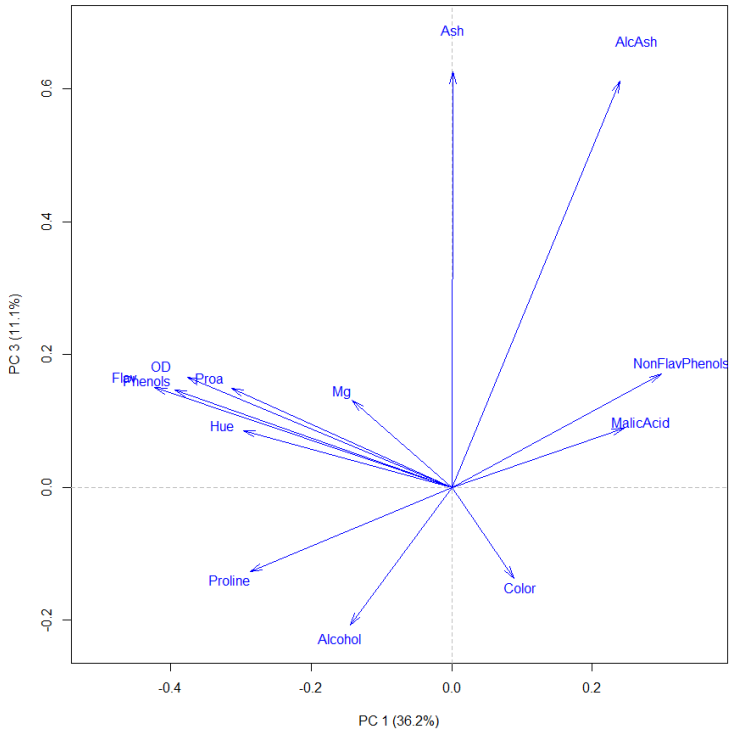
For the 13 variables.

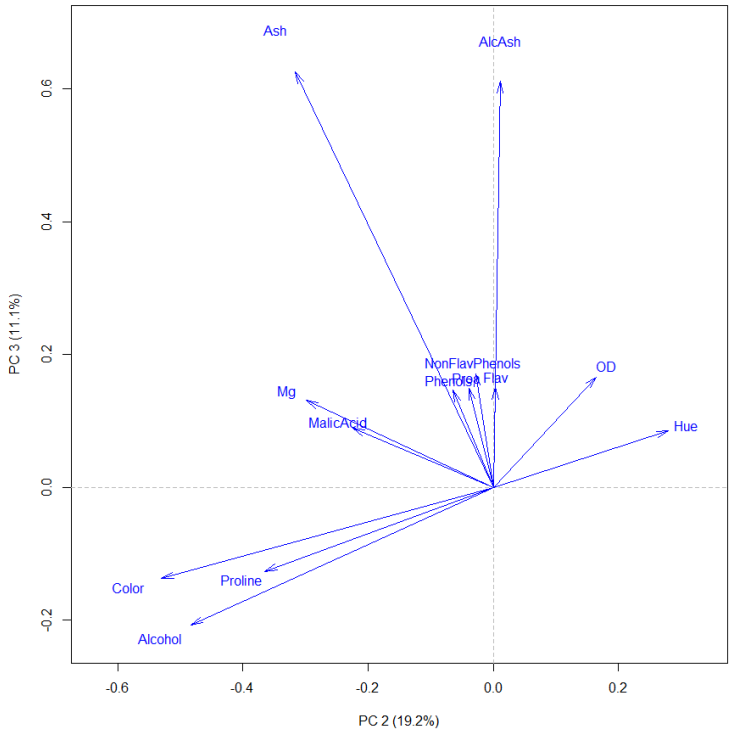
Some considerations (not trivial):

**Barolo:** Sweet taste, alcoholic, low nutritional value, saturated color.
**Grignolino:** Sweet taste, low nutritional value, bright color.
**Barbera:** Bitter taste, high nutritional value, deep color.

# Higher Order Loading Plots

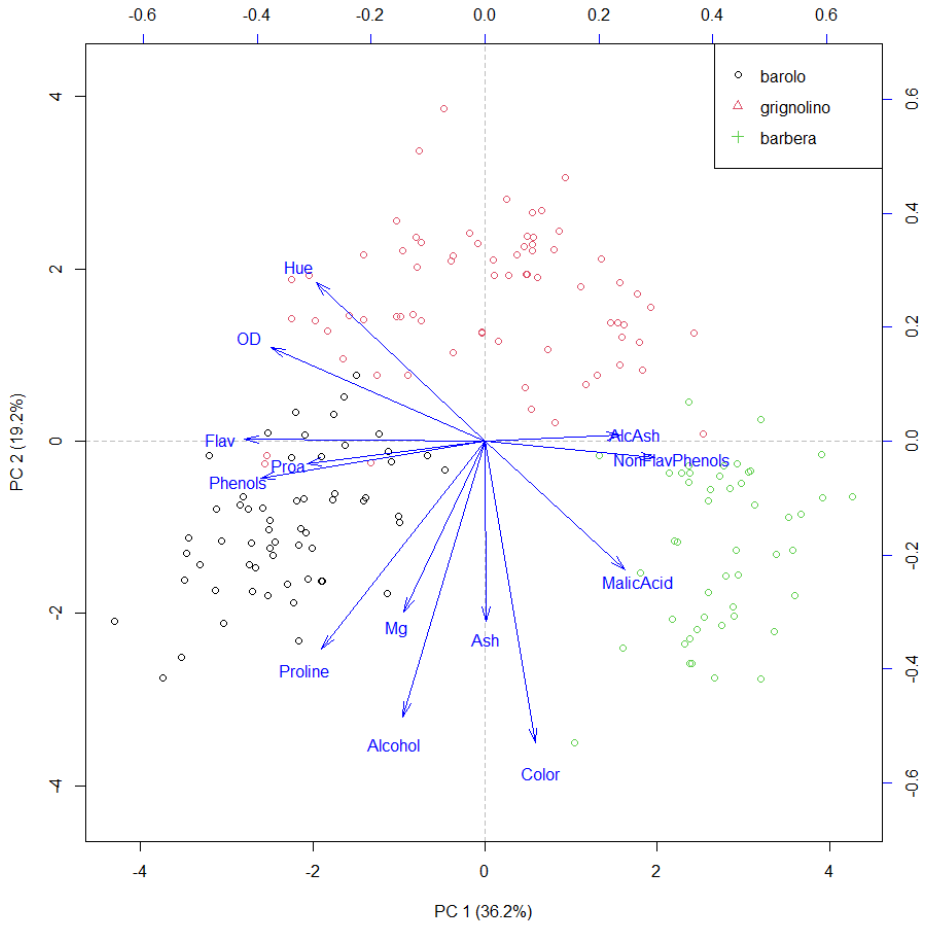
```
par(mfrow=c(1,2))
loadingplot(wines.PC, pc=c(1,3), show.names= TRUE)
loadingplot(wines.PC, pc=c(2,3), show.names= TRUE)
```



