Course: Selected Topics in Statistics - Statistical Analysis of Dependent Data With special emphasis on basics of multivariate data analysis

Jürgen Pilz

Institut für Statistik, Universität Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
juergen.pilz@aau.at

www.jpilz.net

Master Study Program in Mathematics
University of Klagenfurt / Klagenfurt, Austria

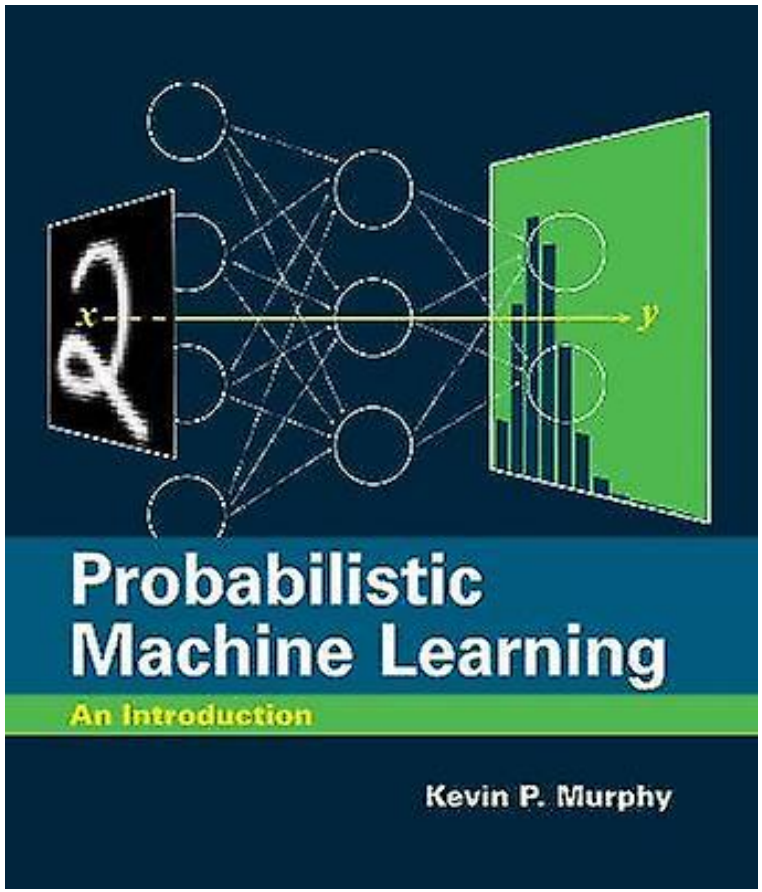
0. What is multivariate data analysis about

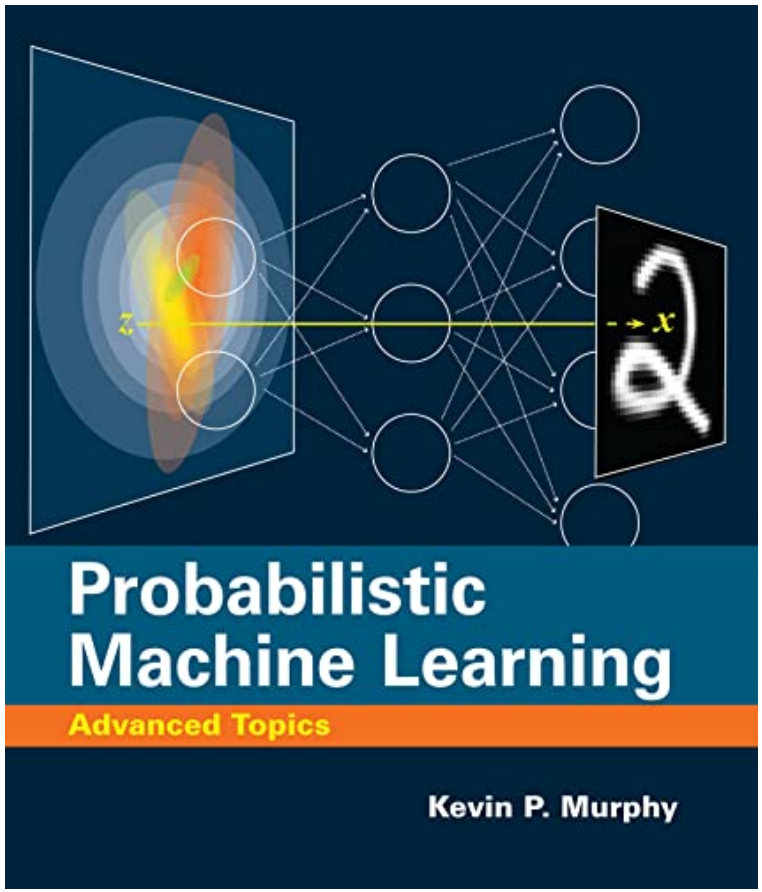
Start our journey through multivariate statistics with its main objectives:

- Study of relationships among $p > 1$ feature variables
- Recognition of essential associations (patterns, dependencies)
- Dimension reduction (high-dimensional feature spaces of genome and neuro-signal data, spatio-temporal models and LLMs)
- Data compression (audio, image and video data, MPEG and JPEG,...)
- MVDA forms the core of modern engineering and social science branches such as "Data Mining", "Big Data Analysis", "Predictive Analytics", "Knowledge Discovery", "(Deep) Machine Learning", "Data Science", "AI and Robotics",...)

Perfect textbook

for interfacing Statistical Learning and (Deep) Probabilistic Machine Learning





K.P. Murphy: Probabilistic Machine Learning. MIT Press 2022, 2023

1. Multivariate Data and Distributions

1.1 Data matrix

We consider $p > 1$ random variables $X_1, \ldots, X_p$, and collect them in a p-dimensional **random vector** $\underline{X} = (X_1, \ldots, X_p)^T$. This is a measurable mapping

$$\underline{X} : \Omega \longrightarrow \mathbb{R}^p$$
$$[\Omega, \mathscr{B}, \mathbb{P}] \longrightarrow [\mathbb{R}^p, \mathscr{B}_p, \mathbb{P}_{\underline{X}}],$$

$X_1, \ldots, X_p : \Omega \longrightarrow \mathbb{R}$ are called the components of $\underline{X}$, $\mathscr{B}_p$ denotes the Borel-σ-algebra of $\mathbb{R}^p$ and $\mathbb{P}_{\underline{X}}$ the induced probability measure.

The realizations of $\underline{X}$ are denoted as

$\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$,

This means: for n objects (individuals) we are recording p features, each. The data matrix then reads

Data matrix

$$\underset{(n,p)}{X} = \begin{pmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{matrix} \text{sample 1} \\ \text{sample 2} \\ \vdots \\ \text{sample } n \end{matrix} .$$

Examples:

1) Medical records of n patients: age, weight, height, blood pressure, cholesterol,... $\widehat{=}$ p variables of n patients
2) Soil samples at n different locations: element contents (N, P, Mg, K, Ca,...), pH-value, humidity,...
3) Weekly closing values of p different stocks (assets) within a year ($n = 52$, $p =$ number of stocks (assets) in a portfolio)
4) Stress/performance test values of n members of some occupational group under p different stress situations, measurements at a discrete scale, usually: $1, 2, 3, 4, 5$ (e.g. $1 =$ low stress, ..., $5 =$ high stress; i.e. $p = 5$).

Data matrix

In practical situations, we often have a mixture of qualitative (categorical) and quantitative variables.

Critical assumption:
Independence of the samples, i.e. the realizations are independent copies of $\underline{X}$

$$\underline{X}_i \overset{i.i.d}{\sim} F_{\underline{X}} \ (i = 1, \ldots, n)$$

Note: The components $X_1, \ldots, X_p$ of $\underline{X}$ are, in general, not independent. They might well be correlated!
I.e., in general, the covariance matrix is non-diagonal, $\mathrm{Cov}(\underline{X}) \neq I_p$.

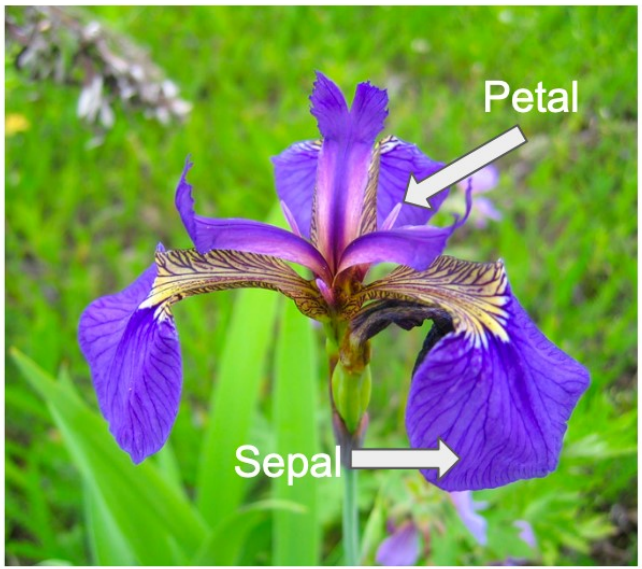
1.2 Graphical displays

We make use of the iris data from the R-package datasets. This is comprised of measurements of four variables of the iris flower: sepal length, sepal width and petal length, petal width (sepal=Kelchblatt, petal=Blütenblatt).

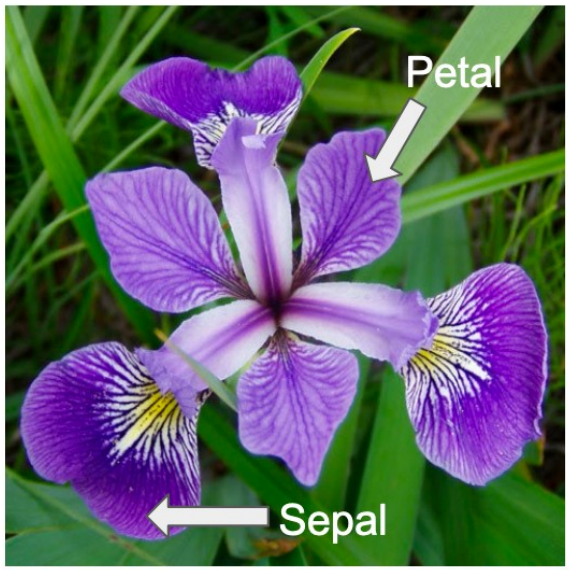
Iris species

There are 50 meaurements for each of the three types (setosa, versicolor, virginica) of the iris flower.

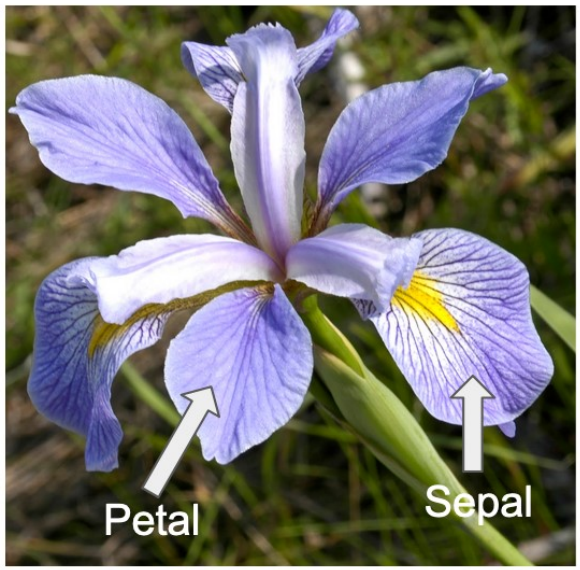
Iris setosa



Iris versicolor



Iris virginica



Data matrix

```
> data(iris)
> head(iris, 5)
  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
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
> str(iris)
'data.frame': 150 obs. of 5 variables:
$ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
$ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
$ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
$ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 ...
```

Data matrix

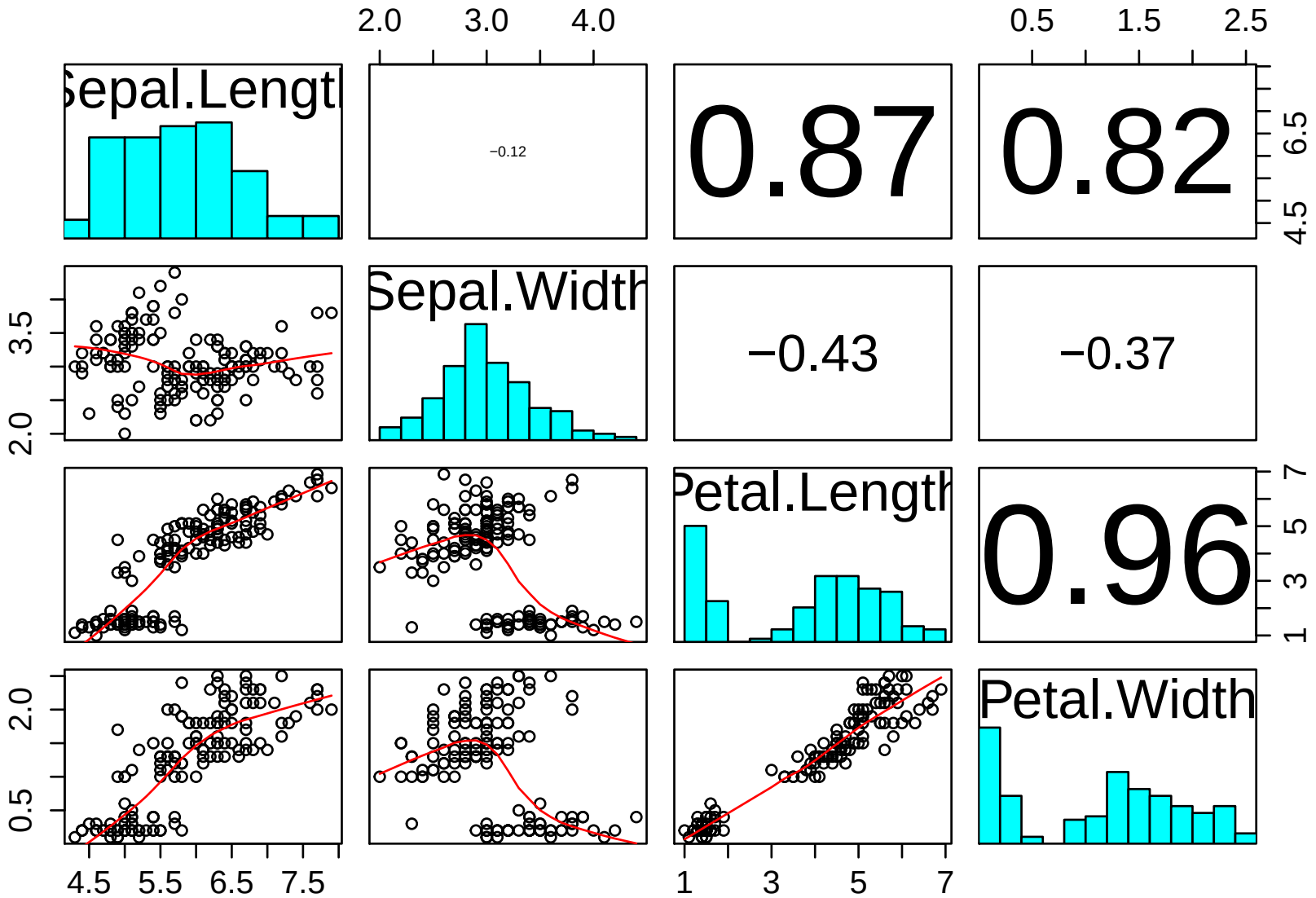
```
> summary(iris[, -5])
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300
Median :5.800	Median :3.000	Median :4.350	Median :1.300
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500

a) Pairwise Scatterplots

```
> pairs(iris[,-5], diag.panel=panel.hist, upper.panel=panel.cor,
+ lower.panel=panel.smooth)
```

Scatterplot with panels



Scatterplot with Species

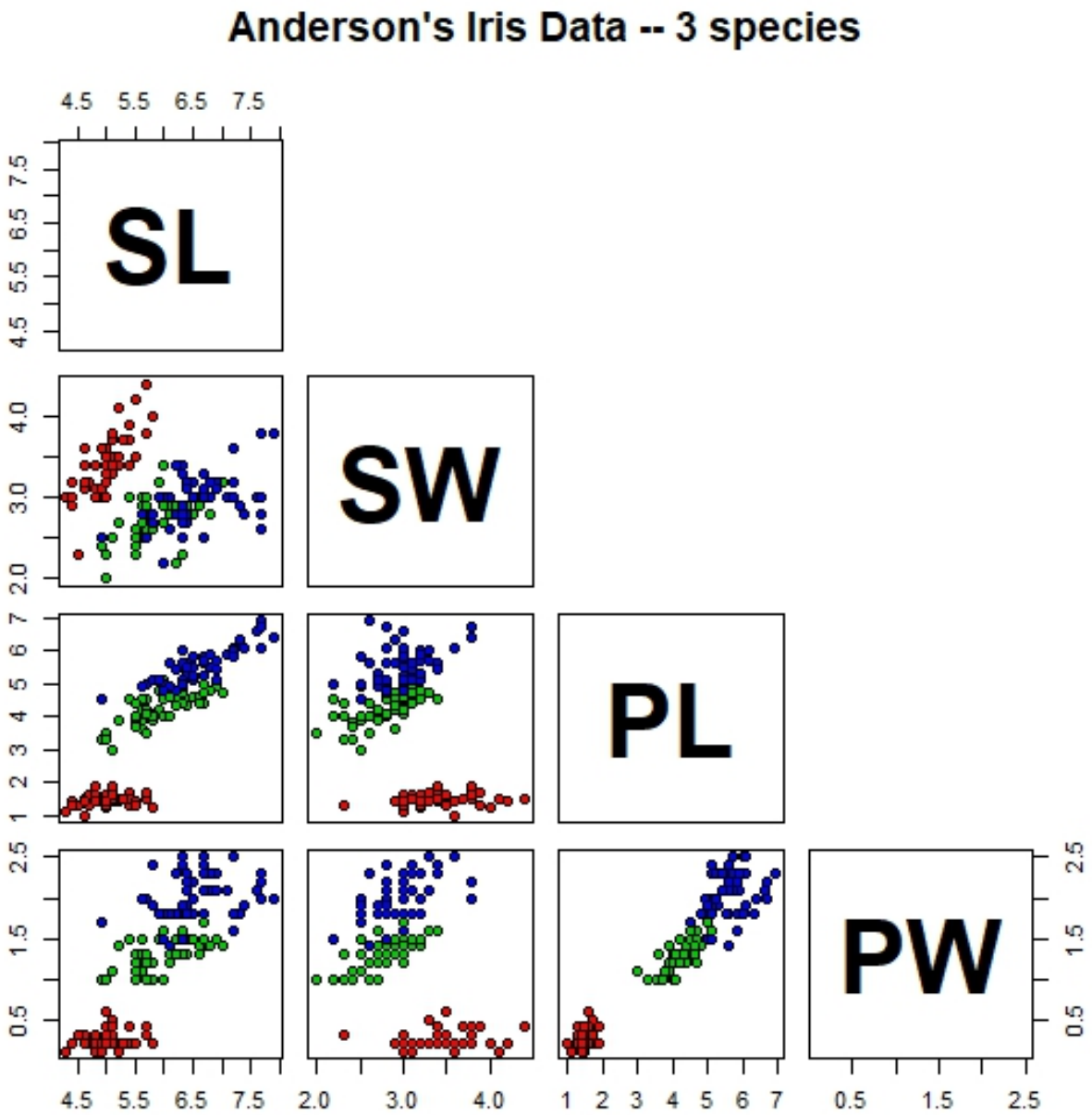
```
> pairs(iris[,-5], main = "Anderson's Iris Data – 3 species", pch = 21,
+ bg = c("red", "green", "blue")[unclass(iris$Species)],
+ upper.panel=NULL, labels=c("SL","SW","PL","PW"),
+ font.labels=2, cex.labels=4.5)
```

The resulting plot can be seen on the next slide.

There are many other options for pairwise scatterplots, in paticular, when we make use of

```
> library(ggplot2)
```

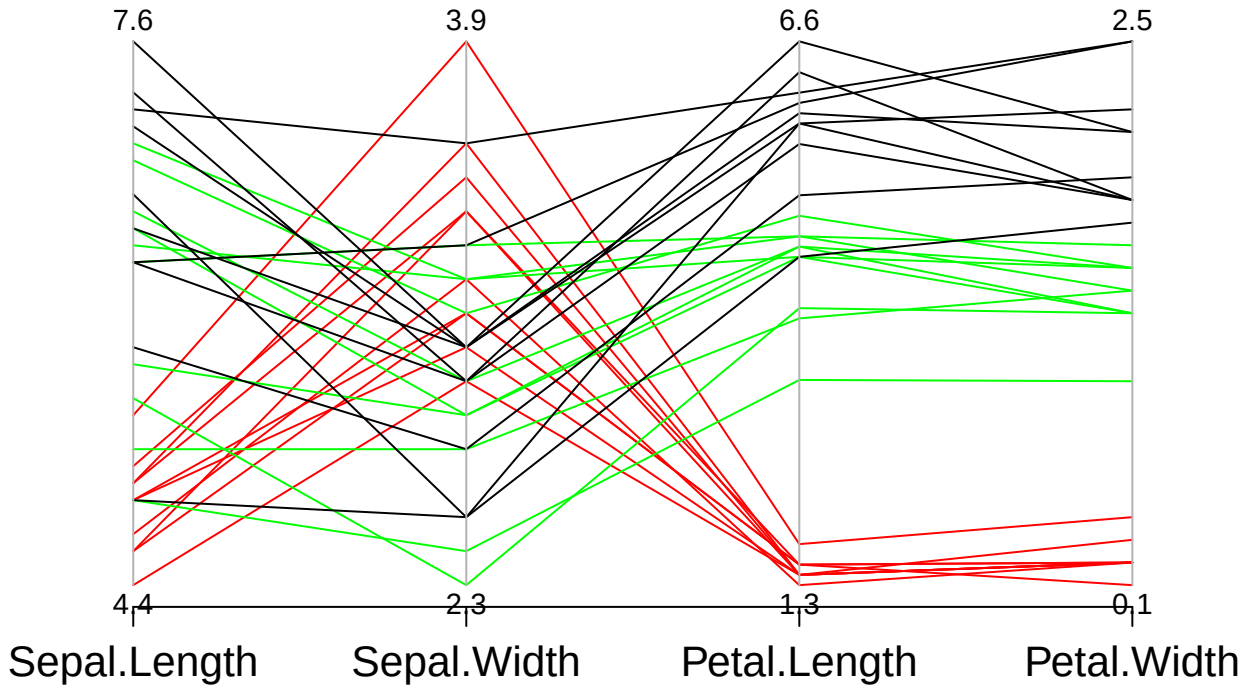
Scatterplot with Species



Scatterplot with Species

b) Parallel Coordinates Plot