# The Biplot

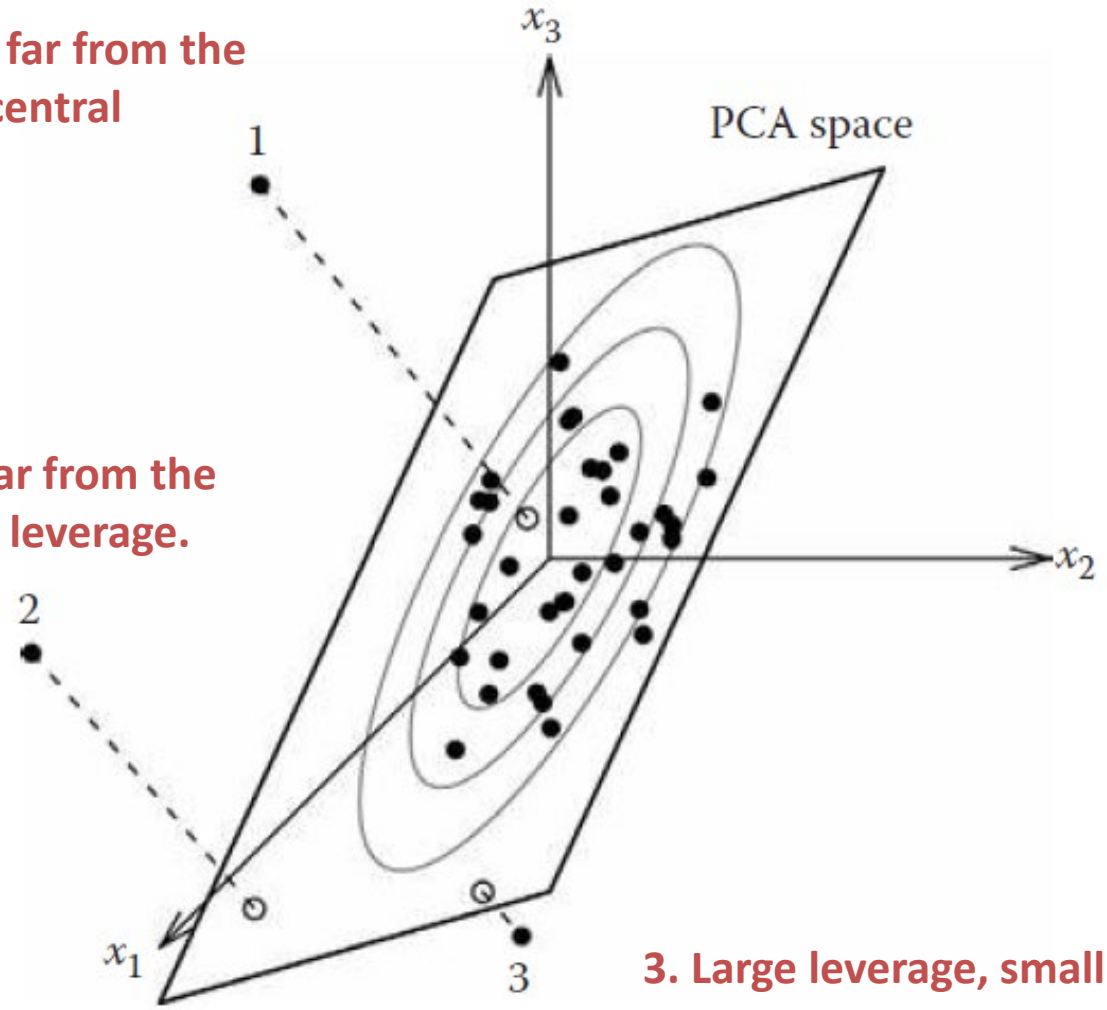
```
biplot(wines.PC, score.col = vintages, show.names = "loadings")
legend("bottomright",levels(vintages), col=1:3,pch=1:3)
```