```
> library(MASS)
> parcoord(iris[c(1:10,51:60,101:110), -5], var.label=T,
+ col=c(rep("red",10), rep("green",10),rep("black",10)), lwd= 1.8)
```



c) Andrews Plot

Andrews (1972) proposed to code multivariate observations $\underline{x}_i = (x_{i1}, \ldots, x_{ip})^T; i = 1, \ldots, n$ via curves of the following form:

$$\underset{-\pi \le t \le \pi}{f_i(t)} = \begin{cases} \frac{1}{\sqrt{2}} x_{i1} + \sum_{k=1}^{\frac{p-1}{2}} \left\{ x_{i,2k} \sin(kt) + x_{i,2k+1} \cos(kt) \right\} & \text{if } p \text{ is odd} \\ \frac{1}{\sqrt{2}} x_{i1} + \sum_{k=1}^{\frac{p}{2}} x_{i,2k} \sin(kt) + \sum_{k=1}^{\frac{p}{2}-1} x_{i,2k+1} \cos(kt) & \text{if } p \text{ is even.} \end{cases}$$

Thus, the observations are acting as Fourier coefficients.

Disadvantage: The shape of the curves strongly depends on the ordering of the variables $x_1, \ldots, x_p$. For increasing p, the last variables contribute only little (falling into the high frequency region of the curve).

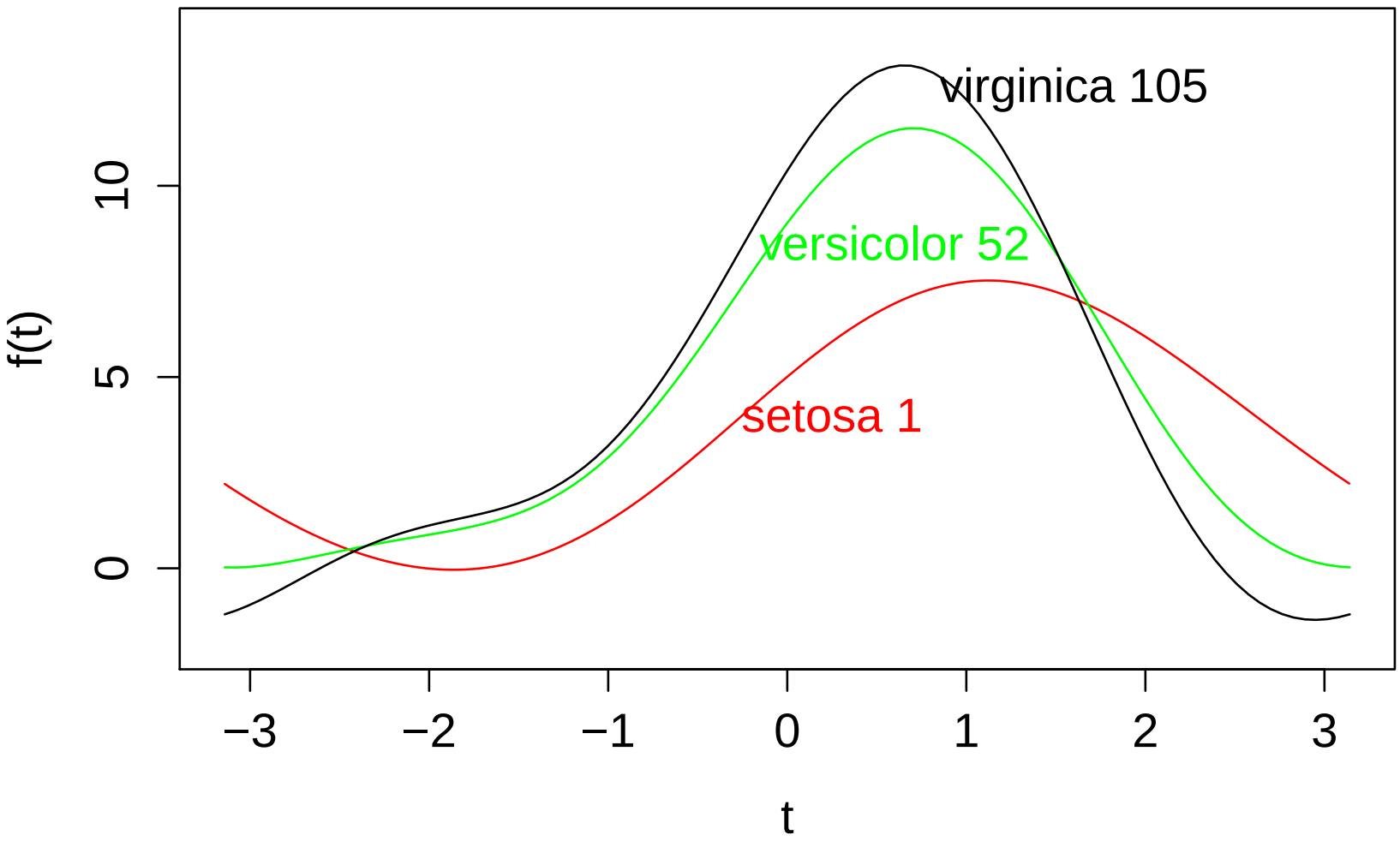
Consequently, Andrews-Plots only make sense for "ordered" variables and moderate sample sizes n.

Andrews Curves

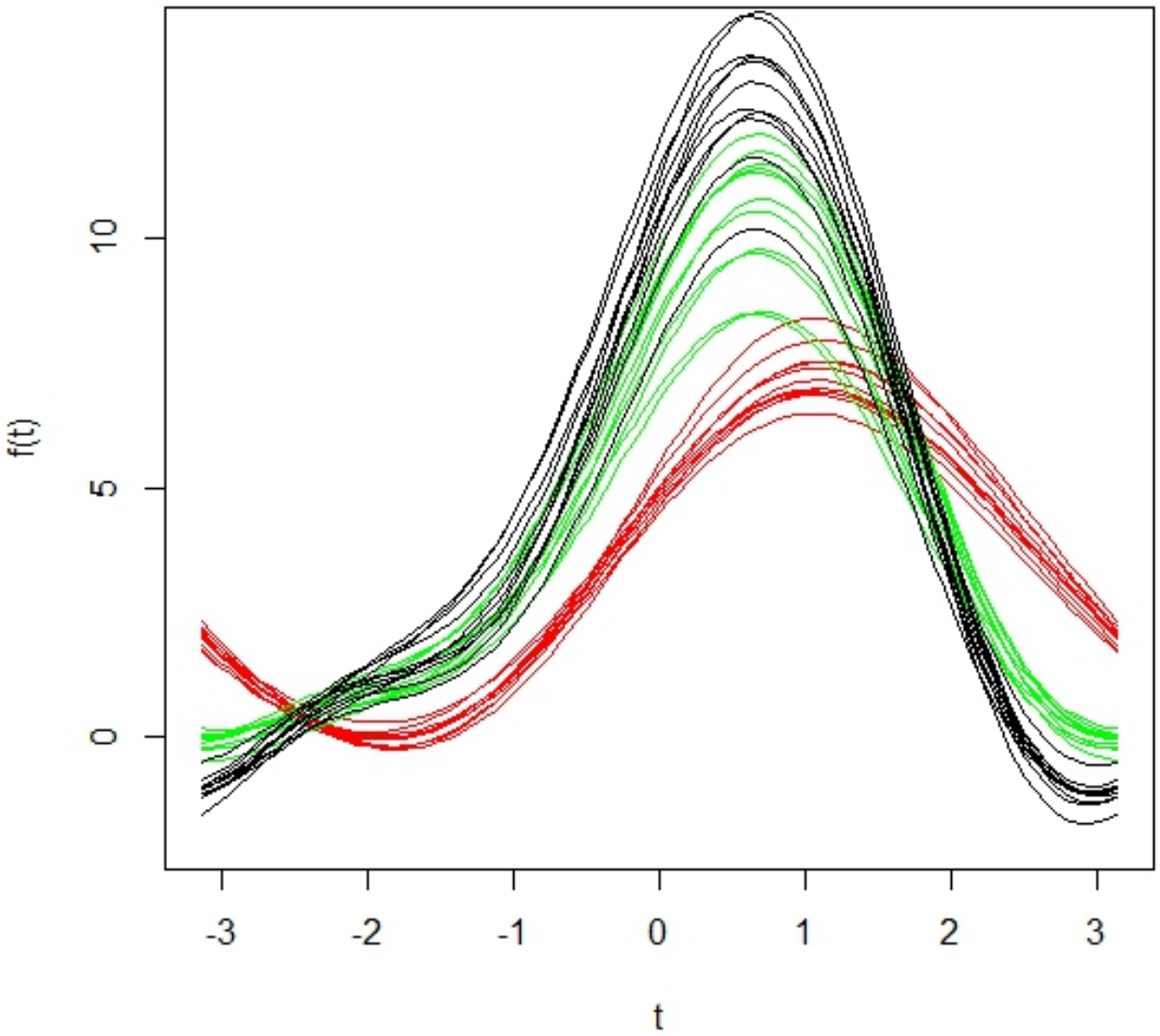
The following code produces the Andrews curves for three selected observations of the iris data ($p = 4$):

```
> t=seq(from=-pi, to=pi, by=0.02)
> plot(t, iris[1,1]/sqrt(2)+iris[1,2]*sin(t)+iris[1,3]*cos(t)+iris[1,4]*sin(2*t),
+ type="l", ylab="f(t)", ylim=c(-2,14), xlim=c(-3.1415,3.1415),
+ col="red")
> curve(iris[52,1]/sqrt(2)+iris[52,2]*sin(t)+iris[52,3]*cos(t)+iris[52,4]*
+ sin(2*t), type="l", xname="t", add=TRUE, from=-pi, to=pi,
+ col="green")
> curve(iris[105,1]/sqrt(2)+iris[105,2]*sin(t)+iris[105,3]*cos(t)+
+ iris[105,4]*sin(2*t), type="l", xname="t",
+ add=TRUE, from=-pi, to=pi)
> text(1.6,12.5, "virginica 105")
> text(0.6,8.5, "versicolor 52", col="green")
> text(0.25,4, "setosa 1", col="red")
```

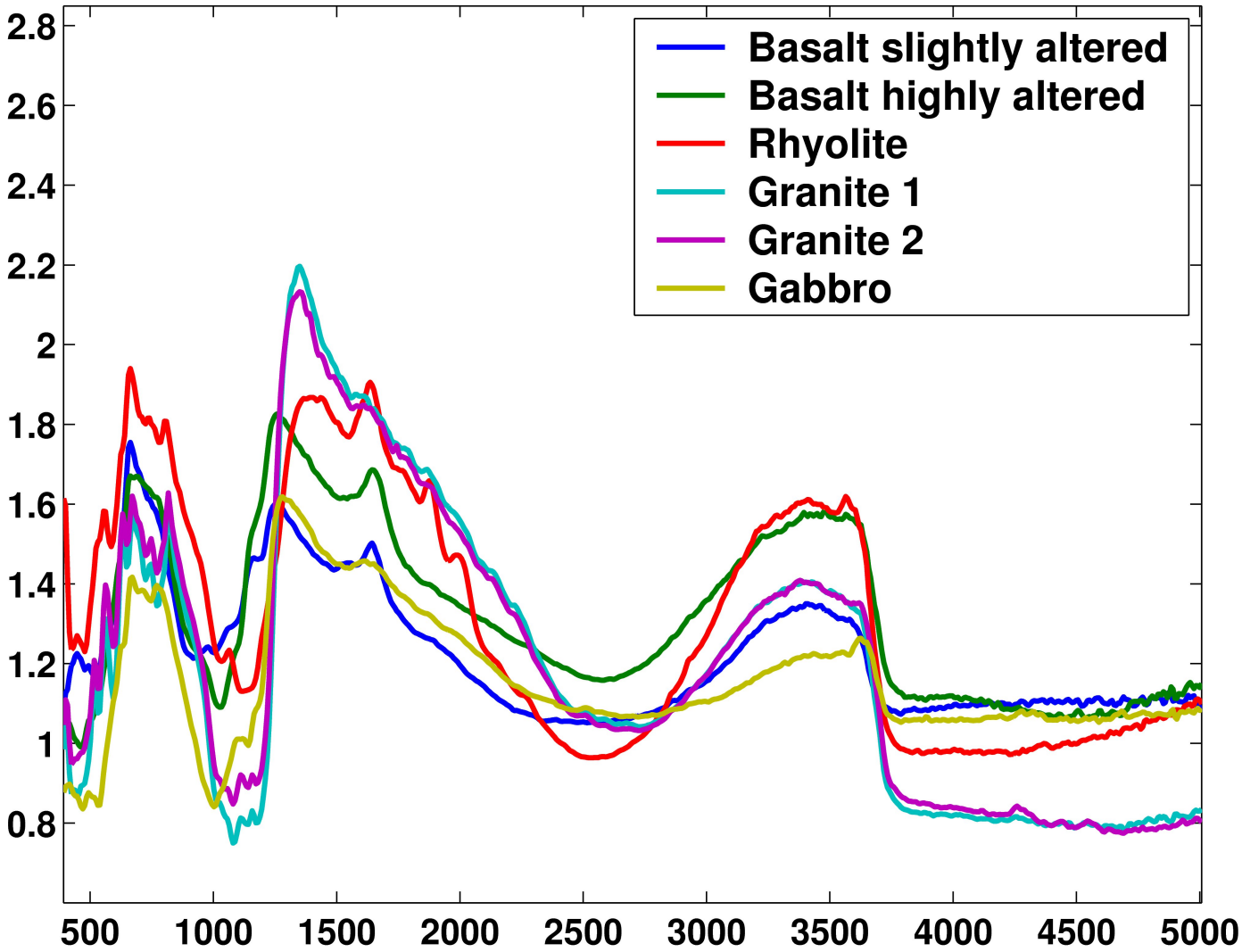
Andrews Curves



More Andrews Curves



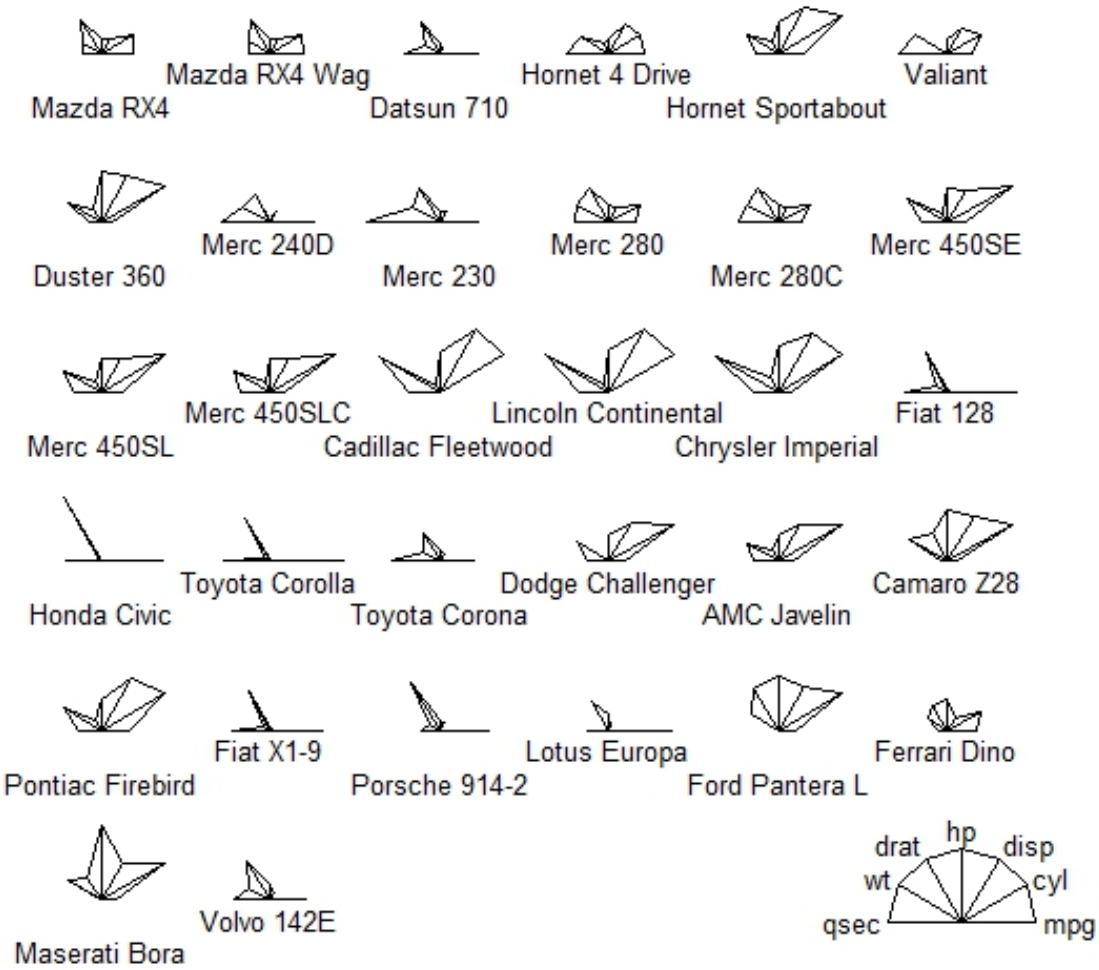
Functional Data Analysis



Stars

d) Stars-Plot

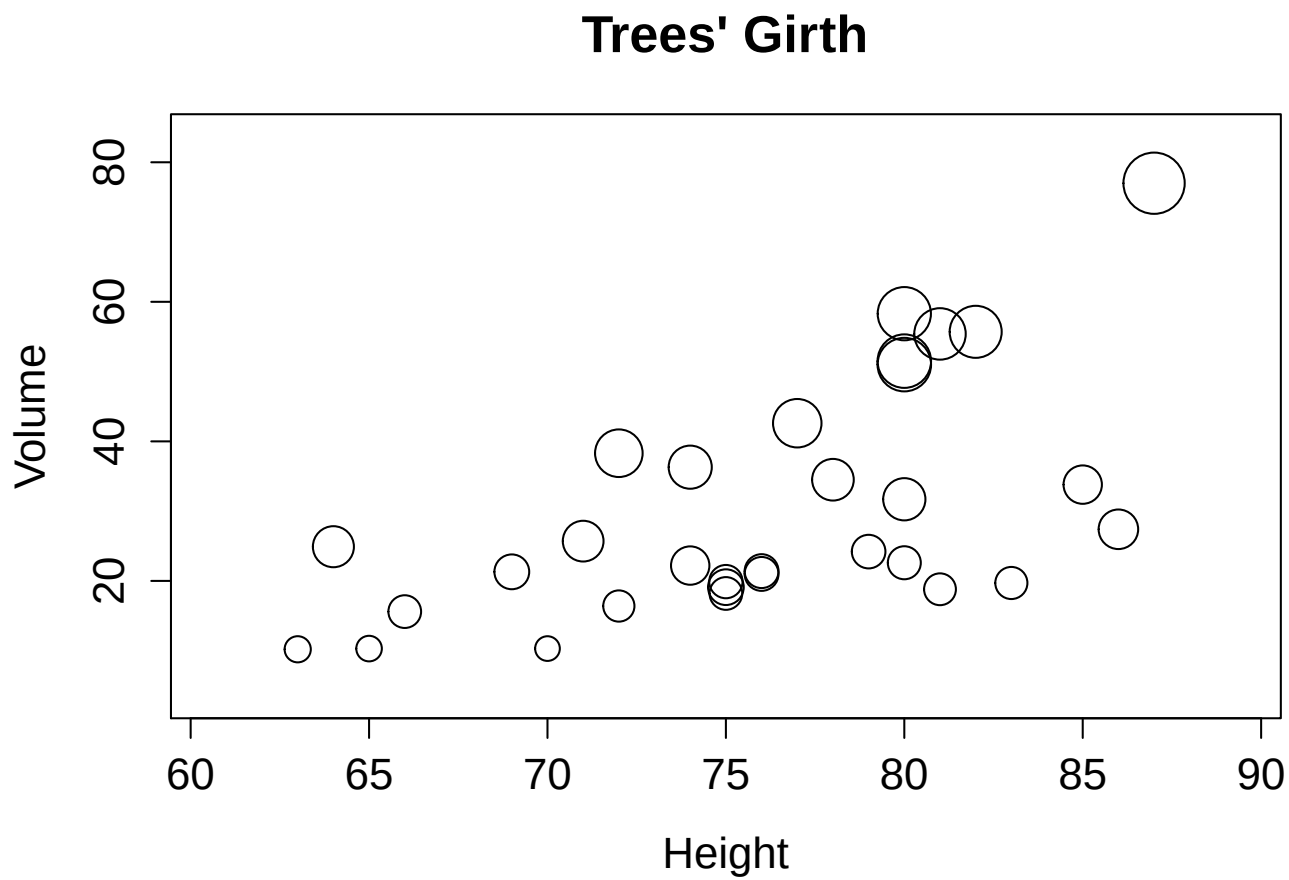
> data(mtcars); stars(mtcars[, 1:7], key.loc = c(14, 2), full = F)



Symbols

e) Symbols-Plot