# Diagnostic Plots – Residuals and Leverage

1. **Large residual, far from the PCA space but central**

2. **Large residual, far from the PCA space, large leverage.**

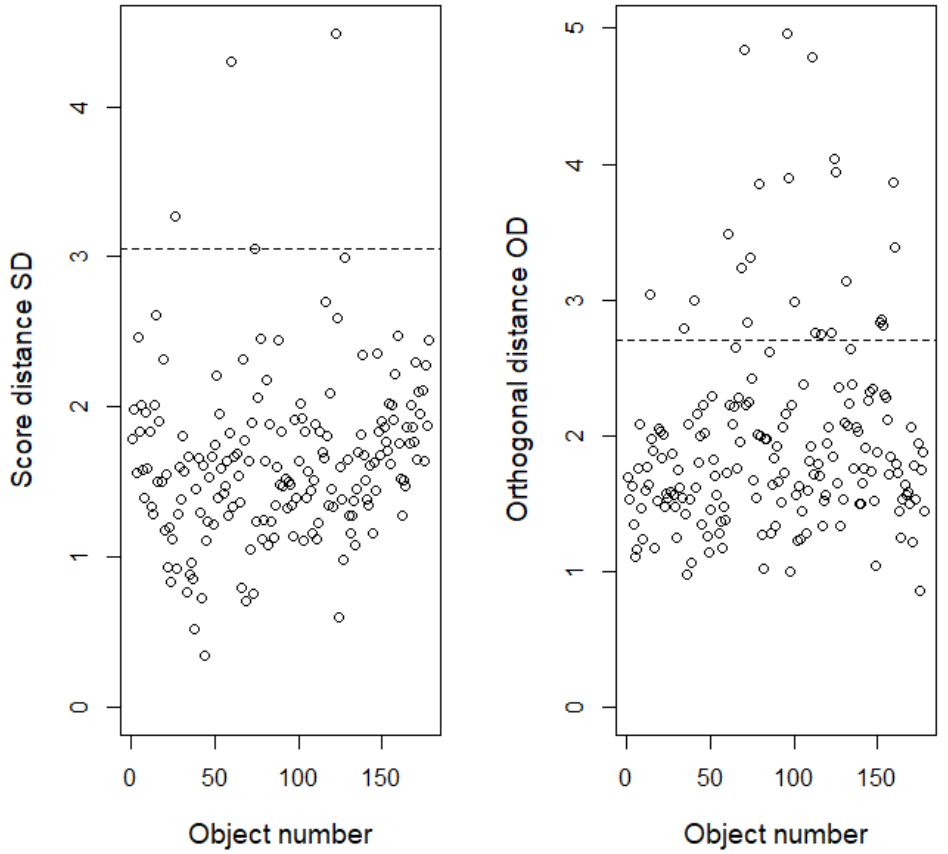



3. **Large leverage, small residual**

# Diagnostic Plots

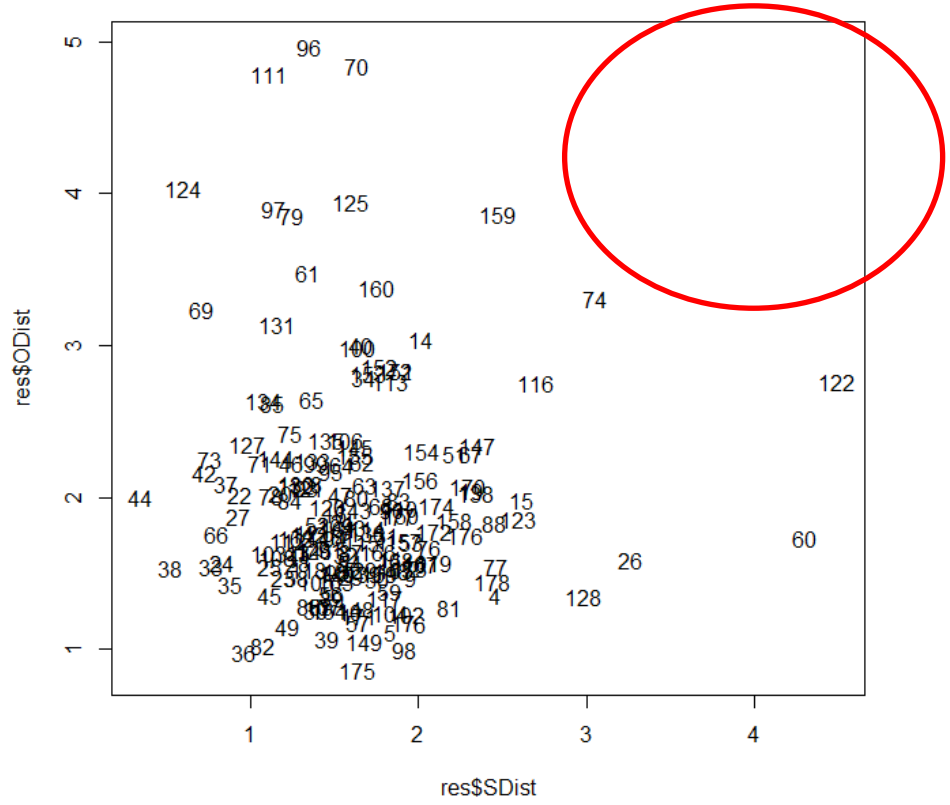
Plotting distances for 3 principal components, leverage (left) and residual (right):

```
wines.PCA<- princomp(wines, cor = TRUE)
res<-pcaDiagplot(wines, wines.PCA, a=3)
```



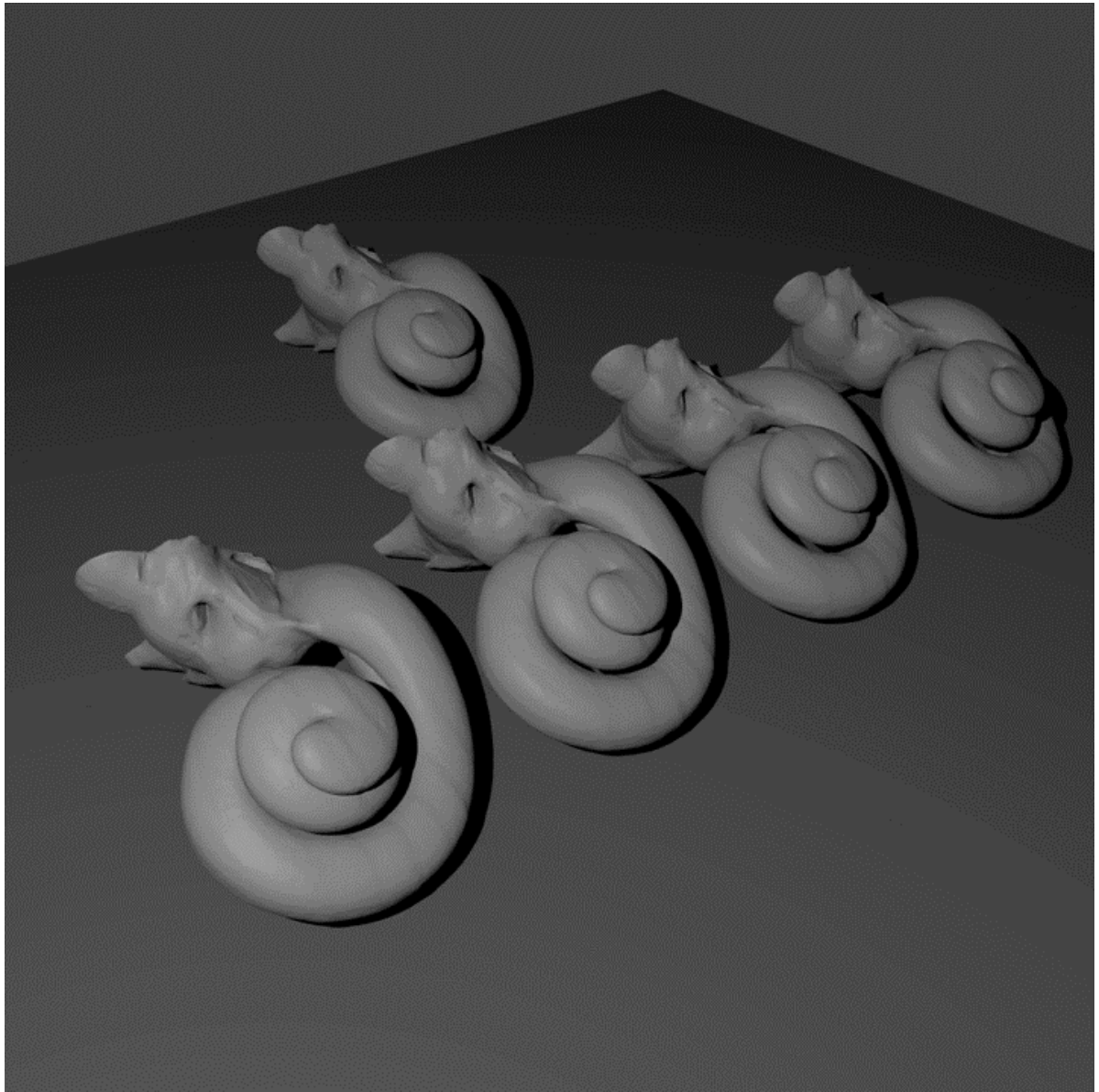
# Diagnostic Plots

```
par(mfrow=c(1,1))
plot(res$SDist, res$ODist, type="n")
text(res$SDist, res$ODist, labels=as.character(1:178))
```

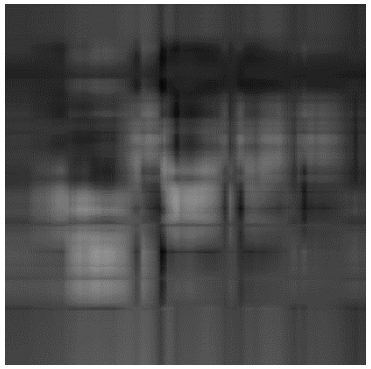



No points have high leverage and also high residual

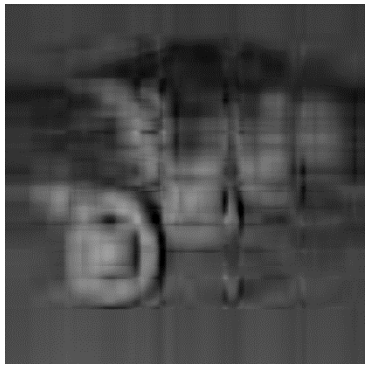
# Complex data



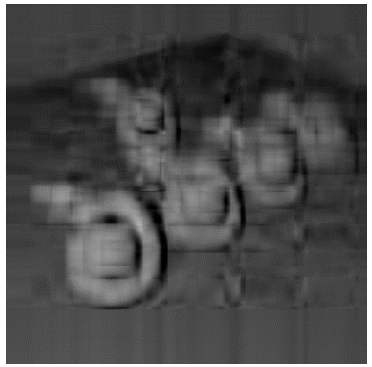
# Image reconstruction from principal components



3 PCs



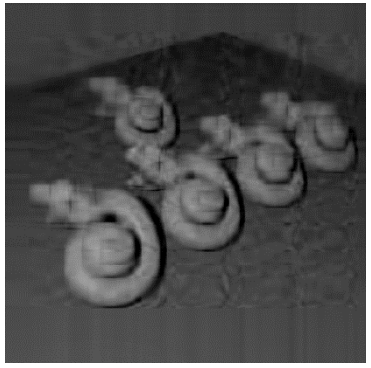
6 PCs



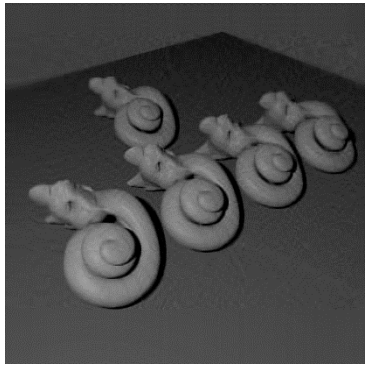
9 PCs

12PCs

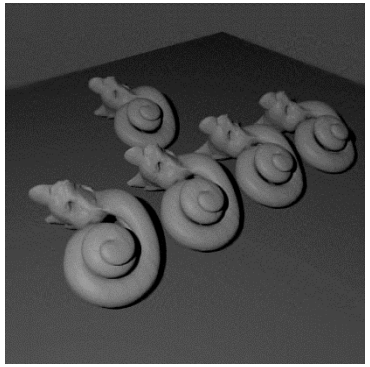
15 PCs



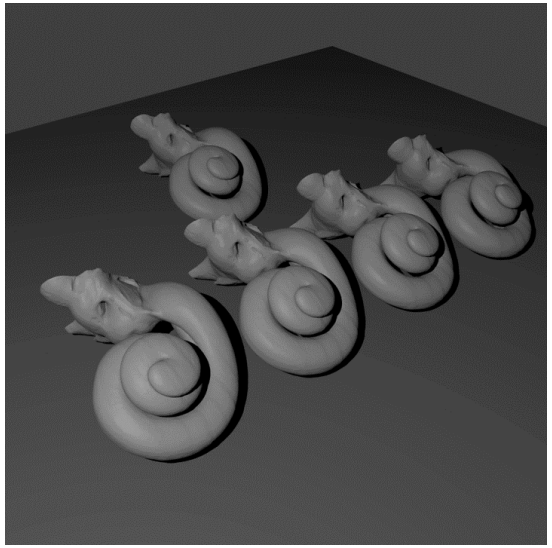
18 PCs



50 PCs



100 PCs

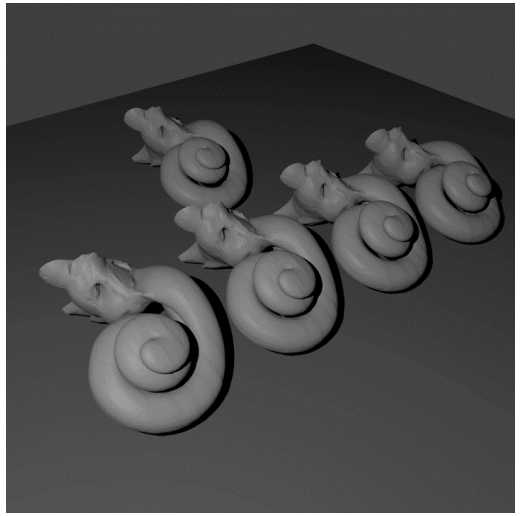


Original: 719x719 pixels

# HOW?

The image is **RGBA** coded:

- **Red**
- **Green**
- **Blue**
- **Alpha - opaqueness**



Red, green and blue is all the colors that the human eye can perceive -
Thus each of the 719x719 pixels is represented by a 4-dimensional vector – 4 layers.

**Reconstruct** the information in each layer from 3, 6, 9, 12, 15, 18, 50 and 100 principal components.

Pixel columns within layers are variables, rows observations.

- 719 principal components capture all information
- But the data is **structured**/essentially contained in a subspace, so much less will extract nearly all the information. More on this later.

# Example 6.9: Boxes with randomly generated sides

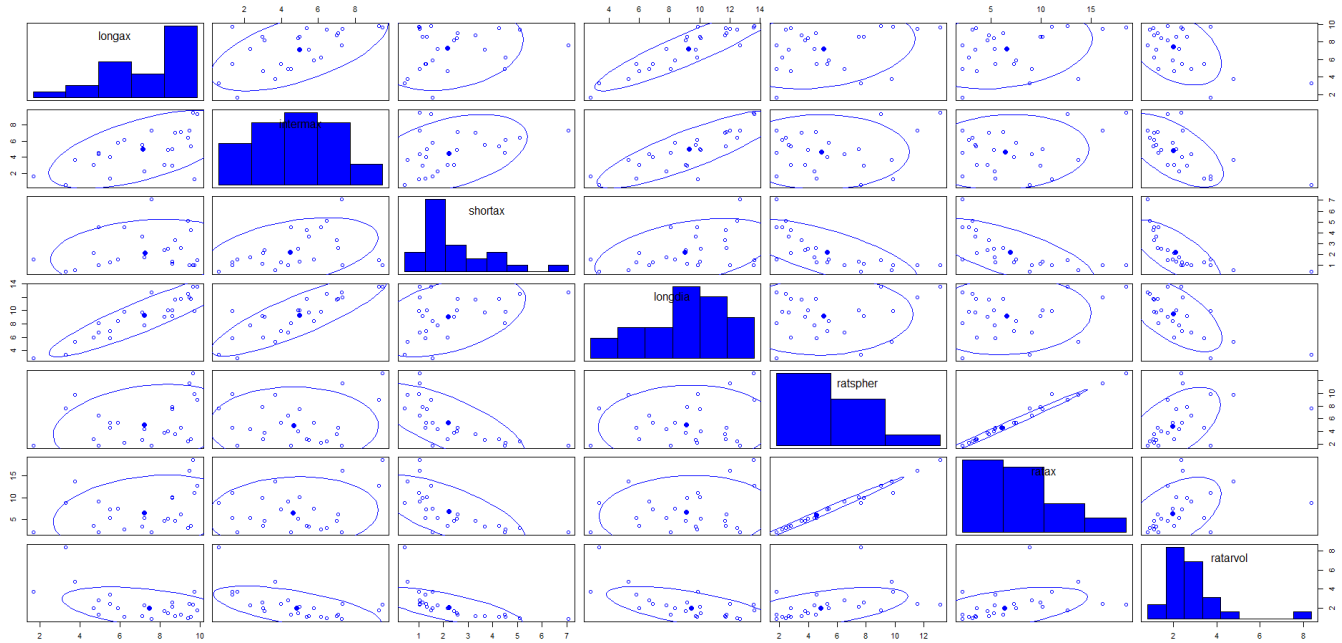
$X_1$ : longest side
$X_2$ : second longest side
$X_3$ : smallest side
$X_4$ : longest diagonal
$X_5$ : radius in the circumscribed sphere divided by radius in the inscribed sphere
$X_6$ : longest side + second longest side)/shortest side
$X_7$ : surface area/volume.