```
> data(trees)
> with(trees, {symbols(Height, Volume, circles = Girth/24,
+ inches =FALSE, main = "Trees' Girth") })
```



Multivariate cdf

1.3 Joint and conditional distributions

1.3.1 Joint distribution

Definition

The multivariate *distribution function (cdf)* of the random vector $\underline{X} = (X_1, ..., X_p)^T$ is defined by

$$F_{\underline{X}}(\underline{x}) = F_{\underline{X}}(x_1, \ldots, x_p) := P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p);\ \underline{x} \in R^p.$$

If all the components of $\underline{X}$ are continuously distributed, then there exists a probability density function (pdf) $f_{\underline{X}}$ such that

a) $f_{\underline{X}}(\underline{x}) = \frac{\partial^p}{\partial x_1 \partial x_2 \cdots \partial x_p} F_{\underline{X}}(x_1, \ldots, x_p)$

b) $f_{\underline{X}}(\underline{x}) \geqslant 0, \forall \underline{x} \in \mathbb{R}^p$ (non-negativity')

c) $\int\limits_{\mathbb{R}^p} f_{\underline{X}}(\underline{x})\, d\underline{x} = \int\limits_{-\infty}^{\infty} \overset{(p)}{\cdots} \int\limits_{-\infty}^{\infty} f_{\underline{X}}(x_1, \ldots, x_p)\, dx_1 dx_2 \ldots dx_p = 1$
(normalization)

Multivariate cdf

If $\underline{X}$ has a discrete distribution, i.e. the realizations of $\underline{X}$ take values in a countable or finite set of points $\mathscr{X} = \{\underline{x}_j\}_{j \in J} \subset \mathbb{R}^p$, then it holds:

$$P(\underline{X} \in D) = \sum_{j:\underline{x}_j \in D} P(\underline{X} = \underline{x}_j), \quad \forall D \subseteq \mathscr{X}.$$

Representation of discrete probability distributions: by means of *tensors*

Properties of the cdf $F_{\underline{X}}$:

1) $F_{\underline{X}}(\underline{x}) \in [0,1], \ \forall \underline{x} \in \mathbb{R}^p$
2) $F_{\underline{X}}$ is monotonically increasing in each component $x_i, i = 1, ..., p$
3) $\lim_{x_j \to -\infty} F_{\underline{X}}(x_1, ..., x_{j-1}, x_j, x_{j+1}, ..., x_p) = 0, \ \forall j \in \{1, ..., p\}$

   briefly: $F_{\underline{X}}(x_1, ..., x_{j-1}, -\infty, x_{j+1}, ..., x_p) = 0$

   (it suffices that $x_j \to -\infty$ for at least one $j \in \{1, ..., p\}$)

Properties of a multivariate cdf

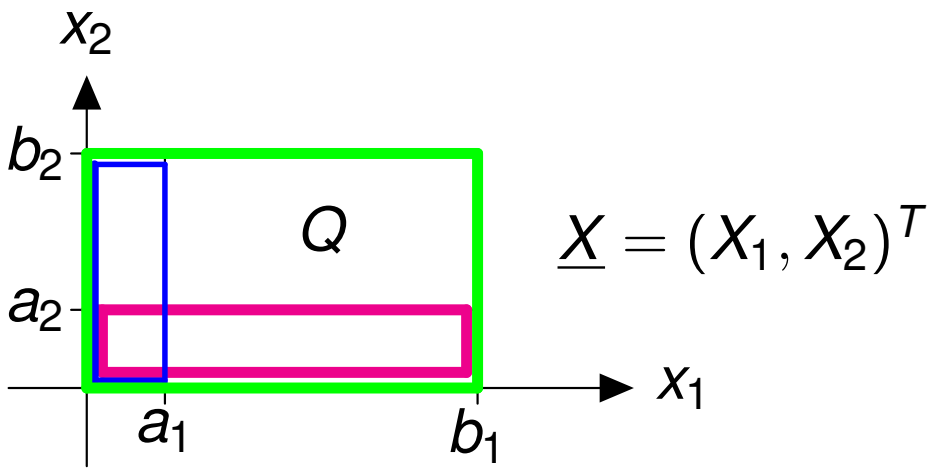
4) $\lim_{x_1 \to \infty, x_2 \to \infty, \ldots, x_p \to \infty} F_{\underline{X}}(x_1, \ldots, x_p) = 1,$

briefly: $F_{\underline{X}}(\infty, \infty, \ldots, \infty) = 1$ (all x_j taken to ∞)

5) $P_{\underline{X}}(Q) = \int_Q dF_{\underline{X}}(\underline{x}) \geqslant 0$ for all hyper-rectangles $Q \subset \mathbb{R}^p$;

$Q = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_p, b_p] \subset \mathbb{R}^p.$

Illustration of property 5) for the case p=2 and non-negative $\underline{X}$:



$$P_{\underline{X}}(Q) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \overset{!}{\geqslant} 0.$$

Properties of a multivariate cdf

It is not hard to find functions F satisfying 1) - 4) but violating 5) (thus, generating negative probabilities!).

Theorem

The conditions 1) - 5) *are necessary and sufficient for* $F = F_{\underline{X}}$ *to be a cdf. In the continuous case, condition* 5) *is equivalent to requiring*

$$\frac{\partial^p F_{\underline{X}}(x_1,\ldots,x_p)}{\partial x_1 \cdots \partial x_p} \geqslant 0.$$

1.3.2. Marginal distribution, conditional distribution

Partitioning of $\underline{X}$ into two sub-vectors: $\underline{X} = (\underbrace{X_1, \ldots, X_r}_{\underline{X}_1 \in \mathbb{R}^r}, \underbrace{X_{r+1}, \ldots, X_p}_{\underline{X}_2 \in \mathbb{R}^{p-r}})^T$

For continuous $\underline{X}$, the **marginal pdf** of $\underline{X}_1$ can be obtained from the joint density $f_{\underline{X}}(x_1, \ldots, x_p)$ as follows:

Marginal densities and distributions

$$f_{\underline{X}_1}(x_1, \ldots, x_r) = \int_{-\infty}^{\infty} \overset{(p-r)}{\cdots} \int_{-\infty}^{\infty} f_{\underline{X}}(x_1, \ldots, x_p)\, dx_{r+1} \ldots dx_p,$$

i.e. we are integrating over the "nuisance" variables (remaining variables which are not of interest).

The corresponding **marginal distribution** of $\underline{X}_1$ then reads

$$F_{\underline{X}_1}(x_1, ..., x_r) = F_{\underline{X}}(x_1, \ldots, x_r, \infty, \ldots, \infty)$$

$$= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_r} \underbrace{\left\{ \int_{-\infty}^{\infty} \overset{(p-r)}{\cdots} \int_{-\infty}^{\infty} f_{\underline{X}}(u_1, ..., u_p)\, du_{r+1} \cdots du_p \right\}}_{f_{\underline{X}_1}(u_1, \ldots, u_r)} du_1 \cdots du_r \,.$$

Conditional densities

In the continuous case, we can also define conditional distributions on the basis of conditional probability densities.
The **conditional pdf** of the sub-vector $\underline{X}_1$ for given realization of $\underline{X}_2 = \underline{x}_2 = (x_{r+1}, \ldots, x_p)^T$ can be obtained in the usual way via

$$f_{\underline{X}_1|\underline{X}_2=\underline{x}_2}(x_1, ..., x_r) = \frac{f_{\underline{X}}(x_1, ..., x_r, x_{r+1}, ..., x_p)}{f_{\underline{X}_2}(x_{r+1}, ..., x_p)}$$

i.e. "joint pdf/ marginal pdf".

Definition

$\underline{X}_1$ and $\underline{X}_2$ are said to be stochastically independent if and only if

$$F_{\underline{X}}(\underline{x}) = F_{\underline{X}_1}(\underline{x}_1) F_{\underline{X}_2}(\underline{x}_2) \quad \forall \underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix} \in \mathbb{R}^p.$$

In the continuous case this is equvivalent to demanding that

$$f_{\underline{X}}(\underline{x}_1, \underline{x}_2) = f_{\underline{X}_1}(\underline{x}_1)\ f_{\underline{X}_2}(\underline{x}_2).$$

Two-dimensional case

Example (Ex. 1.1)

Let $p = 2$ and $\underline{X} = (X_1, X_2)^T$ have the pdf

$$f_{\underline{X}}(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & , 0 \leqslant x_1, x_2 \leqslant 1 \\ 0 & , \textit{else}. \end{cases}$$

$f_{\underline{X}}(x_1, x_2)$ is a pdf, since it is non-negative and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\underline{X}}(x_1, x_2)dx_1 dx_2 = \frac{1}{2}\left[\frac{x_1^2}{2}\right]_0^1 + \frac{3}{2}\left[\frac{x_2^2}{2}\right]_0^1 = \frac{1}{4} + \frac{3}{4} = 1.$$

The marginal densities read

$$f_{X_1}(x_1) = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2\right) dx_2 = \frac{1}{2}x_1 + \frac{3}{4}\,, \quad 0 \leqslant x_1 \leqslant 1\,,$$

Example (Ex. 1.1 cont'd)

$$f_{X_2}(x_2) = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2\right) dx_1 = \frac{3}{2}x_2 + \frac{1}{4}, \quad 0 \leqslant x_2 \leqslant 1$$

and the conditional pdf's are given by

$$f_{X_2|X_1=x_1}(x_2) = \frac{\frac{1}{2}x_1 + \frac{3}{2}x_2}{\frac{1}{2}x_1 + \frac{3}{4}} \, I_{[0,1]\times[0,1]}(x_1, x_2) \, ,$$

$$f_{X_1|X_2=x_2}(x_1) = \frac{\frac{1}{2}x_1 + \frac{3}{2}x_2}{\frac{3}{2}x_2 + \frac{1}{4}} \, I_{[0,1]\times[0,1]}(x_1, x_2) \, .$$

Note: The conditional densities are nonlinear in x_1 and x_2, respectively, although the joint pdf has a simple linear structure.

Two-dimensional case

The following example shows the basic multivariate *reconstruction problem*: the marginalization process is not unique!

Example (Ex. 1.2)

Consider the following pdf's:

$$f_{\underline{X}}(x_1, x_2) = \begin{cases} 1 & , \text{if } \ 0 \leqslant x_1, x_2 \leqslant 1 \\ 0 & , \text{else} \end{cases}$$

(Density of the 2D- uniform distribution over the unit square) and

$$\tilde{f}_{\underline{X}}(x_1, x_2) = \begin{cases} 1 + \alpha(2x_1 - 1)(2x_2 - 1) & , \text{if } \ 0 \leqslant x_1, x_2 \leqslant 1 \\ 0 & , \text{else} \end{cases}$$

for some given $\alpha \in [-1, 1]$, $\alpha \neq 0$. The corresponding marginal pdf's read

Non-uniqueness of marginalization

Example (Ex. 1.2 cont'd)

$$f_{X_1}(x_1) = \begin{cases} 1 & \text{, if } 0 \leqslant x_1 \leqslant 1 \\ 0 & \text{, else} \end{cases} = 1 \cdot I_{[0,1]}(x_1), \quad X_1 \sim \mathrm{U}[0,1]$$

$$f_{X_2}(x_2) = \begin{cases} 1 & \text{, if } 0 \leqslant x_2 \leqslant 1 \\ 0 & \text{, else} \end{cases} = 1 \cdot I_{[0,1]}(x_2), \quad X_2 \sim \mathrm{U}[0,1]$$

and the marginal pdf's corresponding to $\tilde{f}_{\underline{X}}(x_1, x_2)$ are given by:

$$\tilde{f}_{X_1}(x_1) = \int_0^1 \{1 + \alpha(2x_1 - 1)(2x_2 - 1)\}\, dx_2$$

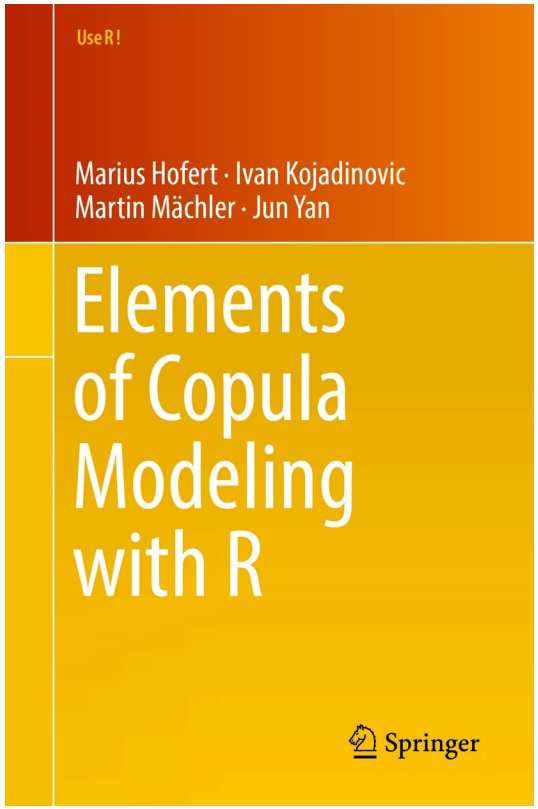
$$= 1 + \alpha(2x_1 - 1)\left[x_2^2 - x_2\right]_0^1 = 1;\ 0 \leqslant x_1 \leqslant 1$$

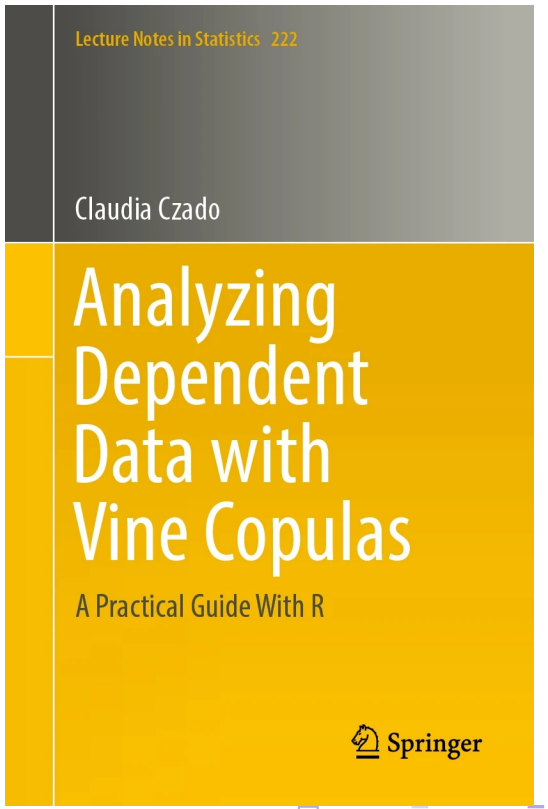
Analogously: $\tilde{f}_{X_2}(x_2) = 1,\ 0 \leqslant x_2 \leqslant 1$, i.e. we have identical marginals belonging to different joint distributions.

Non-uniqueness of marginalization

Reconstruction problem:

- It is impossible to reconstruct the higher-dimensional joint distributions (densities) from the lower-dimensional marginal distributions (densities). For the reconstruction, we need a connecting tool, the so-called **copula**.





Reconstruction

- Conversely, the projection of higher-dimensional distributions into lower-dimensional ones does not cause problems. This is accomplished through marginalization (integrating "out" the nuisance variables).
- The unique reconstruction of the joint distribution from the marginal distributions of sub-vectors is only possible if these are stochastically independent.
- However, we can reconstruct the joint distributions when we make use of marginal distributions in conjunction with conditional distributions.

As an example, we can represent three-dimensional densities ($p = 3$, $\underline{X} = (X_1, X_2, X_3)^T$) on the basis of unidimensional conditional and marginal densities as follows:

Projection theorem

$$\begin{aligned} f_{(X_1,X_2,X_3)}(x_1,x_2,x_3) &= f_{(X_1,X_2)|X_3=x_3}(x_1,x_2)\, f_{X_3}(x_3) \\ &= f_{X_1|X_2=x_2,X_3=x_3}(x_1)\, f_{X_2|X_3=x_3}(x_2)\, f_{X_3}(x_3) \end{aligned}$$

This is used e.g. in 3D image processing (reconstruction on the basis of two-dimensional projections). Further, even high-dimensional distributions can be generated from unidimensional conditional and marginal densities. This method has become known as **Gibbs-Sampling**; it is widely used in **Bayesian Statistics**.

The reconstruction problem becomes completely clear from the following theorem.

Theorem (Cramér-Wold)

The distribution of $\underline{X} = (X_1, ..., X_p)^T$ is completely determined by the set of all (one-dimensional) distributions of projections

$$\left\{ \underline{a}^T \underline{X} = a_1 X_1 + ... + a_p X_p : \underline{a} = (a_1, ..., a_p)^T \in \mathbb{R}^p \right\}.$$

1.4 Expectation vector, covariance matrix and correlation matrix

1.4.1 Unconditional expectation, covariance and correlation

We define the expectation and covariance of the random vector $\underline{X}$ element-wise.

$\underline{X} = (X_1, ..., X_p)^T$ has the **expectation vector**

$$\underline{\mu} = \mathbb{E}\underline{X} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_p \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} =$$

$$\int_{\mathbb{R}^p} (x_1, \ldots, x_p)^T f_{\underline{X}}(x_1, \ldots, x_p)\, dx_1 \cdots dx_p = \int_{\mathbb{R}^p} \underline{x}\, f_{\underline{X}}(\underline{x})\, d\underline{x}$$

with components $\mu_i = \mathbb{E}X_i = \int_{\mathbb{R}^p} x_i\, f_{\underline{X}}(x_1, ..., x_p)\, dx_1 \cdots dx_p \; i = 1, ..., p$.

Covariance matrix

The **covariance matrix** $\Sigma = \mathrm{Cov}(\underline{X})$ is defined element-wise as

$$\mathrm{Cov}(\underline{X}) = \left(\mathrm{Cov}(X_i, X_j)\right)_{i,j=1,\ldots,p}$$

$$= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathrm{Cov}(X_{p-1}, X_p) \\ \mathrm{Cov}(X_p, X_1) & \cdots & \mathrm{Cov}(X_p, X_{p-1}) & \mathrm{Var}(X_p) \end{pmatrix}$$

where
$\mathrm{Cov}(X_i, X_j) = \mathbb{E}\{(X_i - \mathbb{E}X_i)\left(X_j - \mathbb{E}X_j\right)\} = \mathbb{E}\left(X_i X_j\right) - (\mathbb{E}X_i)(\mathbb{E}X_j)$;
$i, j = 1, \ldots, p$.

Correlation matrix

Corollary 1: $\mathrm{Cov}(\underline{X}) = \mathbb{E}\{(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T\} = \mathbb{E}(\underline{X}\underline{X}^T) - \underline{\mu}\,\underline{\mu}^T$.

Correspondingly, we define the **correlation matrix**:

$$\mathrm{P} = \left(\mathrm{Corr}(X_i, X_j)\right)_{i,j=1,\ldots,p} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{p-1,p} \\ \rho_{p1} & \cdots & \rho_{p,p-1} & 1 \end{pmatrix}$$

where

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}};\ i, j = 1, \ldots, p\ .$$

Corollary 2: $\mathrm{P} = D^{-1/2}\Sigma D^{-1/2}$, with $D = \mathrm{diag}((Var(X_1), \ldots, Var(X_p))$.

1.4.2 Conditional expectation, covariance and correlation

The conditional expectations, variances and correlations are defined via the corresponding conditional distributions, i.e. for the partitioned random vector

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_r \\ \cdots \\ X_{r+1} \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}$$

we define the **conditional expectation** of $\underline{X}_2$ for given realization $\underline{x}_1$ of $\underline{X}_1$ as

Conditional expectation and covariance

$$\mathbb{E}(\underline{X}_2|\underline{X}_1 = \underline{x}_1) = \int_{\mathbb{R}^{p-r}} (x_{r+1}, ..., x_p)^T \; f_{\underline{X}_2|\underline{x}_1}(x_{r+1}, ..., x_p) \; dx_{r+1} \cdots dx_p$$

$$= \int_{\mathbb{R}^{p-r}} \underline{x}_2 \; f_{\underline{X}_2|\underline{x}_1}(\underline{x}_2) \; d\underline{x}_2$$

Observing that $\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - (\mathbb{E}X_i)(\mathbb{E}X_j)$, we define the **conditional covariance** matrix of $\underline{X}_2$ for given realization of $\underline{X}_1$ according to

$$\mathrm{Cov}(\underline{X}_2|\underline{X}_1 = \underline{x}_1)$$
$$= \mathbb{E}(\underline{X}_2\underline{X}_2^T|\underline{X}_1 = \underline{x}_1) - [\mathbb{E}(\underline{X}_2|\underline{X}_1 = \underline{x}_1)][\mathbb{E}(\underline{X}_2|\underline{X}_1 = \underline{x}_1)]^T.$$

Analogously, we define the

Conditional correlation matrix

conditional correlation matrix $Corr(\underline{X}_2|\underline{X}_1 = \underline{x}_1)$, element-wise by

$$\rho(X_i, X_j|\underline{X}_1 = \underline{x}_1) = \frac{\text{Cov}(X_i, X_j|\underline{X}_1 = \underline{x}_1)}{\sqrt{\text{Var}(X_i|\underline{X}_1 = \underline{x}_1)\text{Var}(X_j|\underline{X}_1 = \underline{x}_1)}}$$

$$i, j = r + 1, ..., p.$$

Example

Let $\underline{X} = (X_1, X_2, X_3)^T$ have pdf

$$f_{\underline{X}}(x_1, x_2, x_3) = \begin{cases} \frac{2}{3}(x_1 + x_2 + x_3) & \text{, for } 0 \leqslant x_1, x_2, x_3 \leqslant 1 \\ 0 & \text{, else} \end{cases}$$

Note that the density is symmetric w.r.t. x_1, x_2, x_3.
The marginal density of the first two components is obtained as

$$f_{(X_1,X_2)}(x_1, x_2) = \frac{2}{3}\left(x_1 + x_2 + \frac{1}{2}\right) I_{[0,1]^2}(x_1, x_2)$$

Conditional correlation matrix

Example (cont'd)

and the marginal pdf of X_1 reads

$$f_{X_1}(x_1) = \frac{2}{3}(x_1 + 1)\ I_{[0,1]}(x_1).$$

Further,

$$\mathbb{E}X_1 = \int_0^1 \frac{2}{3} x_1 (x_1 + 1)\, dx_1 = \frac{2}{3}\left[\frac{x_1^3}{3}\right]_0^1 + \frac{2}{3}\left[\frac{x_1^2}{2}\right]_0^1 = \frac{5}{9} = \mathbb{E}X_2 = \mathbb{E}X_3$$

$$\mathbb{E}X_1^2 = \int_0^1 \frac{2}{3} x_1^2 (x_1 + 1)\, dx_1 = \frac{2}{3}\left[\frac{x_1^4}{4}\right]_0^1 + \frac{2}{3}\left[\frac{x_1^3}{3}\right]_0^1 = \frac{7}{18} = \mathbb{E}X_2^2 = \mathbb{E}X_3^2$$

Conditional correlation matrix

Example (cont'd)

$$\mathbb{E}(X_1X_2) = \int_0^1 \int_0^1 x_1x_2\frac{2}{3}\left(x_1 + x_2 + \frac{1}{2}\right) dx_1\, dx_2 = \frac{11}{36}$$
$$= \mathbb{E}(X_1X_3) = \mathbb{E}(X_2X_3).$$

Thus, $\underline{\mu} = \mathbb{E}\underline{X} = \frac{5}{9}(1,1,1)^T$, and

$$\Sigma = \mathrm{Cov}(X_1, X_2, X_3) = \mathbb{E}(\underline{X}\underline{X}^T) - \underline{\mu}\underline{\mu}^T = \mathbb{E}\begin{bmatrix} X_1^2 & X_1X_2 & X_1X_3 \\ X_2X_1 & X_2^2 & X_2X_3 \\ X_3X_1 & X_3X_2 & X_3^2 \end{bmatrix} - \underline{\mu}\underline{\mu}^T$$

$$\Longrightarrow \Sigma = \frac{1}{36}\begin{bmatrix} 14 & 11 & 11 \\ 11 & 14 & 11 \\ 11 & 11 & 14 \end{bmatrix} - \frac{25}{81}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \frac{1}{324}\begin{bmatrix} 26 & -1 & -1 \\ -1 & 26 & -1 \\ -1 & -1 & 26 \end{bmatrix}$$

Conditional correlation matrix

Example (cont'd)

Thus, we have

$$\rho_{12} = \mathrm{Corr}(X_1, X_2) = \frac{-1/324}{\sqrt{(26/324)^2}} = \mathrm{Corr}(X_1, X_3) = \mathrm{Corr}(X_2, X_3) = -\frac{1}{26}$$
$$\approx -0.0385.$$

How to determine the conditional covariance matrix

$$\mathrm{Cov}\left(\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} | X_3 = x_3\right) = \begin{bmatrix} \mathrm{Var}(X_1|x_3) & \mathrm{Cov}(X_1, X_2|x_3) \\ \mathrm{Cov}(X_1, X_2|x_3) & \mathrm{Var}(X_2|x_3) \end{bmatrix}?$$

We need the conditional distributions of $X_1|x_3$ and $(X_1, X_2)|x_3$.

$$f_{(X_1,X_2)|x_3}(x_1, x_2) = \frac{x_1 + x_2 + x_3}{1 + x_3} \, I_{[0,1]^2}(x_1, x_2)$$

$$f_{X_1|x_3}(x_1) = \frac{x_1 + x_3 + 1/2}{1 + x_3} \, I_{[0,1]}(x_1).$$

Conditional correlation matrix

Example (cont'd)

From these pdf's we easily obtain

$$\mathbb{E}(X_1|x_3) = \int_0^1 x_1 \frac{x_1 + x_3 + 1/2}{1 + x_3}\, dx_1 = \frac{1}{1 + x_3}\left[\frac{x_1^3}{3} + x_3\frac{x_1^2}{2} + \frac{x_1^2}{4}\right]_{x_1=0}^{x_1=1}$$

$$= \tfrac{1}{12} \cdot \tfrac{6x_3+7}{1+x_3} = \mathbb{E}(X_2|x_3).$$

$$\mathbb{E}(X_1X_2|x_3) = \int_0^1\int_0^1 x_1x_2 \frac{x_1 + x_2 + x_3}{1 + x_3}\, dx_1\, dx_2 = \frac{1}{12} \cdot \frac{3x_3 + 4}{1 + x_3}\,.$$

This yields

$$\mathrm{Cov}\left(\begin{pmatrix}X_1\\X_2\end{pmatrix}|X_3 = x_3\right) = \frac{1}{g(x_3)}\begin{bmatrix}12x_3^2 + 24x_3 + 11 & -1\\ -1 & 12x_3^2 + 24x_3 + 11\end{bmatrix}$$

Conditional correlation matrix

Example (cont'd)

where $g(x_3) = 144(1 + x_3)^2$, observing that

$\mathrm{Cov}(X_1, X_2|x_3) = \mathbb{E}(X_1X_2|x_3) - \mathbb{E}(X_1|x_3)\mathbb{E}(X_2|x_3) = -1/g(x_3)$ and

$$\begin{aligned}\mathrm{Var}(X_1|x_3) = \mathrm{Var}(X_2|x_3) &= \mathbb{E}(X_1^2|x_3) - [\mathbb{E}(X_1|x_3)]^2 \\ &= (12x_3^2 + 24x_3 + 11)/g(x_3)\end{aligned}$$

$$\Rightarrow \mathrm{Corr}(X_1, X_2|X_3 = x_3) = -\frac{1}{12x_3^2+24x_3+11} \in (-0.0909, -0.0213)$$

for all $x_3 \in (0, 1)$.

Depending on the size of x_3, the conditional correlation can be less or greater than the unconditional correlation $\rho_{12} = -0.0385$.

1.4.3 Properties of conditional expectations and covariance matrices

In the previous section, we had considered $\mathbb{E}(\underline{X}_2|\underline{X}_1 = \underline{x}_1)$ and $\mathrm{Cov}(\underline{X}_2|\underline{X}_1 = \underline{x}_1)$ as functions of the given realization $\underline{x}_1 \in \mathbb{R}^r$ of $\underline{X}_1$. We now consider the randomized versions, leading us to the random vector

$$\mathbb{E}(\underline{X}_2|\underline{X}_1) : \Omega \longrightarrow \mathbb{R}^r$$

and the random matrix

$$\mathrm{Cov}(\underline{X}_2|\underline{X}_1) : \Omega \longrightarrow \mathbb{R}^{r\times r},$$

respectively.

Then the following relationships between conditional and total expectation vectors and covariance matrices, resp., can be established:

Total expectation and total covariance

a) **Law of total (iterated) expectation:**

$$\mathbb{E}\underline{X}_2 = \mathbb{E}_{\underline{X}_1}\left\{\mathbb{E}(\underline{X}_2|\underline{X}_1)\right\}$$

b) **Law of total covariance:**

$$\mathrm{Cov}(\underline{X}_2) = \mathbb{E}_{\underline{X}_1}\left\{\mathrm{Cov}(\underline{X}_2|\underline{X}_1)\right\} + \mathrm{Cov}_{\underline{X}_1}(\mathbb{E}(\underline{X}_2|\underline{X}_1))$$

Prove a) and b)!

Example

Let $p = 2, r = 1$ and $\underline{X} = (X_1, X_2)^T$ with pdf

$$f_{\underline{X}}(x_1, x_2) = \begin{cases} 2e^{-x_2/x_1} & \text{for } 0 < x_1 < 1,\ x_2 > 0 \\ 0 & \text{else.} \end{cases}$$

It is easily seen that

$f_{X_1}(x_1) = 2x_1 I_{[0,1]}(x_1)$, $\mathbb{E}X_1 = \frac{2}{3}$, $\mathrm{Var}(X_1) = \frac{1}{18}$ and

Total expectation and total covariance

Example (cont'd)

$f_{X_2|x_1}(x_2) = \frac{1}{x_1}\ e^{-x_2/x_1}\ I_{(0,\infty)}(x_2)$,

$\mathbb{E}(X_2|X_1) = X_1,\ \mathrm{Var}(X_2|X_1) = X_1^2$.

From these results we get (without needing to compute f_{X_2} explicitly!):

$$\mathbb{E}(X_2) \underset{\text{a)}}{=} \mathbb{E}_{X_1}\left\{\mathbb{E}(X_2|X_1)\right\} = \mathbb{E}_{X_1}(X_1) = \frac{2}{3}\ ,$$

$$\mathrm{Var}(X_2) \underset{\text{b)}}{=} \mathbb{E}_{X_1}\left\{\mathrm{Var}(X_2|X_1)\right\} + \mathrm{Var}\left\{\mathbb{E}(X_2|X_1)\right\} = \mathbb{E}(X_1^2) + \mathrm{Var}(X_1)$$

$$= \mathrm{Var}(X_1) + (\mathbb{E}(X_1))^2 + \mathrm{Var}(X_1)) = 2\mathrm{Var}(X_1) + (\mathbb{E}((X_1))^2 = \frac{5}{9}.$$

Total expectation and total covariance

Connection to regression models:

Target (response) variable $Y = Y(\underline{x}) = Y(x_1, ..., x_p)$

LM: $Y(\underline{x}) = \underline{f}(\underline{x})^T \underline{\beta} + \varepsilon$ where $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$

The regression setup then models

$$\mathbb{E}\, Y(\underline{x}) = \underline{f}(\underline{x})^T \underline{\beta} \;=\; \mathbb{E}(Y|\underline{X} = \underline{x}),$$

i.e. the conditional expectation (mean response) of Y given $\underline{x}$.

More general, we call

$$\underline{g}^*(\underline{X}_1) = \mathbb{E}(\underline{X}_2|\underline{X}_1)$$

the (multivariate) regression function of $\underline{X}_2$ onto $\underline{X}_1$, and

$$\underline{\hat{\varepsilon}} = \underline{X}_2 - \mathbb{E}(\underline{X}_2|\underline{X}_1)$$

the approximation error (residual error) which results from the approximation of $\underline{X}_2$ through $\underline{g}^*$.

MSE-approximation

Theorem

Let $\underline{X}_1 \in R^r$, $\underline{X}_2 \in \mathbb{R}^{p-r}$, $p > r > 0$, *and* $\hat{\underline{\varepsilon}} = \underline{X}_2 - \mathbb{E}(\underline{X}_2|\underline{X}_1)$.

Then it holds:

a) $\mathbb{E}(\hat{\underline{\varepsilon}}) = 0$

b) $\underline{g}^*(\underline{X}_1) = \mathbb{E}(\underline{X}_2|\underline{X}_1)$ *is the best MSE- approximation of* $\underline{X}_2$ *among all measurable functions* $\underline{g}(\underline{X}_1)$, $g : \mathbb{R}^r \longrightarrow \mathbb{R}^{p-r}$*:*

$$\mathbb{E}\,||\underline{X}_2 - \underline{g}^*(\underline{X}_1)||^2 = \min_{\underline{g}} \mathbb{E}\,||\underline{X}_2 - \underline{g}(\underline{X}_1)||^2.$$

Consequence: Minimum-MSE estimation and prediction is accomplished through conditional expectation!

Remark: If $\underline{X} = (\underline{X}_1, \underline{X}_2)^T$ follows a p-dimensional normal distribution (see **Ch. 2**) then $\mathbb{E}(\underline{X}_2|\underline{X}_1)$ linear in $\underline{X}_1$.

This is used e.g. in spatial data analysis: **Kriging-predictor** (BLUP).

Sample mean

1.5 Multivariate samples

Samples $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$; $i = 1, ..., n$; are collected in the data matrix

$$\underset{(n,p)}{X} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

The so-called **sample mean** vector is then given by:

$$\underline{\hat{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_p \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \end{pmatrix}.$$

the components $\hat{\mu}_1, \ldots, \hat{\mu}_p$ are just the (column-wise) sample averages

Sample covariance matrix

of the realizations of the corresponding variables.
$\underline{\hat{\mu}}$ acts as (moment-)estimate of $\underline{\mu} = \mathbb{E}\underline{X}$ (in case of normally distributed observations it coincides with the maximum likelihood estimate).

In order to estimate the covariance matrix $\Sigma := \mathrm{Cov}(\underline{X})$, we use the so-called **sample covariance** matrix, which acts as multivariate generalization of the usual empirical variances/covariances:

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \underline{\hat{\mu}})(\underline{x}_i - \underline{\hat{\mu}})^T.$$

In the main diagonal of $\widehat{\Sigma} = (\hat{\sigma}_{kl})_{k,l=1,\ldots,p}$ we have the empirical variances

$$\hat{\sigma}_{kk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \hat{\mu}_k)^2; \quad k = 1, \ldots, p$$

of the single components X_k; $k = 1, \ldots, p$; of $\underline{X}$ and

Sample covariance matrix

in the off-diagonals we have the empirical covariances

$$\hat{\sigma}_{kl} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \hat{\mu}_k)(x_{il} - \hat{\mu}_l); k, l = 1, \ldots, p; k \neq l.$$

This means: $\hat{\sigma}_{kk}$ is an estimate of $\mathrm{Var}(X_k)$ and $\hat{\sigma}_{kl}$ an estimate of $\mathrm{Cov}(X_k, X_l)$:

$$\hat{\sigma}_{kk} = \widehat{\mathrm{Var}(X_k)}; \hat{\sigma}_{kl} = \widehat{\mathrm{Cov}(X_k, X_l)}; \quad k, l = 1, ..., p.$$

By means of the so-called **mean-centered data matrix**

$$\underset{(n,p)}{X_c} = \begin{bmatrix} x_{11} - \hat{\mu}_1 & x_{12} - \hat{\mu}_2 & \cdots & x_{1p} - \hat{\mu}_p \\ x_{21} - \hat{\mu}_1 & x_{22} - \hat{\mu}_2 & \cdots & x_{2p} - \hat{\mu}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \hat{\mu}_1 & x_{n2} - \hat{\mu}_2 & \cdots & x_{np} - \hat{\mu}_p \end{bmatrix}$$

Sample covariance matrix

we may then write

$$\widehat{\Sigma} = \frac{1}{n-1} X_c^T X_c = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \underline{\hat{\mu}})(\underline{x}_i - \underline{\hat{\mu}})^T.$$

From $\widehat{\Sigma}$ we obtain the **sample correlation matrix** as

$$\widehat{\mathrm{P}} = (\hat{\rho}_{k,l})_{k,l=1,\ldots,p} \text{ where } \hat{\rho}_{k,l} = \widehat{\mathrm{Corr}}(X_k, X_l) = \frac{\hat{\sigma}_{kl}}{\sqrt{\hat{\sigma}_{kk}\hat{\sigma}_{ll}}}.$$

This matrix, in turn, may be easily obtained from the **standardized data matrix**

$$X_{\mathrm{st}} = \begin{bmatrix} \frac{x_{11}-\hat{\mu}_1}{\sqrt{\hat{\sigma}_{11}}} & \frac{x_{12}-\hat{\mu}_2}{\sqrt{\hat{\sigma}_{22}}} & \cdots & \frac{x_{1p}-\hat{\mu}_p}{\sqrt{\hat{\sigma}_{pp}}} \\ \frac{x_{21}-\hat{\mu}_1}{\sqrt{\hat{\sigma}_{11}}} & \frac{x_{22}-\hat{\mu}_2}{\sqrt{\hat{\sigma}_{22}}} & \cdots & \frac{x_{2p}-\hat{\mu}_p}{\sqrt{\hat{\sigma}_{pp}}} \\ \vdots & \vdots & & \vdots \\ \frac{x_{n1}-\hat{\mu}_1}{\sqrt{\hat{\sigma}_{11}}} & \frac{x_{n2}-\hat{\mu}_2}{\sqrt{\hat{\sigma}_{22}}} & \cdots & \frac{x_{np}-\hat{\mu}_p}{\sqrt{\hat{\sigma}_{pp}}} \end{bmatrix}$$

Sample correlation matrix

Then it holds:

$$\widehat{\mathrm{P}} = \frac{1}{n-1} X_{\mathrm{st}}^T X_{\mathrm{st}} \ .$$

Remark:

1) By means of the diagonal matrix of the empirical variances $D = \mathrm{diag}(\hat{\sigma}_{11}, \hat{\sigma}_{22}, ..., \hat{\sigma}_{pp})$ we may write $\widehat{\mathrm{P}}$ equivalently as:

$$\widehat{\mathrm{P}} = D^{-1/2} \widehat{\Sigma} D^{-1/2}, \text{ where } D^{-1/2} = \mathrm{diag}\left(\frac{1}{\sqrt{\hat{\sigma}_{11}}}, ..., \frac{1}{\sqrt{\hat{\sigma}_{pp}}} \right).$$

2) The above representations can be easily verified observing that $X_c = X - \underline{1}_p \underline{\hat{\mu}}^T$, where $\underline{1}_p = (1, ..., 1)^T$ denotes the vector of p one's.

Implementation in R:

```
> apply(mtcars,2,mean)
```
Sample mean vector $\underline{\hat{\mu}}$

```
> cov(mtcars)
```
Sample covariance matrix $\widehat{\Sigma}$

```
> cor(mtcars)
```
Sample correlation matrix $\widehat{\mathrm{P}}$

Variability measures

Let $\underline{X} = (X_1, ..., X_p)^T$ be a $p-$dimensional random vector with $\mathbb{E}\underline{X} = \underline{\mu}$, $\mathrm{Cov}(\underline{X}) = \Sigma$.
As a measure of the deviation of $\underline{X}$ from the center $\underline{\mu}$ we make use of the **Mahalanobis distance**

$$d_M^2(\underline{X}, \underline{\mu}) = (\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu}).$$

For a given realization $\underline{x}$ of $\underline{X}$ we compute its estimated Mahalanobis distance using the estimates $\hat{\underline{\mu}}$ and $\widehat{\Sigma}$ of $\underline{\mu}$ and Σ, respectively.

Special cases:

a) $\Sigma = \mathrm{diag}(\sigma_{11}, ..., \sigma_{pp})$; i.e. $\rho(X_i, X_j) = 0$ if $i \neq j$, which leads to

$$\widehat{d_M^2(\underline{x}, \underline{\mu})} = \sum_{i=1}^{p} \hat{\sigma}_{ii}^{-1} (x_i - \hat{\mu}_i)^2 = \sum_{i=1}^{p} \left(\frac{x_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}_{ii}}} \right)^2,$$

the Euclidean distance, weighted by the inverse variances (precisions).

Variability measures

b) $\Sigma = \sigma^2 I_p$, i.e. uncorrelated observations of equal variance σ^2, which leads to

$$\widehat{d_M^2(\underline{x}, \underline{\mu})} = \hat{\sigma}^{-2} \sum_{i=1}^{p} (x_i - \hat{\mu}_i)^2 = \hat{\sigma}^{-2} ||\underline{x} - \underline{\hat{\mu}}||^2$$

and is thus proportional to the Euclidean distance.

Definition

a) $V_g(\underline{X}) = \det(\mathrm{Cov}(\underline{X})) = \det(\Sigma)$

is called the **generalized variance** of $\underline{X}$.

b) $V_t(\underline{X}) = \mathrm{tr}(\mathrm{Cov}(\underline{X})) = \mathrm{tr}(\Sigma)$

is called the **total variance** of $\underline{X}$.

Generalized and total variance

Interpretation of generalized variance:

$$\mathscr{E} = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \leq c\}$$

defines an ellipsoid in $\mathbb{R}^p$ (often called the scatter ellipsoid) with volume

$$\text{Vol}(\mathscr{E}) \sim \sqrt{\det(\Sigma)}.$$

Interpretation of total variance:

$$V_t(\underline{X}) = \text{tr}(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \text{Var}(X_i),$$

represents the sum of the variances of all components of $\underline{X}$.

2. Multivariate normal distribution

2.1 Transformation theorem, characteristic function

We assume $\underline{X}$ to be continuous with pdf $f_{\underline{X}}(x_1, \ldots, x_p)$.

Theorem (Transformation theorem for pdf's)

Let $A \subseteq \mathbb{R}^p$ be an open set with $\int_A f_{\underline{X}}(\underline{x})\, d\underline{x} = 1$ and $\underline{g} : A \longrightarrow \mathbb{R}^p$ a bijective mapping such that both $\underline{g}$ and $\underline{g}^{-1}$ are continuously differentiable. Then the transformed vector $\underline{Y} = \underline{g}(\underline{X}) : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ has the pdf

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{g}^{-1}(\underline{y})) \cdot |\det(\mathcal{J}(\underline{y}))|,$$

where

$$\mathcal{J}(\underline{y}) = \left[\frac{\partial}{\partial y_j}\, g_i^{-1}(\underline{y})\right]_{i,j=1,\ldots,p}$$

*represents the so-called **Jacobi matrix** of the transformation g.*

Jacobi transformation

Briefly: $\underline{X} = \underline{g}^{-1}(\underline{Y}) \Rightarrow \mathcal{J}(\underline{y}) = \frac{\partial\, \underline{x}}{\partial\, \underline{y}} = \left[\frac{\partial\, x_i}{\partial\, y_j}\right]_{i,j=1,\ldots,p}$.

Special case: Linear transformation

$$\underline{Y} = A\underline{X} + \underline{b};\ A \in \mathbb{R}^{p\times p},\ \underline{b} \in \mathbb{R}^p$$

(of particular importance for rescaling, standardization,...).

If A is regular, i.e. $\det(A) \neq 0$, then we have

$$\underline{X} = \underline{g}^{-1}(\underline{Y}) = A^{-1}(\underline{Y} - \underline{b}).$$

This implies

$$\begin{aligned}
\mathcal{J}(\underline{y}) &= \left[\frac{\partial x_i}{\partial y_j}\right]_{i,j=1,\ldots,p} = A^{-1} \\
\Longrightarrow f_{\underline{Y}}(\underline{y}) &= f_{\underline{X}}(A^{-1}(\underline{y} - \underline{b})) \cdot |\det(A^{-1})| \\
&= \frac{1}{|\det(A)|}\, f_{\underline{X}}(A^{-1}(\underline{y} - \underline{b}))
\end{aligned}$$

Linear transformation

Example

Let $p = 2$ and $\underline{X} = (X_1, X_2)^T \sim U([0,1] \times [0,1])$,
Q: Are the sum and the difference of the components stochastically independent?

$$\underline{Y} = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix} \Longrightarrow \underline{Y} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{A} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \underline{b}, \quad \underline{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\Longrightarrow \underline{X} = A^{-1}\underline{Y} = -\frac{1}{2}\begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} Y_1 + Y_2 \\ Y_1 - Y_2 \end{pmatrix}$$

$$\mathcal{J}(\underline{y}) = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = A^{-1}, \ |\det(\mathcal{J}(\underline{y}))| = |(1/2)^2(-2)| = 1/2$$

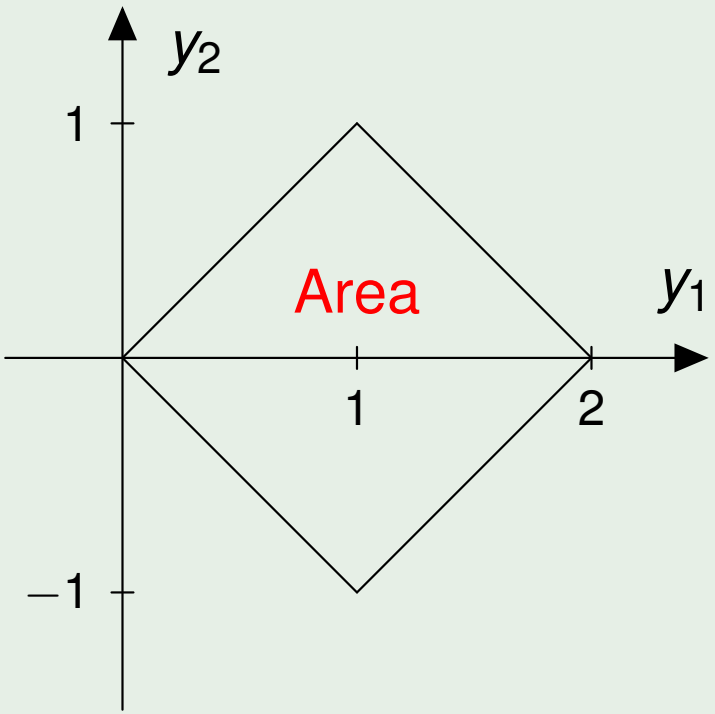
$$f_{\underline{Y}}(\underline{y}) = \frac{1}{2} f_{\underline{X}}(A^{-1}\underline{y}) = \frac{1}{2} f_{\underline{X}}\left(\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2)\right)$$

Linear transformation

Example (cont'd)

$$\Rightarrow f_{\underline{Y}}(\underline{y}) = \begin{cases} \frac{1}{2} & \text{, if } 0 \leqslant y_1 + y_2 \leqslant 2,\ 0 \leqslant y_1 - y_2 \leqslant 2 \\ 0 & \text{, else.} \end{cases}$$

Illustration:



$$|Area| = 2$$

Characteristic function

Characteristic function of random vectors

Recall: The characteristic function of a random variable X is defined as

$$\varphi_X(t) = \mathbb{E}e^{itX} = \int_{\mathbb{R}^1} e^{itx} f_X(x)dx.$$

It is well-known that (φ_X, f_X) form a Fourier-pair.

The extension of this notion to random vectors is straightforward.

Definition

Let $\underline{X} = (X_1, \ldots, X_p)^T$ be a p-dimensional random vector with pdf $f_{\underline{X}}$. Then

$$\varphi_{\underline{X}}(\underline{t}) := \mathbb{E}\left[\exp\left(i\underline{t}^T\underline{X}\right)\right] = \int_{\mathbb{R}^p} \exp\left(i\underline{t}^T\underline{x}\right) f_{\underline{X}}(\underline{x})d\underline{x}$$

is called the **characteristic function** of $\underline{X}$; $\underline{t} \in \mathbb{R}^p$.

chf of linear transformation

It follows immediately from the Cramér-Wold theorem that $\varphi_{\underline{X}}(\cdot)$ characterizes the distribution function $F_{\underline{X}}$ uniquely.
Moreover, $(f_{\underline{X}}, \varphi_{\underline{X}})$ form a Fourier-Paar.

Corollary 2.1: Let $\underline{Y} = \underset{(q,p)}{A}\,\underline{X} + \underline{b}$ be a linear transformation with rank(A) $= q \leqslant p$. Then it holds:

$$\varphi_{\underline{Y}}(\underline{t}) = \exp\left(i\underline{t}^T\underline{b}\right)\varphi_{\underline{X}}(A^T\underline{t}); \quad \underline{t} \in \mathbb{R}^q.$$

Proof:

$$\begin{aligned}
\varphi_{\underline{Y}}(\underline{t}) &= \mathbb{E}\left\{\exp\left(i\underline{t}^T\underline{Y}\right)\right\} = \mathbb{E}\left\{\exp\left[i\underline{t}^T(A\underline{X}+\underline{b})\right]\right\} \\
&= \mathbb{E}\left\{\exp\left(i\underline{t}^T\underline{b}\right)\exp\left(i\underline{t}^TA\underline{X}\right)\right\} \\
&= \exp\left(i\underline{t}^T\underline{b}\right)\mathbb{E}\left\{\exp\left(i[A^T\underline{t}]^T\underline{X}\right)\right\} \\
&= \exp\left(i\underline{t}^T\underline{b}\right)\varphi_{\underline{X}}(A^T\underline{t})
\end{aligned}$$

2.2 Multivariate Gaussian distribution

Normality (Gaussianity) is a distributional assumption which many statistical methods are based upon. Quite often, the normal distribution comes out as a limiting distribution of (normalized) sums and averages, respectively.

Definition

The random vector $\underline{X} = (X_1, \ldots, X_p)^T$ is said to follow a p-dimensional normal distribution with parameters $\underline{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathrm{PD}(p)$, briefly: $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, if it has the pdf

$$f_{\underline{X}}(\underline{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\Big[-\frac{1}{2} \underbrace{(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})}_{d_M^2(\underline{x}, \underline{\mu})} \Big].$$

Let us have a brief look at the special case of a bivariate ($p = 2$) normal distribution.

Bivariate normal

Example (Bivariate normal pdf)

For the normal random vector $\underline{X} = (X_1, X_2)^T \sim N_2\left(\underline{\mu}, \Sigma\right)$

with parameters

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix};$$

where $\sigma_i^2 = Var(X_i); i = 1, 2$; and $\rho = Cor(X_1, X_2)$, we have

$$\det(\Sigma) = \sigma_1^2\sigma_2^2(1-\rho^2) \quad , \quad \Sigma^{-1} = \frac{1}{\det(\Sigma)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

and thus obtain the pdf in a more explicit form as

$$f_{\underline{X}}(t_1, t_2) = \frac{1}{2\pi\sigma_1\sigma_2\tau} \exp\left(-\frac{1}{2\tau^2}\left[t_1^2 - 2\rho t_1 t_2 + t_2^2\right]\right)$$

where $\tau^2 = 1 - \rho^2$, and $t_i = (x_i - \mu_i)/\sigma_i$ $(i = 1, 2)$.

Bivariate normal distribution

Example (cont'd)

From the analytical form of the pdf we easily recognize that it has elliptical contours. The marginal distributions of $\underline{X}$ are univariate normal distributions:

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-(x_i - \mu_i)^2/2\sigma_i^2) \quad i = 1, 2$$

i.e. $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. For the conditional density $f_{X_2|X_1=x_1}$ it holds:

$$f_{X_2|X_1=x_1}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_{2|1}} \exp(-(x_2 - \mu_{2|1})^2/2\sigma_{2|1}^2), \text{ where}$$

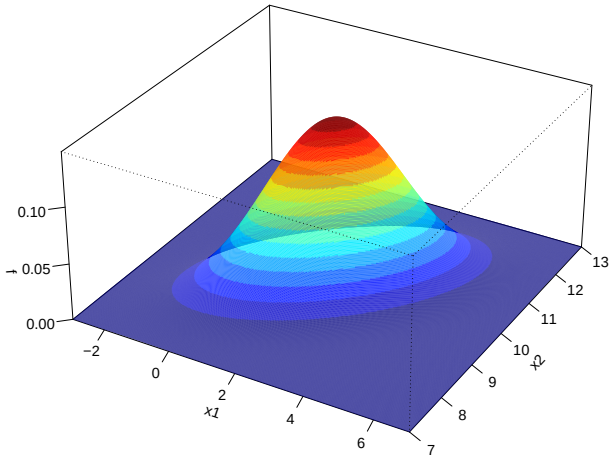
$$\mu_{2|1} = \mathbb{E}(X_2|X_1 = x_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1)$$

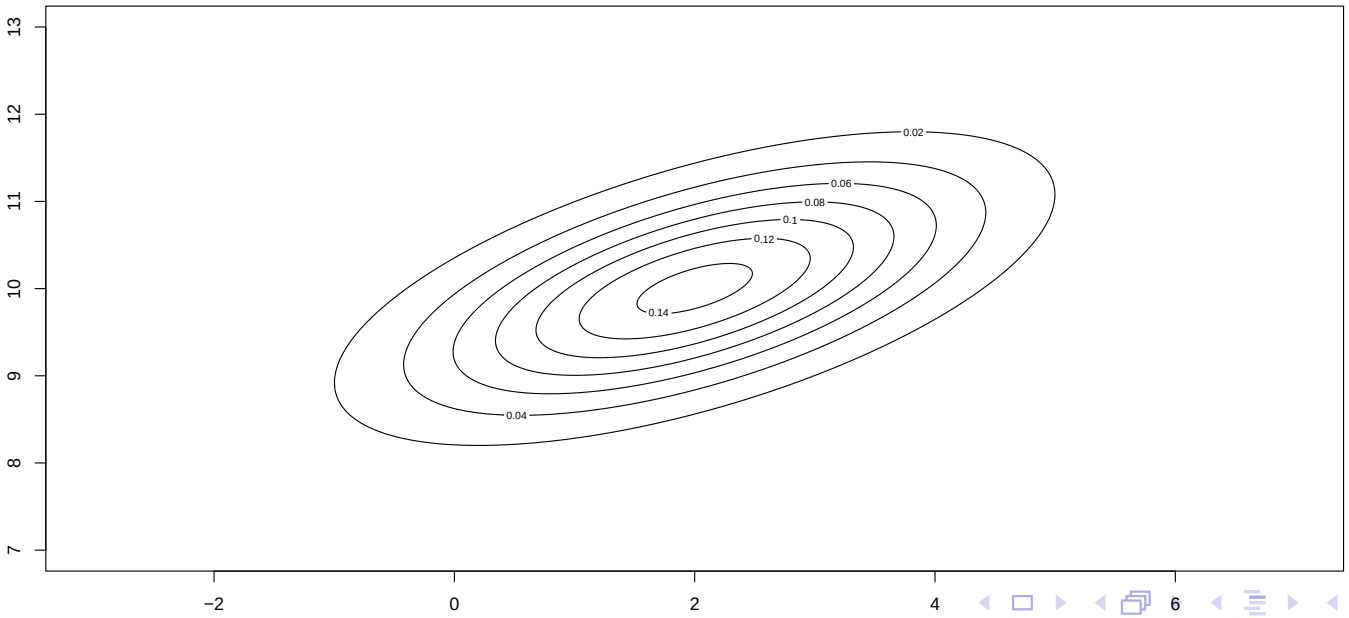
$$\sigma_{2|1}^2 = \mathrm{Var}(X_2|X_1 = x_1) = \sigma_2^2(1 - \rho^2).$$

Visualization in R