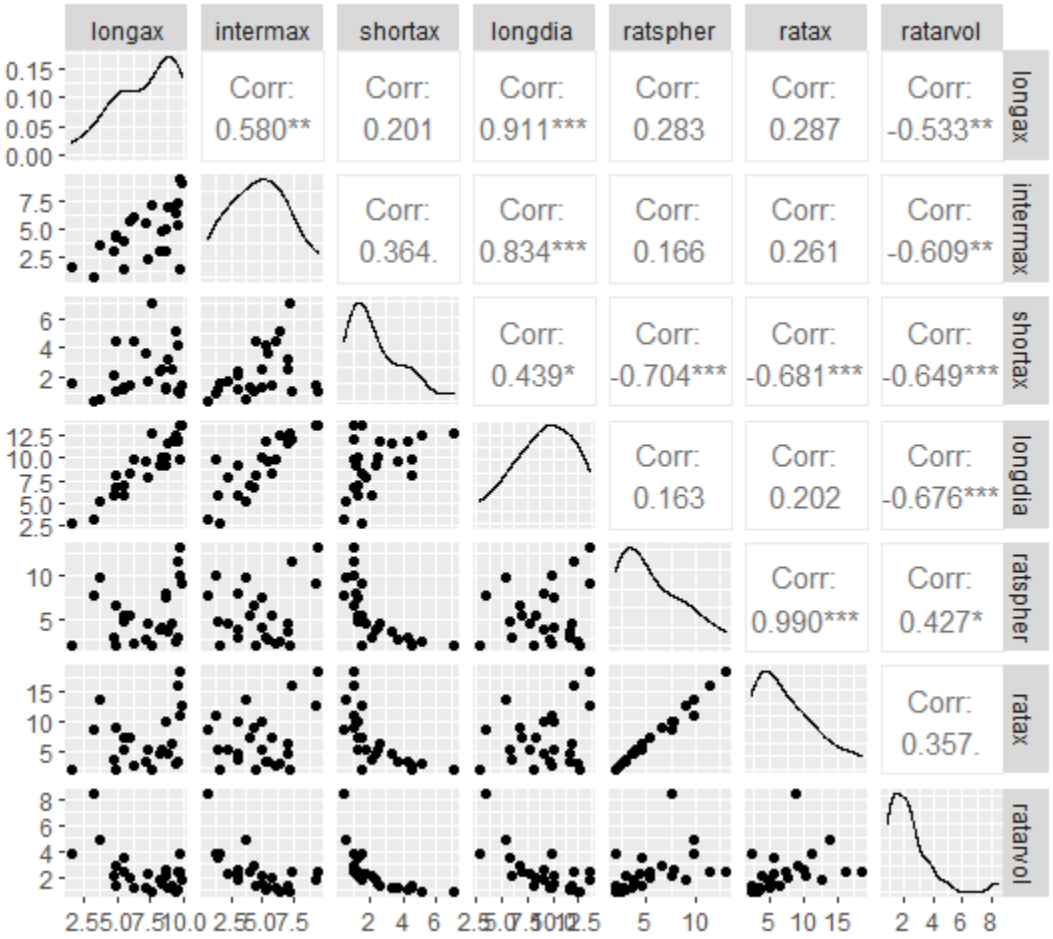
| Box | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| 1 | 3.760 | 3.660 | 0.540 | 5.275 | 9.768 | 13.741 | 4.782 |
| 2 | 8.590 | 4.990 | 1.340 | 10.022 | 7.500 | 10.162 | 2.130 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 24 | 8.210 | 3.080 | 2.420 | 9.097 | 3.753 | 4.657 | 1.719 |
| 25 | 9.410 | 6.440 | 5.110 | 12.495 | 2.446 | 3.103 | 0.914 |

From J.C. Davis. *Statistics and data analysis in geology.* John Wiley, New York 1973

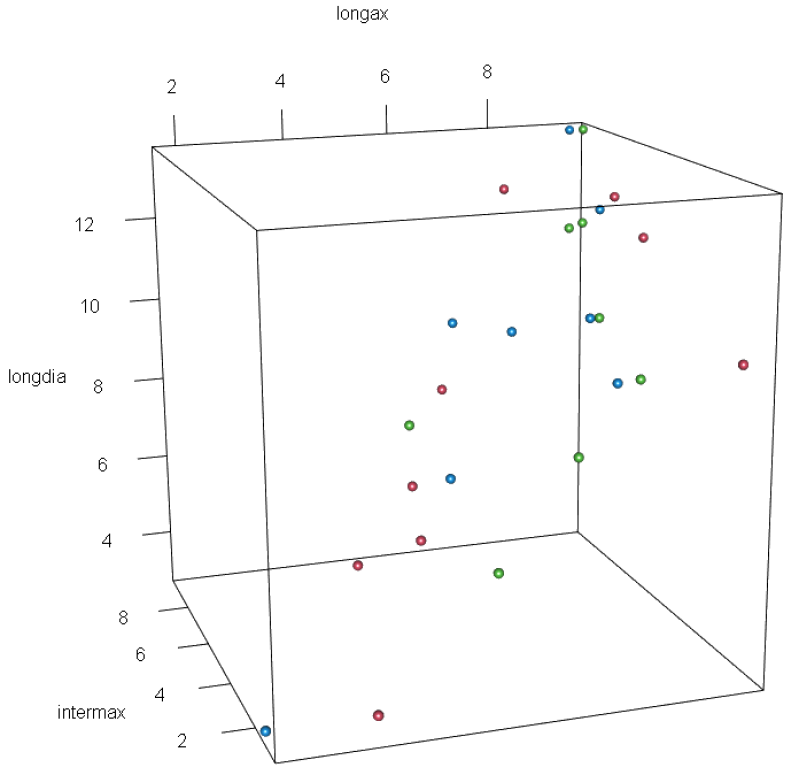
# Pairwise distribution of the 7 variables

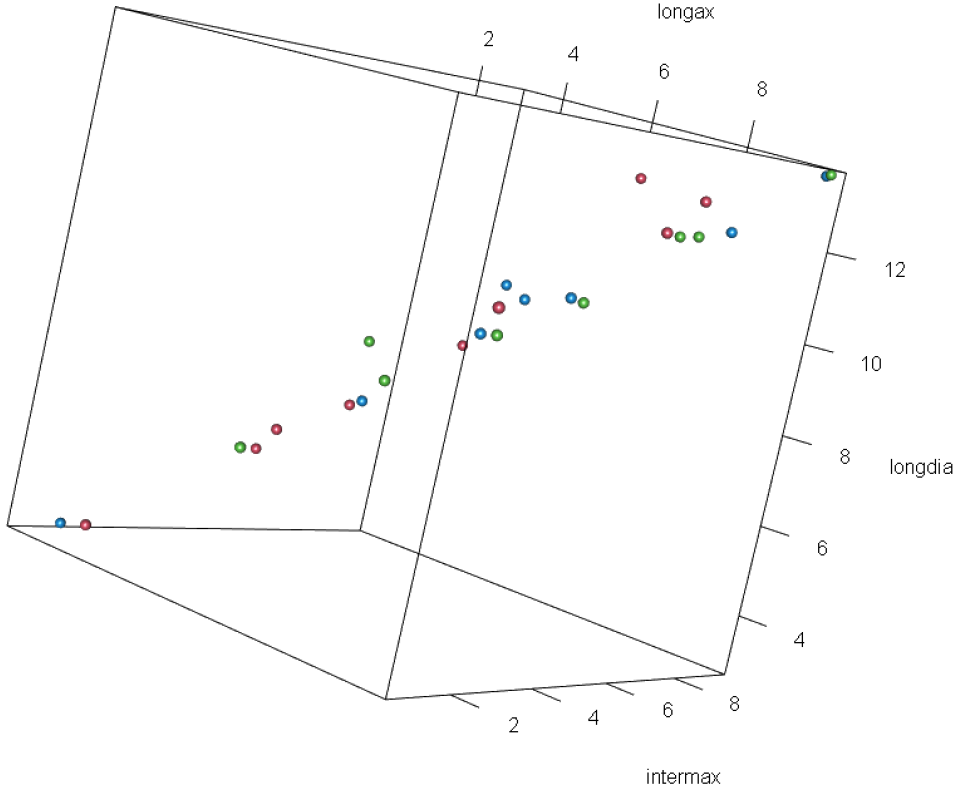
```
library(car)
scatterplotMatrix(BoxData,regLine=F,smooth=F,
  ellipse=list(levels=0.9, robust=TRUE, fill=F),
  diagonal=list(method="histogram", breaks="FD"))

library(GGally)
ggpairs(BoxData)
```



# Distribution of longax, intermax, and longdia




Seen from another angle

# Estimated variance and correlation matrices

## Boxes with randomly generated sides

```
> round(var(BoxData),digits=2)
          longax intermax shortax longdia ratspher ratax ratarvol
longax      5.40     3.26    0.78    6.39     2.16  3.04    -2.00
intermax    3.26     5.85    1.46    6.08     1.31  2.88    -2.37
shortax     0.78     1.46    2.77    2.20    -3.84 -5.17    -1.74
longdia     6.39     6.08    2.20    9.11     1.61  2.78    -3.28
ratspher    2.16     1.31   -3.84    1.61    10.71 14.77     2.25
ratax       3.04     2.88   -5.17    2.78    14.77 20.78     2.62
ratarvol   -2.00    -2.37   -1.74   -3.28     2.25  2.62     2.59
```

```
> round(cor(BoxData),digits=2)
          longax intermax shortax longdia ratspher ratax ratarvol
longax      1.00     0.58    0.20    0.91     0.28  0.29    -0.53
intermax    0.58     1.00    0.36    0.83     0.17  0.26    -0.61
shortax     0.20     0.36    1.00    0.44    -0.70 -0.68    -0.65
longdia     0.91     0.83    0.44    1.00     0.16  0.20    -0.68
ratspher    0.28     0.17   -0.70    0.16     1.00  0.99     0.43
ratax       0.29     0.26   -0.68    0.20     0.99  1.00     0.36
ratarvol   -0.53    -0.61   -0.65   -0.68     0.43  0.36     1.00
```

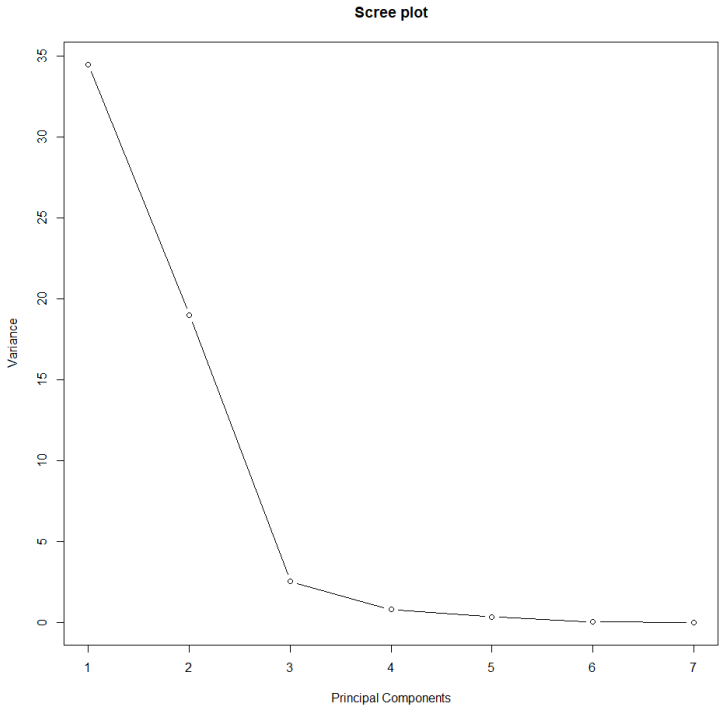
# Eigenvalues and eigenvectors of the correlation matrix

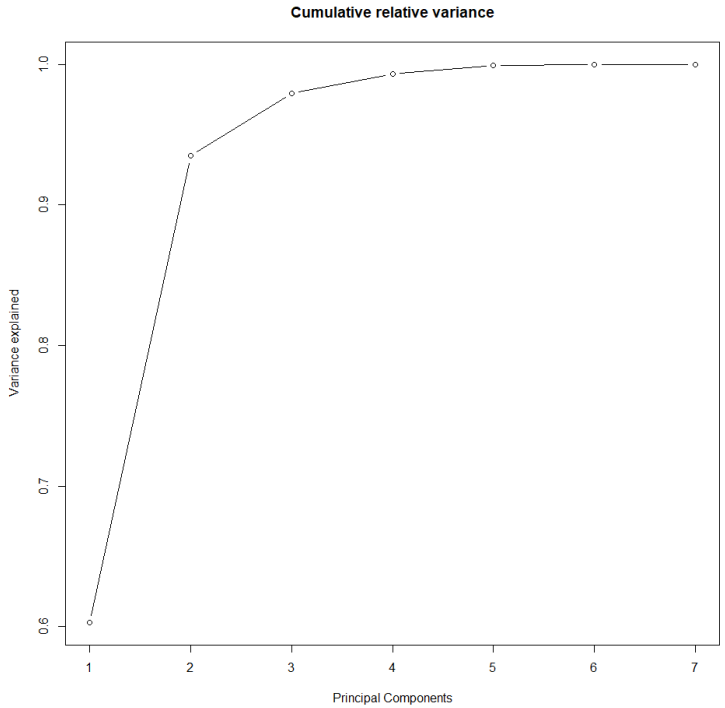
| Eigenvalues of the Correlation Matrix | | | |
|---|---|---|---|
| | Eigenvalue | Proportion | Cumulated |
| 1 | 34.49 | 0.60 | 0.60 |
| 2 | 19.00 | 0.33 | 0.93 |
| 3 | 2.54 | 0.04 | 0.98 |
| 4 | 0.81 | 0.01 | 0.99 |
| 5 | 0.34 | 0.006 | 0.999 |
| 6 | 0.03 | 0.0006 | 0.99996 |
| 7 | 0.003 | 0.00004 | 1.0000 |