```
> # Input
> m1 = 2 #mean of X1
> m2 = 10 #mean of X2
> s1 = 1.5 #standard deviation of X1
> s2 = 0.9 #standard deviation of X2
> r = 0.6 #correlation between X1 and X2
> # 3D and contour plot arrangements
> x1 = seq(m1-5,m1+5, length= 500)
> x2 = seq(m2-3,m2+3, length= 500)
> z= function(x1,x2){z=exp(-(((x1-m1)*(x1-m1)/(s1*
s1))+((x2-m2)*(x2-m2)/(s2*s2))-2*r*(x1-m1)*(x2-m2)/
(s1*s2))/(2*(1-r*r)))/(2*pi*s1*s2*sqrt(1-r*r))}
> f = outer(x1, x2, z)
> persp3D(x1, x2, f, theta=30, phi=30, expand=0.5)
> contour(x=x1, y=x2, z=f)
```

Bivariate normal distribution in R





Standardization

Remark: The contour surfaces $f_{\underline{X}}(\underline{x}) = c > 0$ define ellipsoids in $\mathbb{R}^p$, $p \geq 2$, with center $\underline{\mu}$ and shape determined by the eigenvalues of Σ:

$$\ln f_{\underline{X}}(\underline{x}) = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(\det\Sigma) - \frac{1}{2}d_M^2(\underline{x},\underline{\mu}) = \ln c$$

$$\Longleftrightarrow (\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu}) = b$$

where $b = -p\ln(2\pi) - \ln(\det\Sigma) - 2\ln c$.

Standardization of $\underline{X}$: Transforming $\underline{X}$ into

$$\underline{Y} = \Sigma^{-1/2}\left(\underline{X}-\underline{\mu}\right),$$

where $\Sigma^{1/2}$ denotes the (symmetric) square root of Σ , i.e.

$$\left(\Sigma^{1/2}\right)\left(\Sigma^{1/2}\right)^T = \Sigma \text{ and } \Sigma^{-1/2} = \left(\Sigma^{1/2}\right)^{-1}.$$

The symmetric square root can be obtained from the spectral decomposition of Σ:

$$\Sigma = U\Lambda U^T,$$

Standardization

where U is the matrix of (orthogonal) eigenvectors and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ the diagonal matrix of the eigenvalues of Σ. Then it follows

$$\Sigma^{1/2} = U\Lambda^{1/2}U^T = U \,\text{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}\right) U^T,$$
$$\Sigma^{-1/2} = U\Lambda^{-1/2}U^T = U \,\text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_p}}\right) U^T.$$

Corollary 2.2: Let $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$, then it holds

a) $\mathbb{E}\underline{X} = \underline{\mu}$, $\text{Cov}(\underline{X}) = \Sigma$

b) $\underline{X}$ has characteristic function

$$\varphi_{\underline{X}}(\underline{t}) = \exp\left[i\underline{t}^T\underline{\mu} - \tfrac{1}{2}\underline{t}^T\Sigma\underline{t}\right].$$

c) The standardization of $\underline{X}$ is effected by the transformation

$$\underline{Y} = \Sigma^{-1/2}\left(\underline{X} - \underline{\mu}\right) \sim N_p\left(\underline{0}, I_p\right),$$

i.e. $\underline{Y}$ has components $Y_k \underset{\text{i.i.d.}}{\sim} N(0,1)$; $k = 1, ..., p$.

Standardization

Proof of c): For the linear transformation $\underline{Y} = \Sigma^{-1/2}\left(\underline{X} - \underline{\mu}\right)$ we have:

$$\underline{X} = \Sigma^{1/2}\underline{Y} + \underline{\mu}\ , \det(\mathcal{J}) = |\Sigma|^{1/2}$$

and thus it follows from the transformation theorem that

$$\begin{aligned} f_{\underline{Y}}(\underline{y}) &= |\det(\mathcal{J})| f_{\underline{X}}\left(\Sigma^{1/2}\underline{y} + \underline{\mu}\right) \\ &= (2\pi)^{-p/2} \exp\left[-\frac{1}{2}\left(\Sigma^{1/2}\underline{y} + \underline{\mu} - \underline{\mu}\right)^T \Sigma^{-1}\left(\Sigma^{1/2}\underline{y} + \underline{\mu} - \underline{\mu}\right)\right] \\ &= (2\pi)^{-p/2} \exp\left[-\frac{1}{2}\underline{y}^T\underline{y}\right] \\ &= \prod_{i=1}^{p}\left(\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y_i^2}{2}\right]\right) = \prod_{i=1}^{p} f_{Y_i}(y_i) \end{aligned}$$

with $Y_i \underset{\text{i.i.d}}{\sim} N(0,1);\ i = 1, \ldots, p.$

chf of multivariate normal

Proof of b): Each component Y_k has characteristic function $\varphi_{Y_k}(t) = \exp(-t^2/2)$. From the definition of the chf of random vectors we then have, due to the independence of $Y_1, ..., Y_p$:

$$\varphi_{\underline{Y}}(\underline{t}) = \mathbb{E}\left[\exp\left(i\underline{t}^T\underline{Y}\right)\right] = \prod_{k=1}^{p} \mathbb{E}\left[\exp\left(it_k Y_k\right)\right] = \prod_{k=1}^{p} \varphi_{Y_k}(t_k)$$

$$= \prod_{k=1}^{p} \exp(-t_k^2/2) = \exp(-\tfrac{1}{2}\underline{t}^T\underline{t}).$$

Using Corollary 2.1, we obtain for $\underline{X} = \Sigma^{1/2}\underline{Y} + \underline{\mu}$:

$$\varphi_{\underline{X}}(\underline{t}) = \exp\left[i\underline{t}^T\underline{\mu}\right] \varphi_{\underline{Y}}\left((\Sigma^{1/2})^T\underline{t}\right)$$

$$= \exp\left[i\underline{t}^T\underline{\mu}\right] \exp\left[-\frac{1}{2}\underline{t}^T(\Sigma^{1/2})(\Sigma^{1/2})^T\underline{t}\right]$$

$$= \exp\left[i\underline{t}^T\underline{\mu} - \frac{1}{2}\underline{t}^T\Sigma\underline{t}\right]$$

chf of multivariate normal

Proof of a): For $\underline{Y} = \Sigma^{-1/2}\left(\underline{X} - \underline{\mu}\right) \sim N_p\left(\underline{0}, I_p\right)$ it holds:

$$\mathbb{E}\underline{Y} = \underline{0}_p, \quad \mathrm{Cov}(\underline{Y}) = I_p.$$

For the transformed $\underline{X} = \Sigma^{1/2}\underline{Y} + \underline{\mu}$ we then have :

$$\begin{aligned}
\mathbb{E}\underline{X} &= \Sigma^{1/2}\mathbb{E}(\underline{Y}) + \underline{\mu} = \underline{\mu} \\
\mathrm{Cov}(\underline{X}) &= \mathbb{E}\left[\left(\underline{X} - \mathbb{E}\underline{X}\right)\left(\underline{X} - \mathbb{E}\underline{X}\right)^T\right] \\
&= \mathbb{E}\left[\left(\Sigma^{1/2}\underline{Y} + \underline{\mu} - \underline{\mu}\right)\left(\Sigma^{1/2}\underline{Y} + \underline{\mu} - \underline{\mu}\right)^T\right] \\
&= \mathbb{E}\left[\Sigma^{1/2}\underline{Y}\,\underline{Y}^T\left(\Sigma^{1/2}\right)^T\right] = \Sigma^{1/2}\mathbb{E}\left(\underline{Y}\,\underline{Y}^T\right)\left(\Sigma^{1/2}\right)^T \\
&= \Sigma^{1/2} I_p \left(\Sigma^{1/2}\right)^T = \left(U\Lambda^{1/2}U^T\right)\left(U\Lambda^{1/2}U^T\right) = U\Lambda U^T = \Sigma,
\end{aligned}$$

observing that $\mathbb{E}\left(\underline{Y}\,\underline{Y}^T\right) = \mathrm{Cov}(\underline{Y}) + (\mathbb{E}Y)(\mathbb{E}Y)^T = \mathrm{Cov}(\underline{Y}) = I_p$.

Simulation of normal random vectors

Remark: For the standardization we can also make use of the **Cholesky decomposition**: $\Sigma = R^T R$, with upper triangular matrix R.

a) $\underline{Y} \sim N(\underline{0}, I_p) \Rightarrow \underline{X} = R^T \underline{Y} + \underline{\mu} \sim N(\underline{\mu}, R^T R = \Sigma)$

b) $\underline{X} \sim N(\underline{\mu}, \Sigma) \Rightarrow \underline{Y} = (R^T)^{-1}(\underline{X} - \underline{\mu}) \sim N(\underline{0}, I_p)$

This can also be used for the simulation of normal random vectors.

Example (Simulating normal rv's)