$X_1$: longest side
$X_2$: second longest side
$X_3$: smallest side
$X_4$: longest diagonal
$X_5$: radius in circumscribed sphere divided by radius in inscribed sphere
$X_6$: (longest side + second longest side)/ shortest side
$X_7$: surface area/volume

| Eigenvectors | | | | | | | |
|---|---|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| longax | 0.164247 | 0.422023 | -0.644914 | 0.090057 | 0.225052 | -0.414957 | 0.384782 |
| intermax | 0.142092 | 0.447420 | 0.712955 | 0.049719 | 0.394949 | -0.066404 | 0.329324 |
| shortax | -0.172881 | 0.256798 | 0.130098 | -0.629178 | -0.607154 | -0.279860 | 0.210758 |
| longdia | 0.169853 | 0.649513 | -0.146285 | -0.212135 | 0.032942 | 0.402806 | -0.565282 |
| ratspher | 0.546410 | -0.134786 | -0.105316 | -0.164491 | -0.161154 | 0.596078 | 0.513655 |
| ratax | 0.768262 | -0.133249 | 0.148715 | 0.062070 | -0.206930 | -0.464846 | -0.327335 |
| ratarvol | 0.072837 | -0.313012 | -0.065187 | -0.719473 | 0.595697 | -0.106941 | -0.092389 |

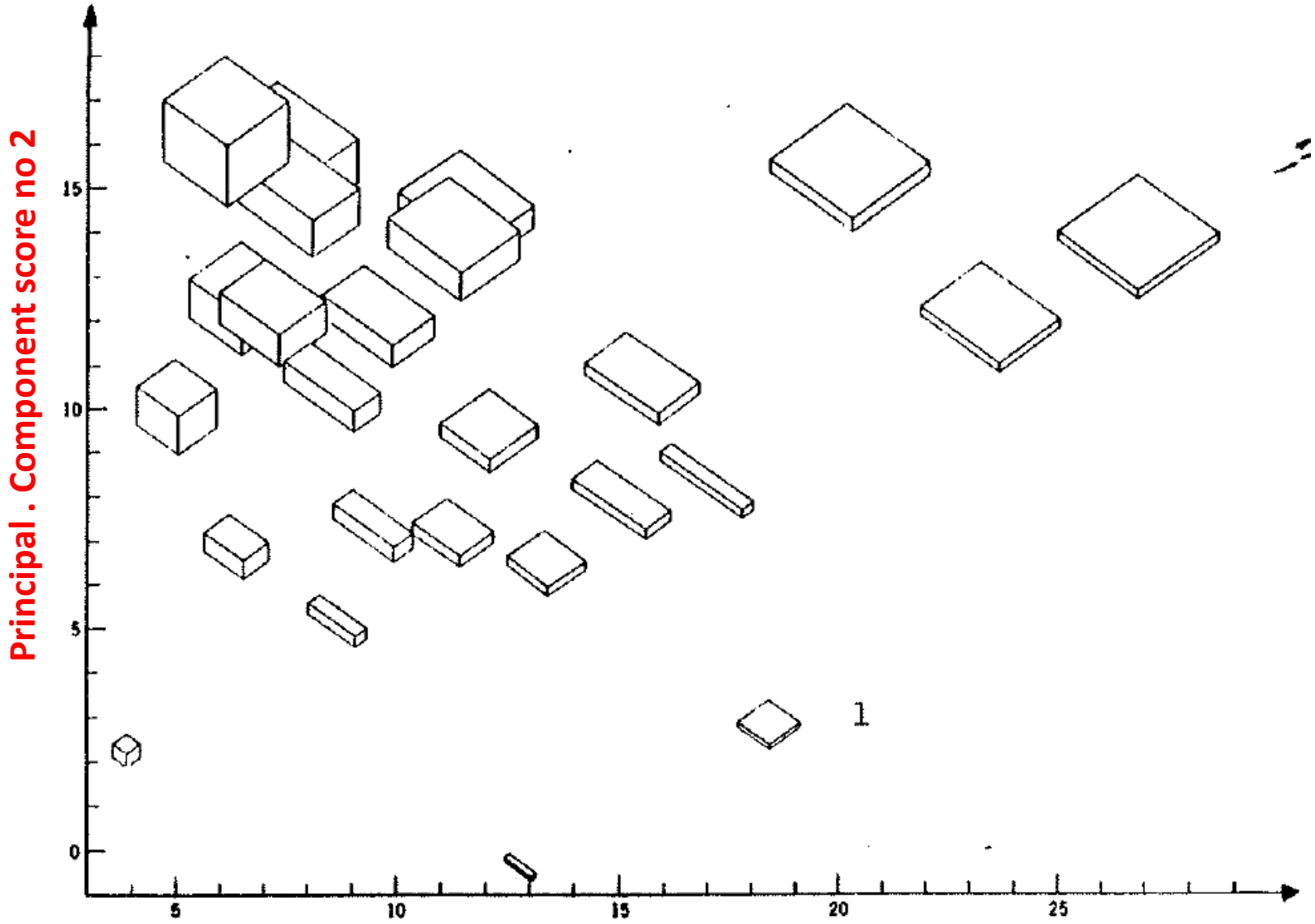
# The Scree and Variance Explained Plots




The **Scree and Variance Explained Plots** show the eigenvalues $\hat{\lambda}_i$, and the cumulative proportion of variance explained $\frac{\hat{\lambda}_1+\cdots+\hat{\lambda}_m}{\hat{\lambda}_1+\cdots+\hat{\lambda}_m+\hat{\lambda}_{m+1}+\cdots+\hat{\lambda}_k}$

# The Component Scores plot with boxes superimposed



# Reconstructing the original observations from the first m principal components

Remark 6.7: Since $\boldsymbol{X} = \boldsymbol{PY}$ awith the notation in the book, an obvious way of **reconstructing $\boldsymbol{X}$ from the first m principal components $\boldsymbol{Y}^{(m)}$** is

$$\boldsymbol{X}^* = \boldsymbol{P}^{(m)}\boldsymbol{Y}^{(m)} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_m]\begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix} = Y_1\boldsymbol{p}_1 + \cdots + Y_m\boldsymbol{p}_m$$

The variance matrix of the reconstructed vector is

$$\begin{aligned} \mathrm{D}(\boldsymbol{X}^*) &= \mathbf{P}\,\mathrm{D}(\boldsymbol{Y}^*)\mathbf{P}^T \\ &= \lambda_1\boldsymbol{p}_1\boldsymbol{p}_1^T + \cdots + \lambda_m\boldsymbol{p}_m\boldsymbol{p}_m^T. \end{aligned}$$

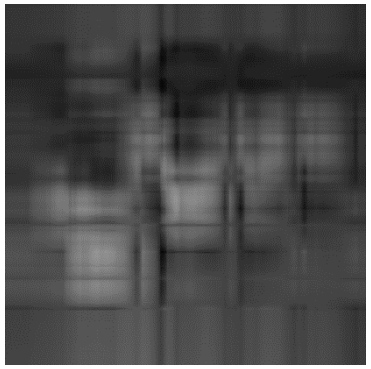
The spectral decomposition of $\boldsymbol{\Sigma}$ is (p. 458)

$$\boldsymbol{\Sigma} = \lambda_1\boldsymbol{p}_1\boldsymbol{p}_1^T + \cdots + \lambda_m\boldsymbol{p}_m\boldsymbol{p}_m^T + \lambda_{m+1}\boldsymbol{p}_{m+1}\boldsymbol{p}_{m+1}^T + \cdots + \lambda_k\boldsymbol{p}_k\boldsymbol{p}_k^T,$$

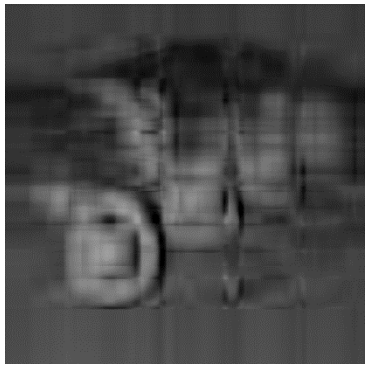
which means that

$$\boldsymbol{\Sigma} - \mathrm{D}(\boldsymbol{X}^*) = \lambda_{m+1}\boldsymbol{p}_{m+1}\boldsymbol{p}_{m+1}^T + \cdots + \lambda_k\boldsymbol{p}_k\boldsymbol{p}_k^T.$$

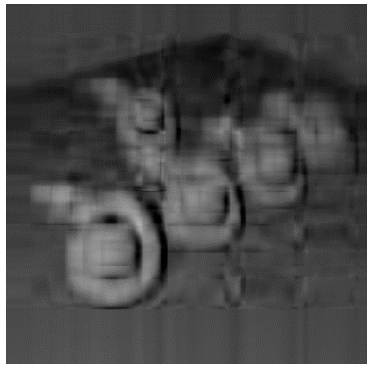
# Image reconstruction from principal components –code in script.



3 PCs

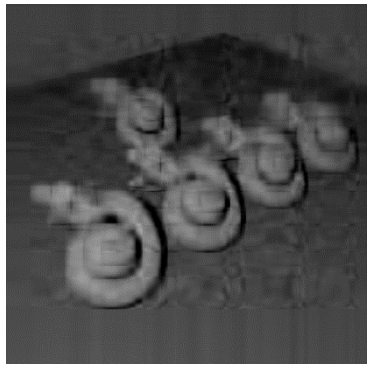


6 PCs

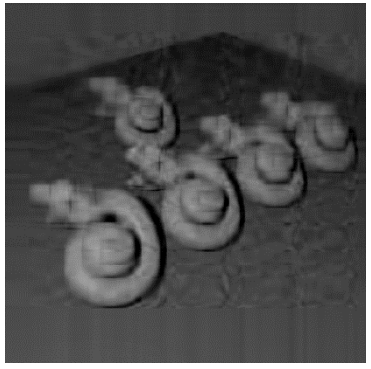


9 PCs

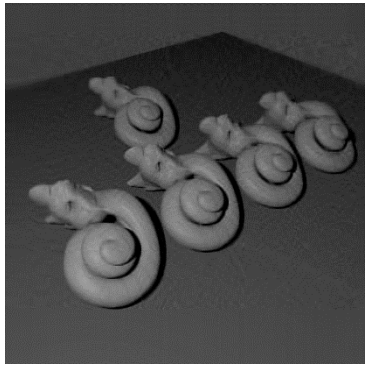
12PCs



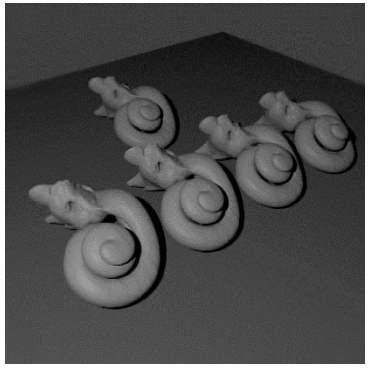
15 PCs



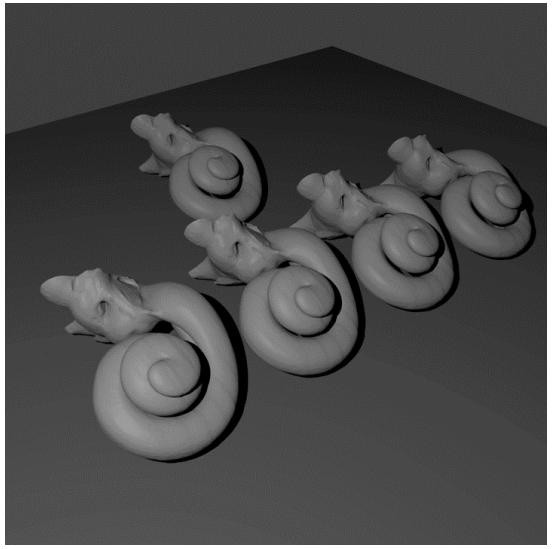
18 PCs



50 PCs



100 PCs



Original: 719x719 pixels

# Test for equality of k-m smallest eigenvalues in the variance matrix

$$H_0 : \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

$$Z_1 = -n^* \log \frac{\det \hat{\mathbf{\Sigma}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n^* \log \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$n^* = n - m - \frac{1}{6}(2(k-m) + 1 + \frac{2}{k-m}),$$

and

$$\hat{\lambda}_* = (\operatorname{tr} \hat{\mathbf{\Sigma}} - \hat{\lambda_1} - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m).$$

The critical region using a test at level $\alpha$ is approximately

$$\{(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) | z_1 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

# Test for equality of k-m smallest eigenvalues in the correlation matrix

$$H_0 : \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

If we instead are using the estimated ***correlation matrix*** $\hat{\mathbf{R}}$ we get the criterion

$$Z_2 = -n \log \frac{\det \hat{\mathbf{R}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n \log \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$\hat{\lambda}_* = (k - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m).$$

The critical region for a test at level $\alpha$ becomes approximately equal to

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | z_2 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

# Test for equality of k-m smallest eigenvalues in BoxData correlation matrix

| Eigenvalues of the Correlation Matrix | | | |
|---|---|---|---|
| | Eigenvalue | Proportion | Cumulated |
| 1 | 3.39 | 0.48 | 0.48 |
| 2 | 2.81 | 0.40 | 0.89 |
| 3 | 0.44 | 0.06 | 0.95 |
| 4 | 0.28 | 0.04 | 0.99 |
| 5 | 0.08 | 0.01 | 0.999 |
| 6 | 0.003 | 0.0005 | 0.99996 |
| 7 | 0.0003 | 0.00004 | 1 |

Can vi assume that the last two values are the same (close to 0)?

$$\prod_{i=6}^{7} \hat{\lambda}_i = 8.942424e-07, \qquad \hat{\lambda}_* = \frac{(\hat{\lambda}_6 + \hat{\lambda}_7)}{2} = 0.001834764, \qquad \mathrm{n} = 25, \qquad -\mathrm{n} * \log\left(\frac{\prod_{i=6}^{7} \hat{\lambda}_i}{\hat{\lambda}_*^2}\right) = 33.14026$$

$$P\left(\chi^2(2) > 33.14\right) = 0.0000006$$

**Conclusion: The statistical test shows that data do not sypport that the 2 smallest eigenvalues may be assumed to be equal.**

# Exercises

- Exercise 7.1
  - Principal Component Analysis on Olympic data
  - Interpretation, number of components
- Exercise 7.2
  - Principal Component Analysis on the 'Fitness' dataset
  - Interpretation, number of components, conditioning on variables
- Exercise 7.3
  - Principal Component Analysis on Beef characterization
  - Interpretation of PCA
- Exercise 7.4
  - PCA by hand and interpretation