```
> p=3; n=25; mu=c(1,0,2)
> Sigma= matrix(c(1,1,1,1,2,1,1,1,3), p, p)
> R= chol(Sigma); R
     [,1] [,2]     [,3]
[1,]    1    1 1.000000
[2,]    0    1 0.000000
[3,]    0    0 1.414214
```

Simulation of normal random vectors

Example (cont'd)

$>$ t(R)% $*$ %R

	[,1]	[,2]	[,3]
[1,]	1	1	1
[2,]	1	2	1
[3,]	1	1	3

$>$ n=25; y= rnorm(n*p) #generates 75 N(0,1)-distributed random numbers

$>$ Y= matrix(y, p, n) #Y contains 25 standard normals $\sim N_p(\underline{0}, I_p)$

$>$ X= t(R)% $*$ %Y + mu #X contains 25 normals $N_p(\underline{\mu}, \Sigma)$

Alternatively, we may generate normally distributed random vectors directly using the command "mvrnorm" in library(MASS):

$>$ mvrnorm(n=25, mu, Sigma) #generates 25 rv's $\sim N_p(\underline{\mu}, \Sigma)$

Lower-dimensional projections

From Corollary 2.1 and Corollary 2.2 we can easily deduce that any subset or linear combination of components of a normally distributed random vector again has a normal distribution.

Corollary 2.3: Let $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$, $A \in \mathbb{R}^{q \times p}$ a matrix with $\operatorname{rank}(A) = q, 1 \leqslant q \leqslant p$ (i.e. A has full row rank) and $\underline{b} \in \mathbb{R}^q$ (translation vector). Then it holds:

$$\underline{Y} = A\underline{X} + \underline{b} \sim N_q\left(A\underline{\mu} + \underline{b}, A\Sigma A^T\right)$$

Proof:

$$\begin{aligned}
\varphi_{\underline{Y}}(\underline{t}) &\underset{Cor.2.1}{=} \exp\left[i\underline{t}^T\underline{b}\right]\varphi_{\underline{X}}(A^T\underline{t}) \\
&\underset{\text{Cor.2.2b)}}{=} \exp\left[i\underline{t}^T\underline{b}\right]\exp\left[i\left(A^T\underline{t}\right)^T\underline{\mu} - \frac{1}{2}\left(A^T\underline{t}\right)^T\Sigma\left(A^T\underline{t}\right)\right] \\
&= \exp\left[i\underline{t}^T\left(A\underline{\mu} + \underline{b}\right) - \frac{1}{2}\underline{t}^T A\Sigma A^T\underline{t}\right].
\end{aligned}$$

This is, however, the characteristic function of $N_q\left(A\underline{\mu} + \underline{b}, A\Sigma A^T\right)$.

Lower-dimensional projections

In particular, setting $A = \underline{e}_i^T = (0, 0, ..., 0, \underbrace{1}_{i}, 0, ..., 0)$, $b = 0$, means:

$$\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right) \Longrightarrow X_i \sim N\left(\mu_i, \sigma_{ii}\right) \ \forall i = 1, \ldots, p \ ,$$

where σ_{ii} represents the i-th diagonal element of Σ.

Note that the converse is not true, from the normality of the sub-vectors of $\underline{X}$ we cannot infer the full distribution of $\underline{X}$.

Further, setting $A = \underline{a}^T = (a_1, \ldots, a_p); \underline{a} \in \mathbb{R}^p$ and $b = 0$, we obtain

$$Y = \underline{a}^T \underline{X} \Longrightarrow Y \sim N(\underline{a}^T \mu, \underline{a}^T \Sigma \underline{a}) \ \ \forall \underline{a} \in \mathbb{R}^p \quad (P)$$

This, in conjunction with the Cramér-Wold theorem, means that the multivariate normal distribution is completely determined by its first two moments $\underline{\mu}$ and Σ.
It also demonstrates that the covariance matrix is positive semidefinite:

$$\text{Var}(\underline{a}^T \underline{X}) = \underline{a}^T \,\text{Cov}(\underline{X}) \underline{a} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j \,\text{Cov}(X_i, X_j) \geq 0 \ \ \forall \underline{a} \in \mathbb{R}^p.$$

Portfolio optimization

Property (P) finds important applications in the so-called **Portfolio optimization** problem.

Example (Portfolio optimization)

Let $X_1, X_2, \ldots X_p$ represent stock values and $a_1, \ldots, a_p$ the weights (percentages) allocated to the corresponding stocks in the portfolio $Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p$. Clearly, for the weights we have: $a_i \geq 0$ and $\sum_{i=1}^{p} a_i = 1$. The risk of the portfolio can be measured by its variance, thus we have to solve the following optimization problem:

$$\text{Var}\,(a_1X_1 + \cdots + a_pX_p) \to \min \text{ subject to}$$

$$\sum_{i=1}^{p} a_i\mu_i \geq \mu_0;\ a_1, \ldots, a_p \geq 0;\ \sum_{i=1}^{p} a_i = 1,$$

where μ_0 stands for the least revenue to be earned by the portfolio.

2.3 Independence, conditional normals

When are two sub-vectors of $\underline{X}$ independent?

Theorem (Independence of sub-vectors)

Let $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ and $A \in \mathbb{R}^{r \times p}$, $B \in \mathbb{R}^{q \times p}$; $r, q \geq 1$; such that $r + q \leqslant p$. Then the random vectors $A\underline{X}$ and $B\underline{X}$ are stochastically independent if and only if

$$A\Sigma B^T = \underset{(r,q)}{\mathbb{O}} \quad \textit{(generalized orthogonality)} \qquad (\star)$$

Proof: a) $''\Longrightarrow''$

Assume $\underline{Y} = A\underline{X}$, $\underline{Z} = B\underline{X}$ to be independent. This implies:

$$\mathrm{Cov}(\underline{Y}, \underline{Z}) = \big(\mathrm{Cov}(Y_i, Z_j)\big)_{\substack{i=1,\ldots,r \\ j=1,\ldots,q}} = \underset{(r,q)}{\mathbb{O}}$$

Further, it follows

Independence of sub-vectors

$$
\begin{aligned}
\mathbb{O} = \mathrm{Cov}(\underline{Y}, \underline{Z}) &= \mathbb{E}\left\{(\underline{Y} - \mathbb{E}\underline{Y})\,(\underline{Z} - \mathbb{E}\underline{Z})^T\right\} \\
&= \mathbb{E}\left\{(A\underline{X} - A\underline{\mu})\,(B\underline{X} - B\underline{\mu})^T\right\} \\
&= \mathbb{E}\left\{A(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T B^T\right\} \\
&= A\left\{\mathbb{E}((\underline{X} - \underline{\mu})\,(\underline{X} - \underline{\mu})^T)\right\} B^T \\
&= A\mathrm{Cov}(\underline{X})B^T = A\Sigma B^T.
\end{aligned}
$$

b) [$''\Longleftarrow''$] Let $A\Sigma B^T = \mathbb{O}$.

W.l.o.g. we assume $\mathrm{rank}(A) = r$, $\mathrm{rank}(B) = q$, i.e. A, B have full rank.

Define $C = \begin{bmatrix} A \\ B \end{bmatrix}$. From $A\Sigma B^T = \left(A\Sigma^{1/2}\right)\left(\Sigma^{1/2}B^T\right) = \mathbb{O}$ we have:

$$
C\underline{X} = C\Sigma^{1/2}\left(\Sigma^{-1/2}\underline{X}\right) = \begin{bmatrix} A\underline{X} \\ B\underline{X} \end{bmatrix} \sim N_{q+r}\left(C\underline{\mu}, C\Sigma C^T\right) \ ,
$$

Independence of sub-vectors

where

$$C\Sigma C^T = \begin{bmatrix} A\Sigma A^T & \mathbb{O} \\ \mathbb{O} & B\Sigma B^T \end{bmatrix}.$$

This, in turn, implies that $|C\Sigma C^T| = |A\Sigma A^T| * |B\Sigma B^T| \neq 0$, and thus:

$$\begin{aligned} f_{(A\underline{X},B\underline{X})}(\underline{y},\underline{z}) &= (2\pi)^{-(r+q)/2} \left(\left| A\Sigma A^T \right| \cdot \left| B\Sigma B^T \right| \right)^{-1/2} \\ &\quad \cdot \exp\left[-\frac{1}{2} \left(C\underline{x} - C\underline{\mu} \right)^T \left(C\Sigma C^T \right)^{-1} \left(C\underline{x} - C\underline{\mu} \right) \right] \\ &= (2\pi)^{-\frac{r}{2}} |A\Sigma A^T|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\underline{y} - A\underline{\mu})^T (A\Sigma A^T)^{-1} (\underline{y} - A\underline{\mu}) \right] \\ &\quad \cdot (2\pi)^{-\frac{q}{2}} |B\Sigma B^T|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\underline{z} - B\underline{\mu})^T (B\Sigma B^T)^{-1} (\underline{z} - B\underline{\mu}) \right] \\ &= f_{A\underline{X}}(\underline{y}) \cdot f_{B\underline{X}}(\underline{z}), \text{ i.e. } A\underline{X} \text{ und } B\underline{X} \text{ are stoch. independent.} \end{aligned}$$

Conditional normal distribution

We will now consider conditional distributions, starting with the partitioning

$$\underline{Z} = \begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix} \begin{matrix} \in \mathbb{R}^r \\ \in \mathbb{R}^q \end{matrix} \sim N_p\left(\underline{\mu}, \Sigma\right), \quad r + q = p.$$

$$\mathbb{E}\underline{Z} = \underline{\mu} = \begin{pmatrix} \underline{\mu}_X \\ \underline{\mu}_Y \end{pmatrix}, \quad \mathrm{Cov}(\underline{Z}) = \Sigma = \begin{bmatrix} \underset{(r,r)}{\Sigma_{\underline{XX}}} & \underset{(r,q)}{\Sigma_{\underline{XY}}} \\ \underset{(q,r)}{\Sigma_{\underline{YX}}} & \underset{(q,q)}{\Sigma_{\underline{YY}}} \end{bmatrix},$$

$$\Sigma_{\underline{XY}} = \Sigma_{\underline{YX}}^T = \mathrm{Cov}(\underline{X}, \underline{Y}).$$

The marginal distributions read

$$\underline{X} \sim N_r\left(\underline{\mu}_X, \Sigma_{\underline{XX}}\right) \text{ and } \underline{Y} \sim N_q\left(\underline{\mu}_Y, \Sigma_{\underline{YY}}\right).$$

Obviously, $\underline{X}$, $\underline{Y}$ are uncorrelated (independent) if $\Sigma_{\underline{XY}} = \Sigma_{\underline{YX}}^T = \underset{(r,q)}{\mathbb{O}}$.

Conditional normal distribution

Theorem (Conditional distribution of $\underline{Y}|\underline{X} = \underline{x}$)

Let $\underline{X} \sim N_p(\mu, \Sigma)$. *Then*

$$P_{\underline{Y}|\underline{X}=\underline{x}} = N_r\left(\underline{\mu}_{\underline{Y}|\underline{x}}, \Sigma_{\underline{Y}|\underline{x}}\right) ,$$

$$\textit{where} \quad \underline{\mu}_{\underline{Y}|\underline{x}} = \mathbb{E}(\underline{Y}|\underline{X} = \underline{x}) = \Sigma_{\underline{Y}\underline{X}}\, \Sigma_{\underline{X}\underline{X}}^{-1}\left(\underline{x} - \underline{\mu}_X\right) + \underline{\mu}_Y$$

$$\textit{and} \quad \Sigma_{\underline{Y}|\underline{x}} = \Sigma_{\underline{Y}\underline{Y}} - \Sigma_{\underline{Y}\underline{X}}\, \Sigma_{\underline{X}\underline{X}}^{-1}\, \Sigma_{\underline{X}\underline{Y}} \quad (\textit{Schur complement}).$$

Proof: Straightforward computation of the conditional pdf via

$$f_{\underline{Y}|\underline{x}} = \frac{f_{\underline{X},\underline{Y}}(\underline{x},\underline{y})}{f_{\underline{X}}(\underline{x})} = \frac{\text{pdf of } N_p\left(\underline{\mu},\Sigma\right)}{\text{pdf of } N_r\left(\underline{\mu}_X,\Sigma_{\underline{X}\underline{X}}\right)}$$

Remarks:

a) $\underline{\mu}_{\underline{Y}|\underline{x}} = \mathbb{E}(\underline{Y}|\underline{X} = \underline{x})$ is called the multivariate regression function of $\underline{Y}$ on $\underline{X}$.

Conditional normal distribution

b) $$\underset{\substack{\text{total}\\ \text{variability}}}{\Sigma_{\underline{Y}\underline{Y}}} = \underset{\substack{\text{residual}\\ \text{variability}}}{\Sigma_{\underline{Y}|\underline{X}}} + \underset{\substack{\text{explained}\\ \text{variability}}}{\Sigma_{\underline{Y}\underline{X}}\Sigma_{\underline{X}\underline{X}}^{-1}\Sigma_{\underline{X}\underline{Y}}} \qquad \text{(generalized } R^2\text{)}$$

c) Special case: $q = 1$, prediction of Y for given realization $\underline{X} = \underline{x}$.

$$\mathbb{E}(Y|\underline{X} = \underline{x}) = \underline{c}^T \Sigma_{\underline{X}\underline{X}}^{-1}(\underline{x} - \mu_{\underline{X}}) + \mu_Y \qquad (\star)$$

$$= \beta_0 + \underline{\beta}^T \underline{x} \qquad \text{multiple linear regression}$$

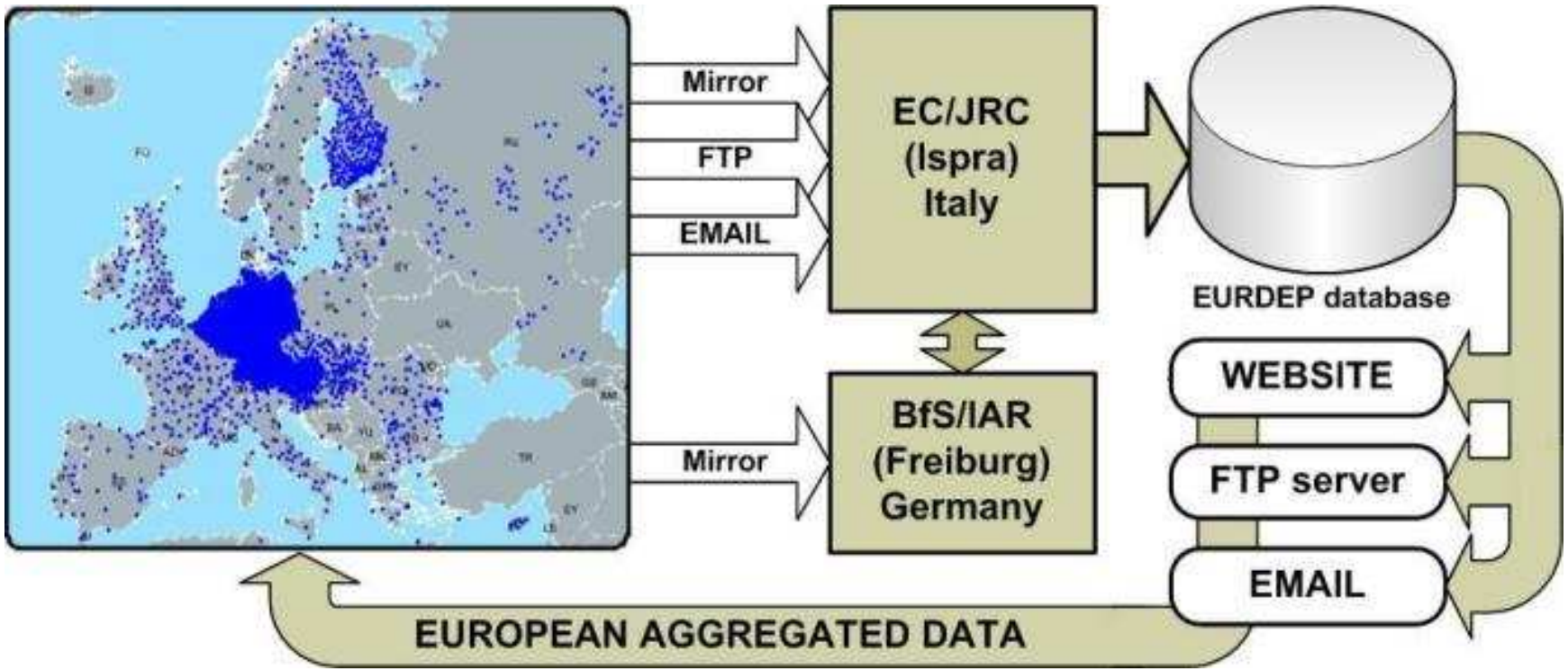
where $\beta_0 = \mu_Y - \underline{c}^T \Sigma_{\underline{X}\underline{X}}^{-1} \mu_{\underline{X}}$ intercept

$\underline{\beta} = \Sigma_{\underline{X}\underline{X}}^{-1} \underline{c}$ slope vector

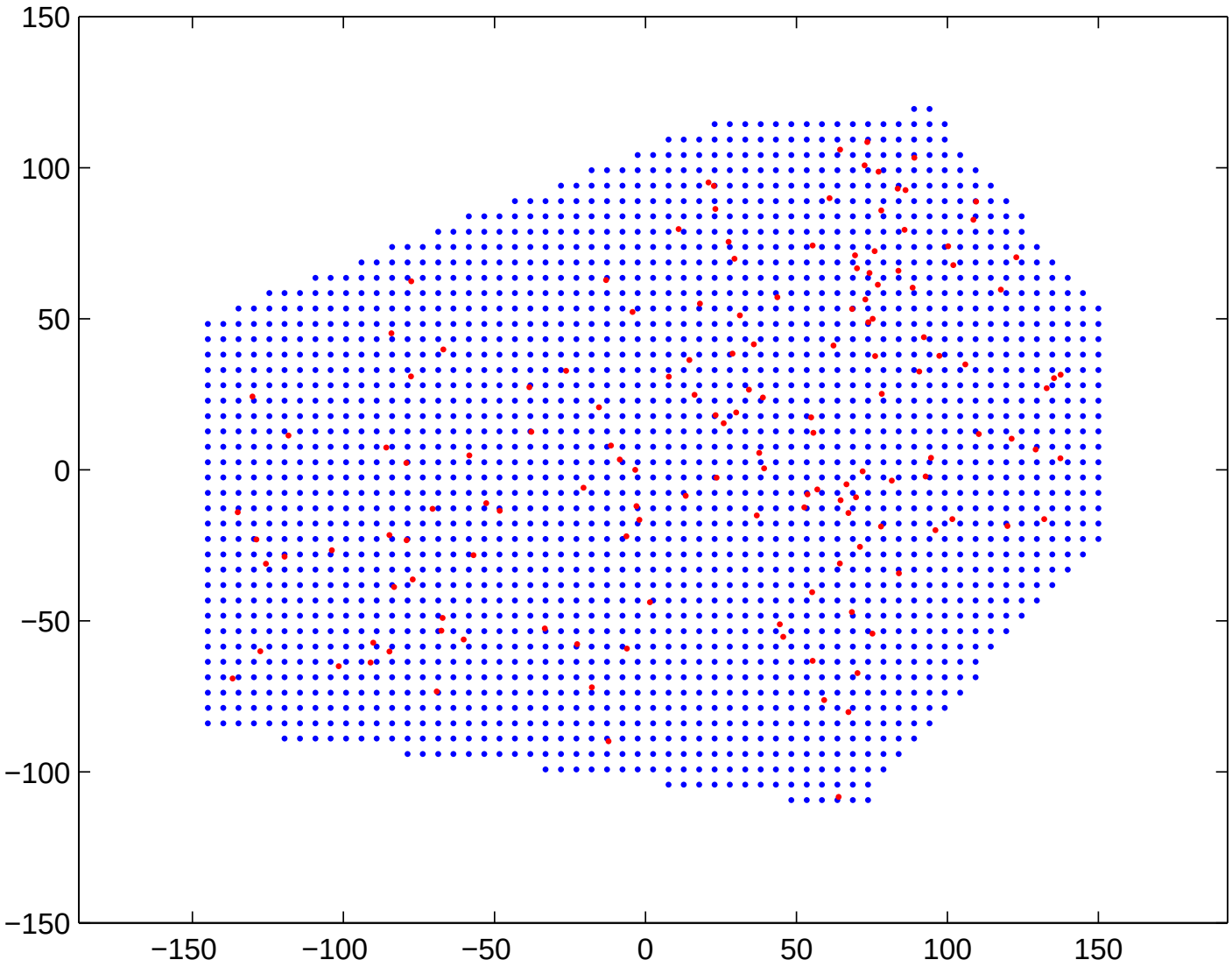
$\underline{c} = (\text{Cov}(Y, X_1), ..., \text{Cov}(Y, X_q))^T$

Application: spatial interpolation, $(\star)$ defines a Kriging predictor (spatial BLUP).

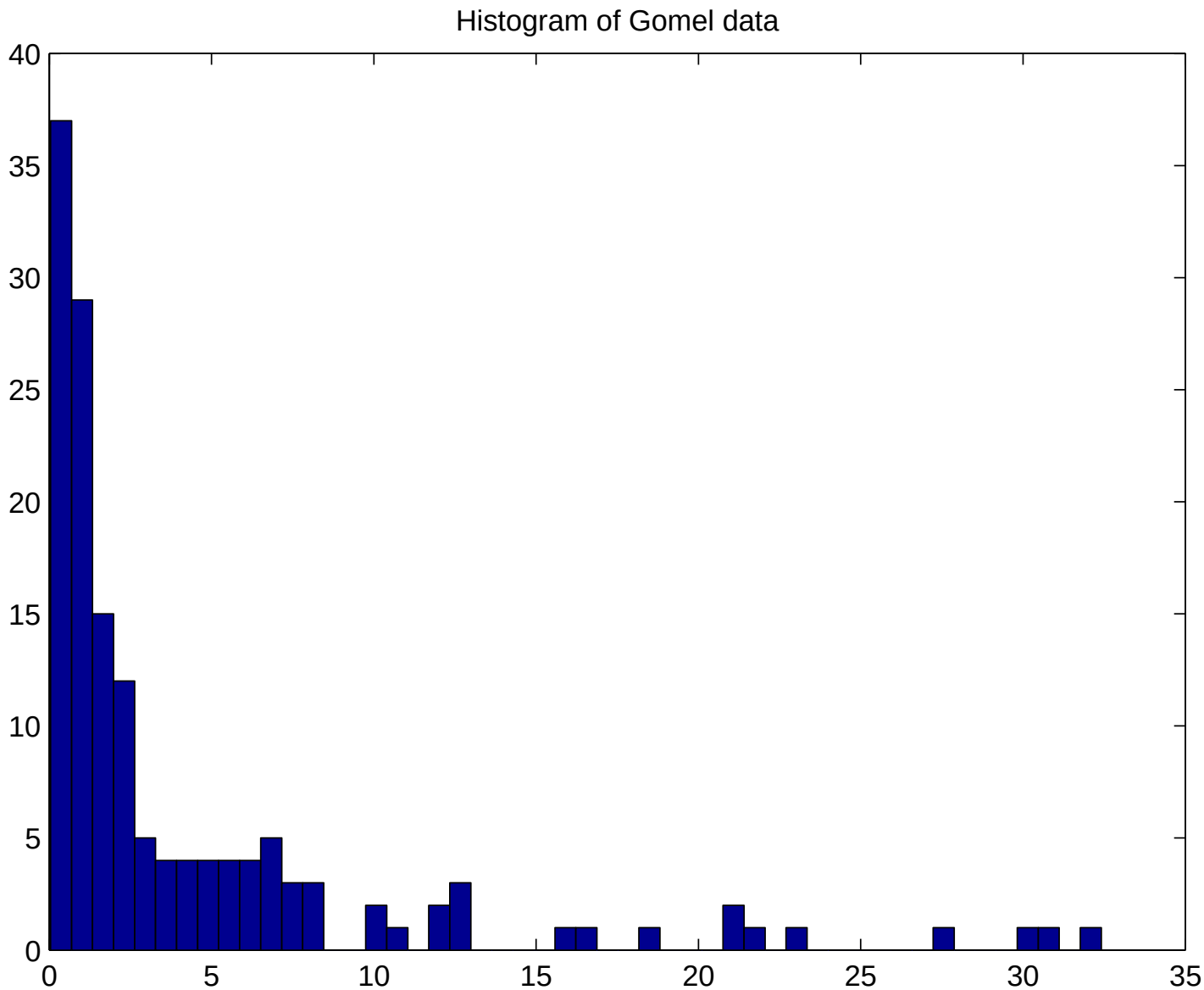
Kriging finds a lot of applications in spatial statistics for mining, environmental monitoring and engineering.

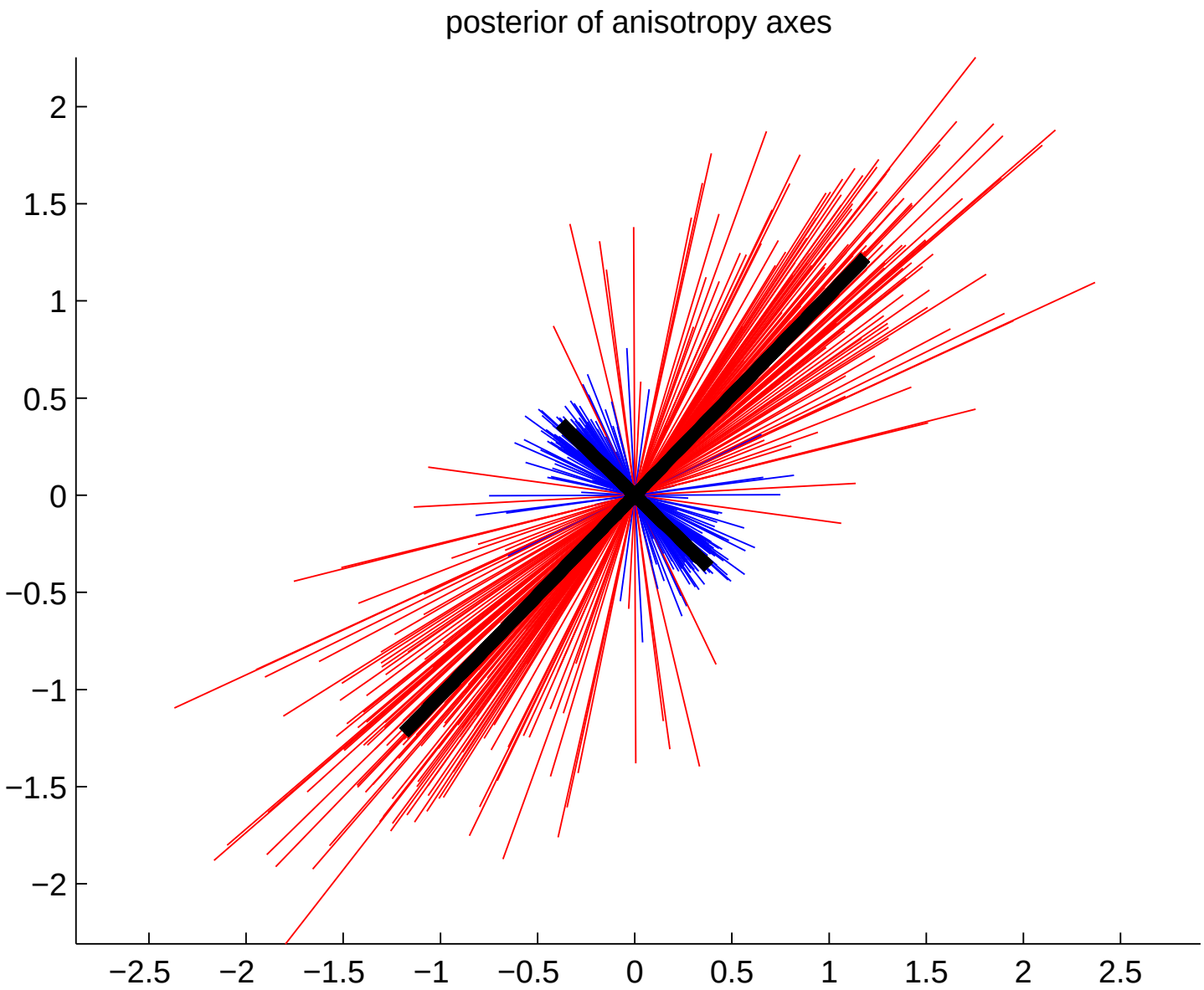


Prediction grid (blue dots), locations of observations (red dots)

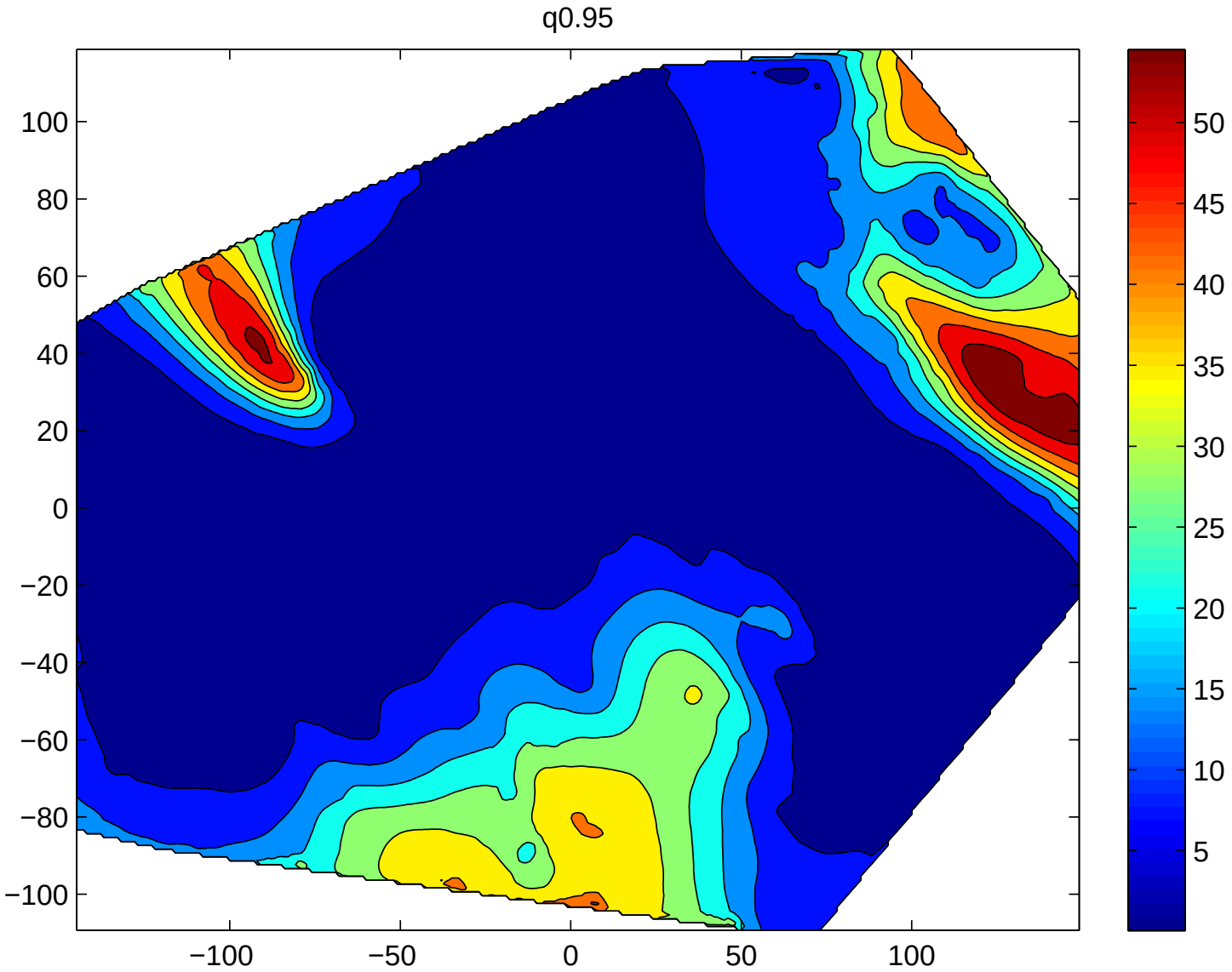


Histogram of Cs 137 data near Gomel (Belarus)

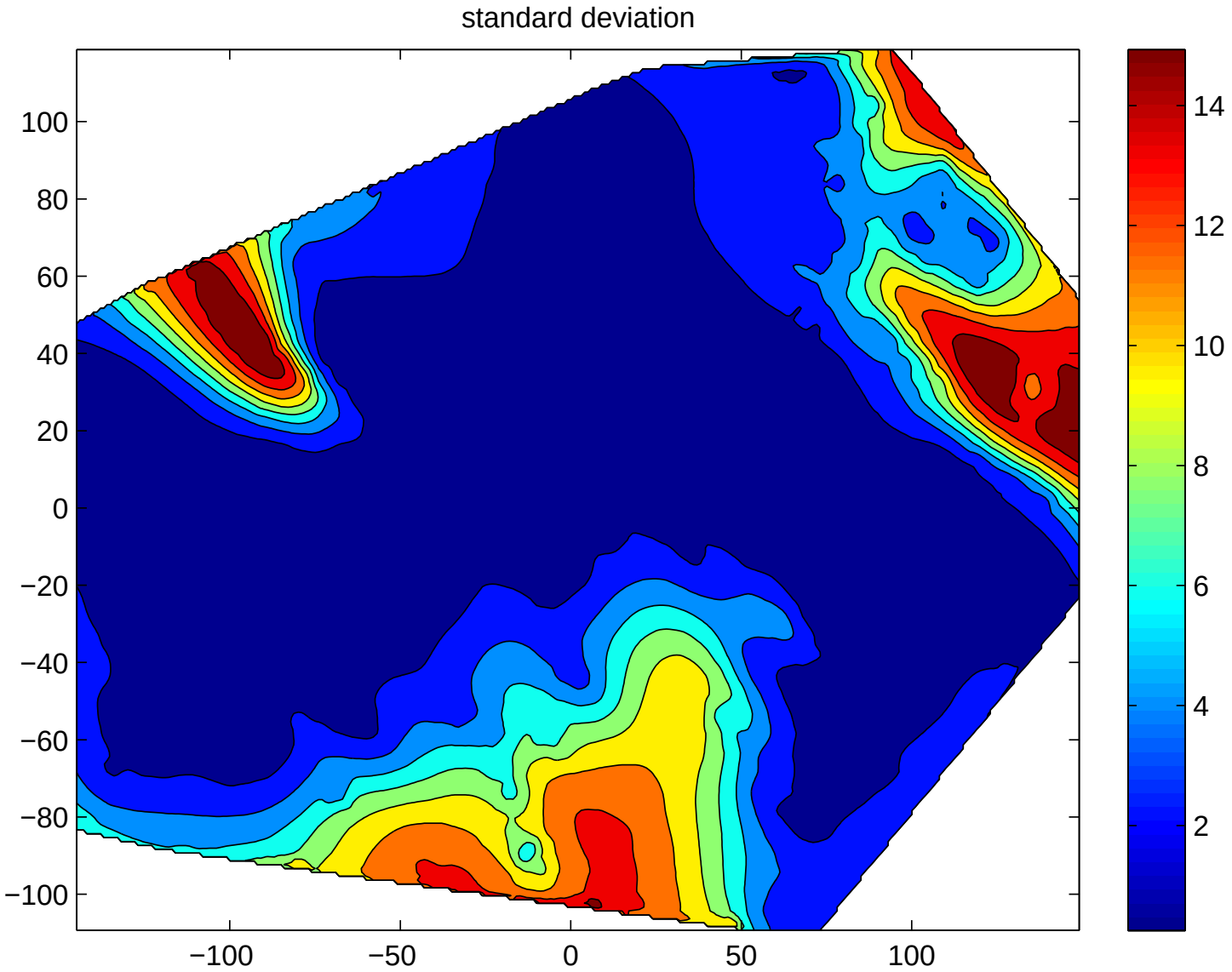



posterior of anisotropy axes
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

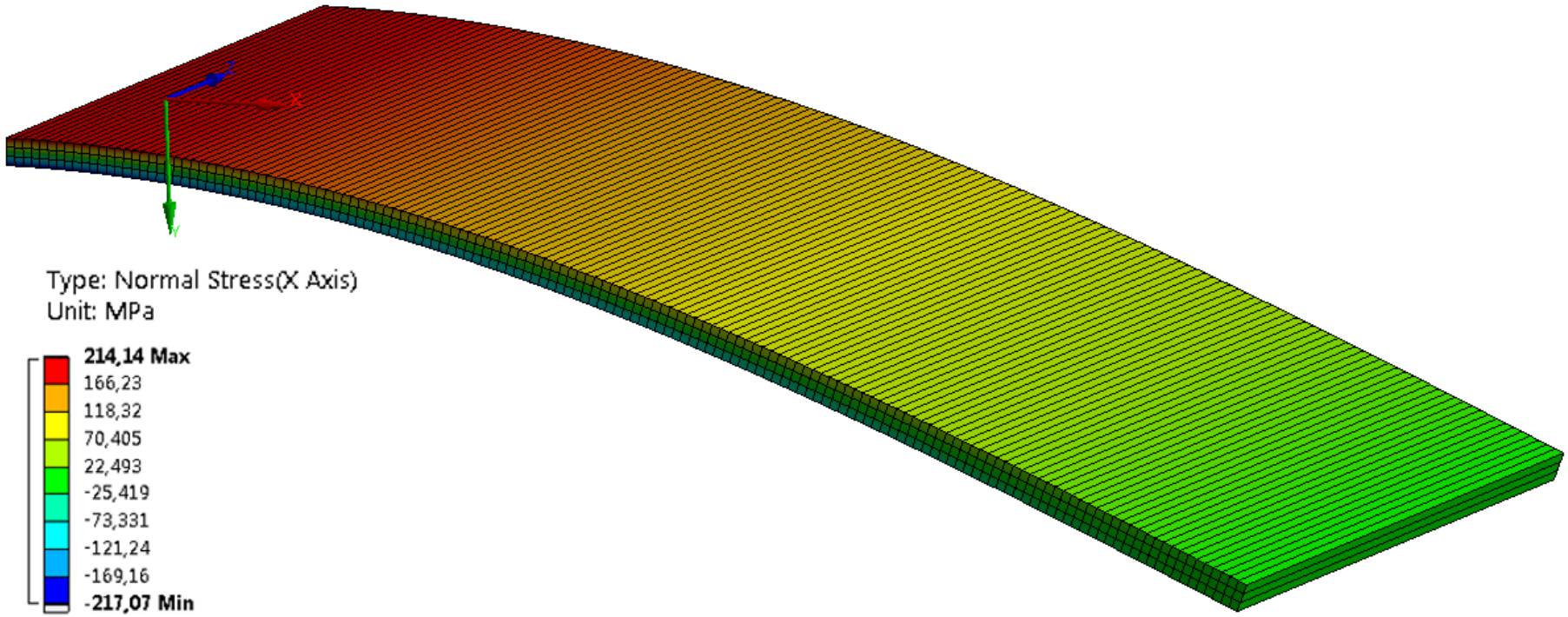
95% posterior quantile of Cs 137 distribution



Standard deviation of Cs 137 posterior distribution



Stress modeling of thin wafers (complementing FEM)



3. Estimation, Testing and Confidence Regions

3.1 Outlier (anomaly) testing

Q: How can we test whether a given observation comes from a normally distributed population with mean $\mathbb{E}\underline{X} = \underline{\mu}$ and $\mathrm{Cov}(\underline{X}) = \Sigma$?

A: Outlyingness judged on the basis of Mahalanobis distance $d_M^2(\underline{X}, \underline{\mu})$

Recall: Sum of squares of p random variables $X_i \overset{iid}{\sim} \mathrm{N}(0,1)$ follows a Chi-Squared-distribution with p degrees of freedom:

$$X_i \overset{iid}{\sim} N(0,1) \Longrightarrow X_1^2 + X_2^2 + \ldots + X_p^2 \sim \chi_p^2$$

Corollary 3.1: If $\underline{Y} \sim N_p(\underline{0}, I_p)$ then it follows

$$\underline{Y}^T \underline{Y} = Y_1^2 + \ldots + Y_p^2 \sim \chi_p^2 .$$

Proof: The components of $\underline{Y}$ are iid - $N(0,1)$ distributed.

Outlier testing

Theorem (Distribution of d_M^2)

$$\underline{X} \sim N_p(\underline{\mu}, \Sigma) \Longrightarrow (\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu}) = d_M^2(\underline{X}, \underline{\mu}) \sim \chi_p^2$$

Proof: After standardization of $\underline{X}$, according to

$$\underline{Y} = \Sigma^{-\frac{1}{2}} (\underline{X} - \underline{\mu}) \sim N_p(\underline{0}, I_p) \ ,$$

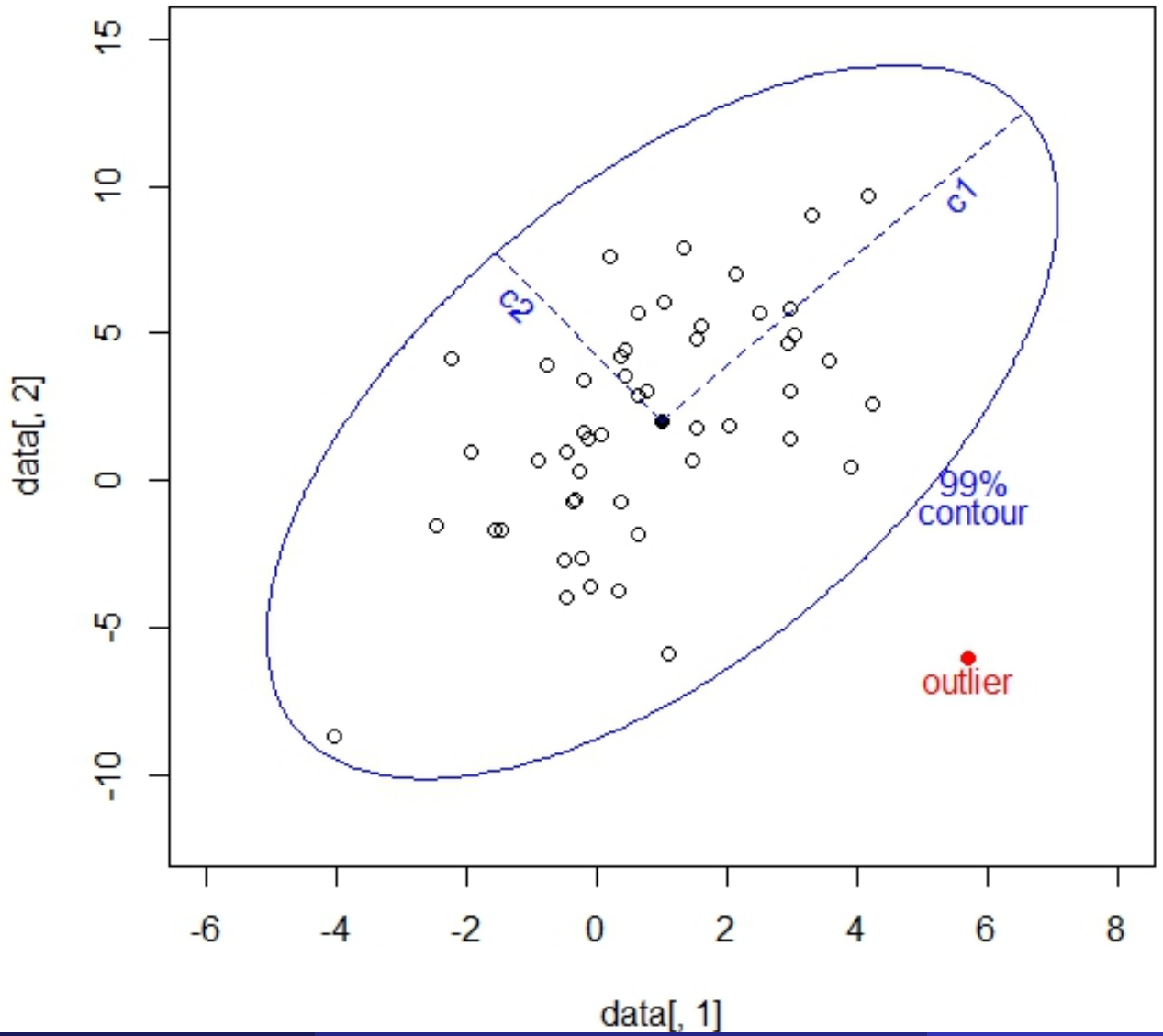
the proof follows from Corollary 3.1:

$$\underline{Y}^T \underline{Y} = (\underline{X} - \underline{\mu})^T \underbrace{\Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}}_{=\Sigma^{-1}} (\underline{X} - \underline{\mu}) = d_M^2(\underline{X}, \underline{\mu}) \sim \chi_p^2 \ .$$

This result can be used for checking the outlyingness of an observation on the basis of the contour levels of this distribution:

$$P((\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu}) \leqslant \chi_{p;1-\alpha}^2) = 1 - \alpha$$

outlier testing: p=2



outlier testing

More generally,

$$\mathscr{E} = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \leq c\}$$

defines a scatter ellipsoid with center $\underline{\mu} \in \mathbb{R}^p$ and semi-axes lengths $c_i = \sqrt{c\lambda_i}$, where $\lambda_i = \lambda_i(\Sigma)$ are the eigenvalues of Σ; $i = 1, ..., p$ and c = $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$-quantile of the Chi-Squared-distribution with p degrees of freedom.

The following code snippet was used to generate the plot above:

```
> n=50; p=2
> y=rnorm(n*p); Y=matrix(y,p,n)
> mu=c(1,2); Sigma=matrix(c(4.,4.8,4.8,16.), p, p)
> R=chol(Sigma); X=t(R)%*%Y+mu; data=t(X)
> d2.99= qchisq(0.99, df = p)
> plot(data, xlim=c(-6,6), ylim=c(-10,12))
> library(cluster); ellipsoidPoints(Sigma, d2.99, loc=mu)
```

Chi-Squared-Probability Plot

Conversely, the χ^2-outlier test is often used in practice for visually checking the normality of data by means of the so-called χ^2-probability plot of the ordered Mahalanobis distances:

$$d^2_{[i]} = (\underline{x}_{[i]} - \underline{\hat{\mu}})^T \, \widehat{\Sigma}^{-1} (\underline{x}_{[i]} - \underline{\hat{\mu}}); \, i = 1, \ldots, n$$

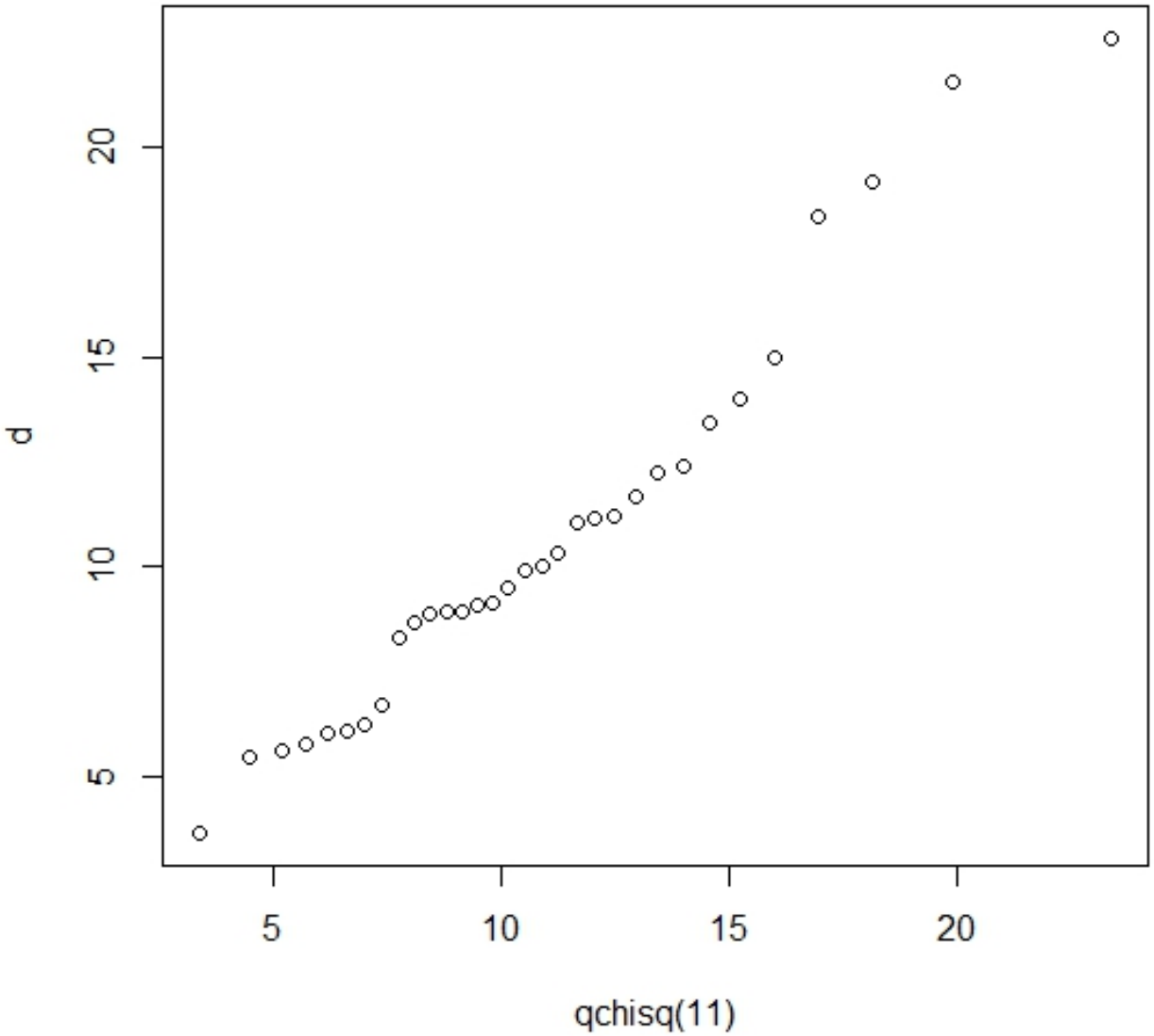
$$d^2_{[1]} \leq d^2_{[2]} \leqslant \ldots \leqslant d^2_{[n]} \text{ und } \underline{d} = (d^2_{[1]}, \ldots, d^2_{[n]})^T$$

where $\underline{x}_{[i]}; \, i = 1, \ldots, n$ are the observed realizations of $\underline{X} \sim N(\mu, \Sigma)$.

The plot of the Mahalanobis distances vs. the ordered χ^2- quantiles proceeds in R as follows:

```
> X = mtcars; a=dim(X)
> d = mahalanobis(X, colMeans(X), cov(X))
> qqplot(qchisq(ppoints(a[1]), df=a[2]), d, xlab="qchisq(df=11)")
```

Chi-Squared-Probability Plot



Additional visual checks should be made via pairwise scatterplots.

Note: Non-centered quadratic forms $\underline{X}^T \Sigma^{-1} \underline{X}$ lead to so-called non-central χ^2-distributions: χ^2_λ with non-centrality parameter $\lambda = \frac{1}{2} \underline{\mu}^T \underline{\mu}$

In generalization of the preceding Corollary 3.1 it holds:

Theorem (Idempotent quadratic forms)

Let $\underline{Y} \sim N_p(\underline{0}, I_p)$. Then it holds:

$\underline{Y}^T A \underline{Y} \sim \chi^2_k$ with k = rank(A) $\Longleftrightarrow$ A is idempotent.

Corollary 3.1 represents the special case in which $A = I_p$: this is the trivial and only idempotent matrix which has full rank.

Remark: An idempotent matrix A has the following properties:
a) A ist quadratic, symmetric, and positive semidefinite.
b) The eigenvalues are either zero or one: $\lambda_i(A) \in \{0, 1\}; i = 1, ..., p$.

Projection matrices

c) $rank(A) = tr(A)$

d) Idempotent and symmetric matrices are projection matrices.

For example, the Hat-Matrix $H = X(X^TX)^{-1}X^T$, known from linear regression, is a projection matrix and effects a decomposition $\underline{Y} = H\underline{Y} + (I - H)\underline{Y} = X(X^TX)^{-1}X^T\underline{Y} + (I - H)\underline{Y}$, where

$$\underline{\hat{Y}} = X\underline{\hat{\beta}} = X(X^TX)^{-1}X^T\underline{Y} = H\underline{Y}$$

is the Gauss-Markov-Predictor of $\mathbb{E}\underline{Y} = X\underline{\beta}$.

The following theorem tells us when quadratic forms and linear combinations $C\underline{Y}$ with $C \in \mathbb{R}^{q\times p}$, $q \leqslant p$, are stoch. independent.

Theorem (Independence of quadratic forms)

Let $\underline{Y} \sim N(\underline{0}, I_p)$ and $A, B \in \mathbb{R}^{p\times p}$ be projection matrices. Then it holds:

a) $\underline{Y}^TA\underline{Y}$ and $\underline{Y}^TB\underline{Y}$ are stochastically independent $\Leftrightarrow A \cdot B = \underset{p\times p}{\mathbb{O}}$

b) $\underline{Y}^TA\underline{Y}$ and $\underset{q\times p}{C}\underline{Y}$ are stochastically independent if $C \cdot A = \underset{q\times p}{\mathbb{O}}$.

Independence of quadratic forms

This theorem follows from a more general result due to Cochran (see G.F.A. Seber: Multivariate Observations. Wiley 2004, Section 9.6).

Assertion a) finds many applications in (M)ANOVA. For an application of assertion b) consider, again, a linear regression model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}\,, \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_p)$$

with $(p \times r)$ model matrix X. The mle of σ^2 can be written as

$$\hat{\sigma}^2 = \frac{1}{p-r}\left(\underline{Y} - X\underline{\hat{\beta}}\right)^T \left(\underline{Y} - X\underline{\hat{\beta}}\right) = \frac{1}{p-r}\,\underline{Y}^T(I_p - H)\underline{Y}\,,$$

which is stoch. independent of the mle $\underline{\hat{\beta}} = \left(X^T X\right)^{-1} X^T Y =: C\underline{Y}$, observing that $X^T H = X^T$ and thus $C \cdot (I_p - H) = \underset{r\times p}{\mathbb{O}}$.

3.2 Maximum-Likelihood-Estimation of expectation and covariance matrix

Given: data $\underline{x}_1, \ldots, \underline{x}_n$ as realizations of $\underline{X}_i \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma); i = 1, \ldots, n$
Goal: Estimation of $\underline{\mu}$ und Σ
The likelihood function (product of the pdf's of $\underline{X}_i$) is obtained as:

$$L(\underline{\mu}, \Sigma; \underline{x}_1, \ldots, \underline{x}_n) = \prod_{i=1}^{n} \left\{ (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu})) \right\}$$

$$= (2\pi)^{-np/2} |\Sigma|^{-n/2} exp(-\frac{1}{2} \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}))\}.$$

The log-likelihood-function $l(\underline{\mu}, \Sigma; \underline{x}_1, \ldots, \underline{x}_n) = logL(\underline{\mu}, \Sigma; \underline{x}_1, \ldots, \underline{x}_n)$ then reads

$$l(\underline{\mu}, \Sigma; \underline{x}_1, \ldots, \underline{x}_n) = -\frac{np}{2} log(2\pi) - \frac{n}{2} log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}).$$

In order to determine the mle's of $\underline{\mu}$ and Σ we have to solve the likelihood equations

$$\frac{\partial}{\partial \underline{\mu}} l(\underline{\mu}, \Sigma; \underline{x}_1, \ldots, \underline{x}_n) = \underline{0}_p \ (*)$$

$$\frac{\partial}{\partial \Sigma} l(\underline{\mu}, \Sigma; \underline{x}_1, \ldots, \underline{x}_n) = \underset{p \times p}{\mathbb{O}} \ (**)$$

Taking the gradient on the lhs of $(*)$ we obtain

$$\nabla_{\underline{\mu}} \left[-\frac{np}{2} log(2\pi) - \frac{n}{2} log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}) \right]$$

$$= -\frac{1}{2} \nabla_{\underline{\mu}} \left[\sum_{i=1}^{n} \underline{x}_i{}^T \Sigma^{-1} \underline{x}_i - 2\underline{\mu}^T \Sigma^{-1} \sum_{i=1}^{n} \underline{x}_i + n\underline{\mu}^T \Sigma^{-1} \underline{\mu} \right]$$

$$= -\frac{1}{2} \left[-2\Sigma^{-1} \sum_{i=1}^{n} \underline{x}_i + 2n\Sigma^{-1} \underline{\mu} \right] .$$

MLE of $\underline{\mu}$

Setting this expression equal to the vector of zeros leads to

$$\Sigma^{-1} \sum_{i=1}^{n} \underline{x_i} = n\Sigma^{-1} \underline{\mu}.$$

Multiplying both sides with Σ from the left we get the solution

$$\underline{\hat{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x_i} \, .$$

In order to solve (**) we need rules for computing derivatives of (scalar) functions of matrices:

1. $\frac{\partial}{\partial \Sigma} \log(\det(\Sigma)) = (\Sigma^T)^{-1}$
2. $\frac{\partial}{\partial \Sigma} \operatorname{tr}(C\Sigma) = C^T$
3. $\frac{\partial}{\partial \Sigma} \operatorname{tr}(C\Sigma^{-1}) = -(\Sigma^{-1} C \Sigma^{-1})^T,$

see e.g. J.R. Magnus, H. Neudecker: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley 2019.

MLE of Σ

Now we turn to eq. (**); the lhs reads:

$$\frac{\partial}{\partial \Sigma} l(\underline{\mu}, \Sigma, \underline{x}_1, \ldots, \underline{x}_n) = \frac{\partial}{\partial \Sigma} \left[-\frac{n}{2} \log(\det \Sigma) - \frac{1}{2} \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}) \right]$$

Observing that $(\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}) = \operatorname{tr}\left[\Sigma^{-1} (\underline{x}_i - \mu)(\underline{x}_i - \underline{\mu})^T\right]$, we further obtain

$$\frac{\partial}{\partial \Sigma}\, l(\underline{\mu}, \Sigma, \underline{x}_1, \ldots, \underline{x}_n) = \frac{\partial}{\partial \Sigma} \left[-\frac{n}{2} \log(\det \Sigma) - \frac{1}{2} \operatorname{tr}\left[\Sigma^{-1} C(\underline{\mu})\right] \right]$$

where we have set

$$C(\underline{\mu}) = \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T \,.$$

MLE of Σ

Using the rules given above, the differentiation w.r.t. Σ then yields

$$\frac{\partial}{\partial \Sigma} l(\underline{\mu}, \Sigma, \underline{x}_1, \ldots, \underline{x}_n) = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1} C(\underline{\mu})\Sigma^{-1} \ .$$

Equating the rhs to the $(p \times p)$-matrix of zeros amounts to

$$n\Sigma^{-1} = \Sigma^{-1} C(\underline{\mu})\Sigma^{-1} \Leftrightarrow n\Sigma = C(\underline{\mu}) \ .$$

Plugging in the mle of μ, we finally obtain the mle of Σ:

$$\hat{\Sigma} = \frac{1}{n} C(\hat{\underline{\mu}}) = \frac{1}{n} \sum_{i=1}^{n} (\underline{x}_i - \hat{\underline{\mu}})(\underline{x}_i - \hat{\underline{\mu}})^T \ .$$

Thus, both the mle's of μ and Σ coincide with the corresponding moment estimators. Whereas $\hat{\underline{\mu}}$ is unbiased, this is not true for $\hat{\Sigma}$.

Bias correction of $\hat{\Sigma}$

Theorem (MLE's of $\underline{\mu}$ and Σ)

The ML-estimators of the parameters $\underline{\mu}$ and Σ of independent data $\underline{X}_i \sim \mathrm{N}_p(\underline{\mu}, \Sigma)$ are given by:

$$\underline{\hat{\mu}}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \text{ with } \mathbb{E}\,\underline{\hat{\mu}}_{ML} = \underline{\mu}$$

$$\hat{\Sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\underline{X}_i - \underline{\hat{\mu}}_{ML})(\underline{X}_i - \underline{\hat{\mu}}_{ML})^T \text{ with } \mathbb{E}\,\hat{\Sigma}_{ML} = (1-\tfrac{1}{n})\Sigma.$$

Proof: (a) $\underline{X}_i \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$

$$\Rightarrow \mathbb{E}(\underline{\hat{\mu}}_{ML}) = \mathbb{E}\,\left(\frac{1}{n}\sum_{i=1}^{n}\underline{X}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\underline{X}_i) = \underline{\mu}.$$

(b)
$$\mathbb{E}(\hat{\Sigma}_{ML}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\underline{X}_i - \underline{\hat{\mu}})(\underline{X}_i - \underline{\hat{\mu}})^T$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\underline{X}_i - \underline{\mu}) + (\underline{\mu} - \underline{\hat{\mu}})\right]\left[(\underline{X}_i - \underline{\mu}) + (\underline{\mu} - \underline{\hat{\mu}})\right]^T$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{(\underline{X}_i - \underline{\mu})(\underline{X}_i - \underline{\mu})^T - 2(\underline{X}_i - \underline{\mu})(\underline{\hat{\mu}} - \underline{\mu}) + (\underline{\hat{\mu}} - \underline{\mu})(\underline{\hat{\mu}} - \underline{\mu})^T\right\}$$

$$= \frac{1}{n}\left\{n\,\mathrm{Cov}(\underline{X}_1) - 2n\,\mathrm{Cov}(\underline{X}_1, \underline{\hat{\mu}}) + n\,\mathrm{Cov}(\underline{\hat{\mu}})\right\}$$

Bias correction of $\hat{\Sigma}$

$$= \Sigma - 2\,\mathrm{Cov}(\underline{X}_1, \underline{\hat{\mu}}) + \mathrm{Cov}(\underline{\hat{\mu}})$$

$$= \Sigma - \tfrac{2}{n}\,\Sigma + \tfrac{1}{n}\,\Sigma = (1 - \tfrac{1}{n})\,\Sigma,$$

observing that $\mathrm{Cov}(\hat{\mu}_{ML}) = \frac{1}{n^2}(n\,\Sigma) = \frac{1}{n}\,\Sigma$ and

$$\mathrm{Cov}(\underline{X}_1, \underline{\hat{\mu}}) = \mathrm{Cov}\left(\underline{X}_1, \tfrac{1}{n}(X_1 + \ldots + X_n)\right) = \tfrac{1}{n}\,\mathrm{Cov}(\underline{X}_1) = \tfrac{1}{n}\,\Sigma.$$

Corollary 3.2: An unbiased estimator of Σ is given by

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\underline{X}_i - \underline{\hat{\mu}}_{ML})(\underline{X}_i - \underline{\hat{\mu}}_{ML})^T\ .$$

Remark: The random matrix $\sum_{i=1}^{n}(\underline{X}_i - \underline{\hat{\mu}})(\underline{X}_i - \underline{\hat{\mu}})^T$ follows a so-called Wishart-distribution (matrix-valued generalization of the χ^2-distribution) with $n-1$ degrees of freedom:

$$(n-1)\,\hat{\Sigma} \sim W_p((n-1), \Sigma)$$

(see Section 3.4).

3.3 Confidence regions and Tests

We first consider confidence regions and tests for the mean vector $\underline{\mu}$ when the covariance matrix Σ is **known**. Starting from the fact that

$$\underline{\hat{\mu}}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \sim N(\underline{\mu}, \Sigma/n) \text{ where } \underline{X}_i \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$$

it follows immediately for its Mahalanobis distance to $\underline{\mu}$:

$$(\underline{\hat{\mu}}_{ML} - \underline{\mu})^T (\Sigma/n)^{-1} (\underline{\hat{\mu}}_{ML} - \underline{\mu}) \sim \chi^2_p \quad (*)$$

This leads us to the following confidence ellipsoid for $\underline{\mu}$ at the confidence level $1 - \alpha$:

$$\mathcal{K}_{\underline{\mu}} = \left\{ \underline{\mu} \in \mathbb{R}^p : (\underline{\mu} - \underline{\hat{\mu}}_{ML})^T \Sigma^{-1} (\underline{\mu} - \underline{\hat{\mu}}_{ML}) \leqslant \frac{1}{n} \chi^2_{p;1-\alpha} \right\},$$

with center $\underline{\hat{\mu}}_{ML}$ and semi-axes lengths $c_i = \sqrt{c\lambda_i}$, where $c = \chi^2_{p;1-\alpha}/n$ and λ_i are the eigenvalues of Σ; $i = 1, ..., p$.

Hotelling T^2

When Σ is **unknown**, then we replace Σ in $(*)$ by the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \underline{\hat{\mu}})(\underline{x}_i - \underline{\hat{\mu}})^T$$

und thus arrive at the so-called **Hotelling-T^2-statistic**:

$$T^2 = n\,(\underline{\hat{\mu}} - \underline{\mu})^T\, \widehat{\Sigma}^{-1} (\underline{\hat{\mu}} - \underline{\mu})\,.$$

This statistic, however, no longer follows a χ^2-distribution as in $(*)$.

Theorem (Distribution of Hotelling's T^2)

The Hotelling-T^2-statistic follows a Fisher-F-distribution:

$$\frac{n-p}{(n-1)p}\, T^2 = \frac{n}{n-1} \cdot \frac{n-p}{p}\, (\underline{\hat{\mu}} - \underline{\mu})^T\, \widehat{\Sigma}^{-1} (\underline{\hat{\mu}} - \underline{\mu}) \;\sim\; F_{p,n-p}$$

For a proof see G.F.A. Seber: Multivariate Observations, Wiley 2004.

Confidence region for $\underline{\mu}, \Sigma$ unknown

Thus, for unknown Σ, the confidence ellipsoid for $\underline{\mu}$ at the confidence level $1-\alpha$ is given by:

$$\mathcal{K}_{\underline{\mu}} = \left\{ \underline{\mu} \in \mathbb{R}^p : (\underline{\mu} - \underline{\hat{\mu}})^T \, \widehat{\Sigma}^{-1} (\underline{\mu} - \underline{\hat{\mu}}) \leq \frac{n-1}{n} \frac{p}{n-p} \, F_{p,n-p;1-\alpha} \right\}.$$

This is an ellipsoid with center $\underline{\hat{\mu}}$ as before, but with semi-axes lengths $\sqrt{c\lambda_i}$ where $\lambda_i = \lambda_i(\widehat{\Sigma})$ and $c = \frac{n-1}{n} \frac{p}{n-p} F_{p,n-p;1-\alpha}$.

Specialization: Hotelling-distribution in case of $p = 1$:

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2) \; ; i = 1, \ldots, n$$

$$\Longrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2; \quad \frac{(n-1)}{\sigma^2} \widehat{\sigma}^2 \sim \chi^2_{n-1}$$

After standardization of $\hat{\mu}$ and squaring the score we have

Hotelling distribution when $p = 1$

$$\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \Longrightarrow \left(\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \right)^2 = n \, \frac{(\hat{\mu} - \mu)^2}{\sigma^2} \sim \chi^2_1 \, .$$

Observing that $\widehat{\sigma}^2$ and $\hat{\mu}$ are stochastically independent, it follows that:

$$T^2 = \frac{\left(\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \right)^2 / 1}{\frac{(n-1)}{\sigma^2} \widehat{\sigma}^2 / (n-1)} = n \, \frac{(\hat{\mu} - \mu)^2}{\widehat{\sigma}^2} \sim F_{1,n-1}$$

This proves the above theorem in the special case $p = 1$. Finally,

$$T = \frac{\hat{\mu} - \mu}{\sqrt{\hat{\sigma}^2/n}} \sim t_{n-1}$$

from which the denotation T^2 and the relationship between the Student-t- and Fisher-F-distribution with first $df = 1$ become obvious.

Hotelling-T^2- test

Concluding, we consider the problem of **testing** the hypothesis

$$H_0 : \underline{\mu} = \underline{\mu}_0 \text{ vs. } H_1 : \underline{\mu} \neq \underline{\mu}_0,$$

where $\underline{\mu}_0 \in \mathbb{R}^p$ is some given (quality) standard. The corresponding Hotelling-T^2-test plays an important role in the area of statistical process control (SPC). We have used it extensively in our project cooperation with Infineon Austria AG for testing quality standards in several production units of semiconductor manufacturing (lithography, etching, burning processes, packaging, life and reliability testing,...)

$$H_0 : \underline{\mu} = \begin{bmatrix} \text{mean etch rate} \\ \text{mean operating temperature} \\ \vdots \\ \text{mean burning duration} \\ \text{mean voltage} \end{bmatrix} = \begin{bmatrix} 0.6\ \mu\text{m/min} \\ 80\ ^\circ\text{C} \\ \vdots \\ 20\ \text{min} \\ 64.3\ \text{mV} \end{bmatrix} = \underline{\mu}_0$$

Hotelling-T^2- test

As a test statistic for testing $H_0 : \underline{\mu} = \underline{\mu}_0$ vs. H_1 we use

$$T_0^2 = n\,(\underline{\hat{\mu}} - \underline{\mu}_0)^T\,\widehat{\Sigma}^{-1}(\underline{\hat{\mu}} - \underline{\mu}_0) \overset{H_0}{\sim} \frac{(n-1)p}{n-p}\,F_{p,n-p}\,.$$

Decision: we reject H_0 if

$$T_0^2 = n\,(\underline{\hat{\mu}} - \underline{\mu}_0)^T\,\widehat{\Sigma}^{-1}(\underline{\hat{\mu}} - \underline{\mu}_0) > \frac{(n-1)p}{n-p}\,F_{p,n-p;1-\alpha}\,.$$

For applications in semiconductor manufacturing process control see

Sonja Muringer: Advanced Process Control based on Multivariate Control Charts and Dynamic Linear Models. PhD Dissertation, AAU Klagenfurt, 2007

3.4 Testing uncorrelatedness

Let $\underline{X}_i \overset{iid}{\sim} N_p(\underline{0}, \Sigma)$; $i = 1, ..., n$; and consider the sum

$$\underset{(p\times p)}{Y} = \sum_{i=1}^{n} \underline{X}_i \, \underline{X}_i^T$$

of n (random) matrices. What can be said about the distribution of Y?

Definition

The random matrix Y is said to be **Wishart-distributed** of dimension p with parameters n (degrees of freedom) and Σ, $Y \sim W_p(n, \Sigma)$, if it has the pdf

$$f_Y(A) \propto \begin{cases} |A|^{(n-p-1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}A)\right) & \text{, for } A \in PSD(p) \\ 0 \text{ , else} \end{cases}$$

Special case: $p = 1$, $\Sigma = (\sigma^2)$. Then it follows

Testing diagonality of Σ

$$X_i \overset{iid}{\sim} N(0, \sigma^2) \Rightarrow \frac{X_i}{\sigma} \overset{iid}{\sim} N(0,1); i = 1, ..., n$$

$$\Rightarrow Y = \sigma^{-2} \sum_{i=1}^{n} X_i^2 \sim \chi_n^2 = Ga\left(\frac{n}{2}, \frac{1}{2}\right)$$

$$\Rightarrow f_Y(y) \propto y^{n/2-1} \exp\left(-\frac{y}{2}\right) I_{(0,\infty)}(y).$$

Therefore, $W_p(n, \Sigma)$ represents the matrix-valued generalization of the χ_n^2-distribution. Accordingly, n is called the number of degrees of freedom (= number of independent summands).

Corollary 3.3: $Y \sim W_p(n, \Sigma) \Longrightarrow \mathbb{E}\, Y = n\, \Sigma$.
Proof :
$\mathbb{E}\, Y = \sum_{i=1}^{n} \mathbb{E}\left(\underline{X}_i \underline{X}_i^T\right) = \sum_{i=1}^{n} \left[\mathrm{Cov}(\underline{X}_i) + (\mathbb{E}\underline{X}_i)(\mathbb{E}\underline{X}_i)^T\right] = n\, \Sigma$.
observing that $\mathbb{E}X_i = \underline{0},\ \mathrm{Cov}(X_i) = \Sigma$ for all $i = 1, ..., n$.

Testing diagonality of Σ

The Wishart distribution finds an important application in testing the uncorrelatedness of the components of $\underline{X} = (X_1, ..., X_p)^T$, which means testing the hypothesis

$$H_0 : \Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2) \Longleftrightarrow H_0 : P = I_p.$$

This hypothesis can be tested by means of the eigenvalues $\lambda_j(R)$ of the sample correlation matrix

$$R = \frac{1}{n-1} X_{st}^T X_{st}$$

and using the test statistic

$$T_R = -\left(n - 2 - \frac{2p-1}{6}\right) \sum_{j=1}^{p} \log(\lambda_j(R)) \overset{as}{\sim} \chi_k^2$$

where $k = \frac{p(p-1)}{2}$.

4. Multivariate Analysis of Variance (MANOVA)

Available: $p \geqslant 1$ variables (features), $g \geqslant 2$ groups of individuals (objects). In each of the groups $k = 1, \ldots, g$ we observe realizations of all of the p features:

$$\underline{x}_{ik} = \begin{pmatrix} x_{i1k} \\ x_{i2k} \\ \vdots \\ x_{ipk} \end{pmatrix} \quad k = 1, \ldots, g; \; i = 1, \ldots, n_k$$

denotes the i-th observation vector of the k-th group; $x_{i1k} \ldots$ denotes the measurement of the 1st feature of the i-th object (person) in the k-th group.

The total number of observation vectors is then given by $n = n_1 + n_2 + \ldots + n_g$, where n_k stands for the number of observation vectors in the k-th group; $k = 1, ..., g$.

One-way MANOVA

In this chapter, we will restrict ourselves to one-way- classification.

Problem: Are there statistically significant group differences w.r.t. the entity of $p \geqslant 1$ features?

4.1 Model of one-way MANOVA

(M) $\underline{X}_{ik} = \underline{\mu}_k + \underline{\epsilon}_{ik}$ with $\underline{X}_{ik} \sim N_p(\underline{\mu}_k, \Sigma)$ (*Variance homogeneity*).

Equivalently, the error terms satisfy

$$\underline{\epsilon}_{ik} \sim N(\underline{0}, \Sigma) \text{ and } \mathbb{E}(\underline{\epsilon}_{ik}) = \underline{0},\ \mathrm{Cov}(\underline{\epsilon}_{ik}) = \Sigma;\ \ k = 1, \ldots, g;\ \ i = 1, \ldots, n_k$$

We are interested in deviations from the **grand mean** $\underline{\mu}$, i.e.

$$\underline{\mu}_k = \underline{\mu} + \underline{\alpha}_k; k = 1, ..., g$$

where $\underline{\alpha}_k$ represents the **group effect** of the k-th group.

Therefore, we test the hypothesis

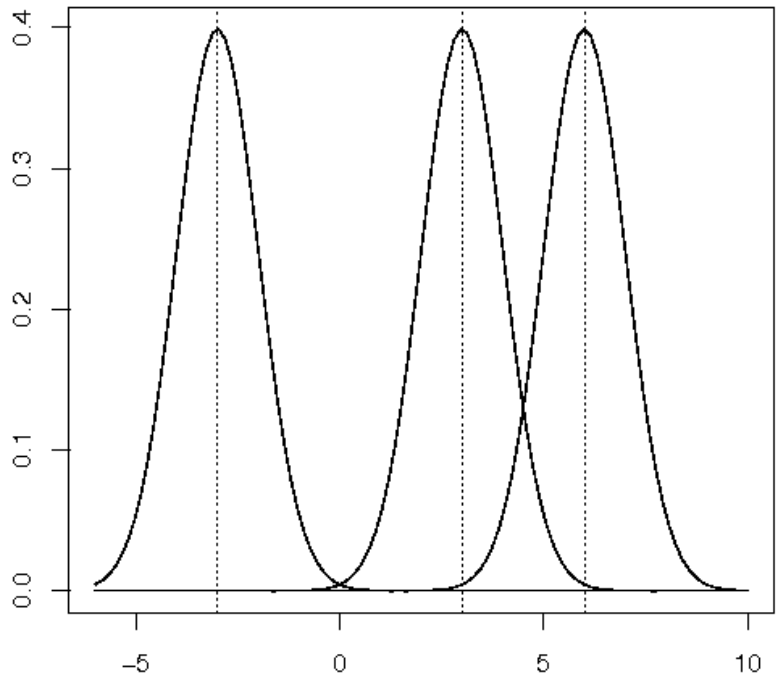
$$H_0 : \underline{\mu}_1 = \underline{\mu}_2 = \ldots = \underline{\mu}_g (= \underline{\mu}) \Longleftrightarrow \tilde{H}_0 : \underline{\alpha}_1 = \underline{\alpha}_2 = \ldots = \underline{\alpha}_g = \underline{0}$$

against the alternative hypothesis

$$H_A : \underline{\mu}_i \neq \underline{\mu}_j \text{ for at least one pair } (i, j) \; ; \; i \neq j.$$

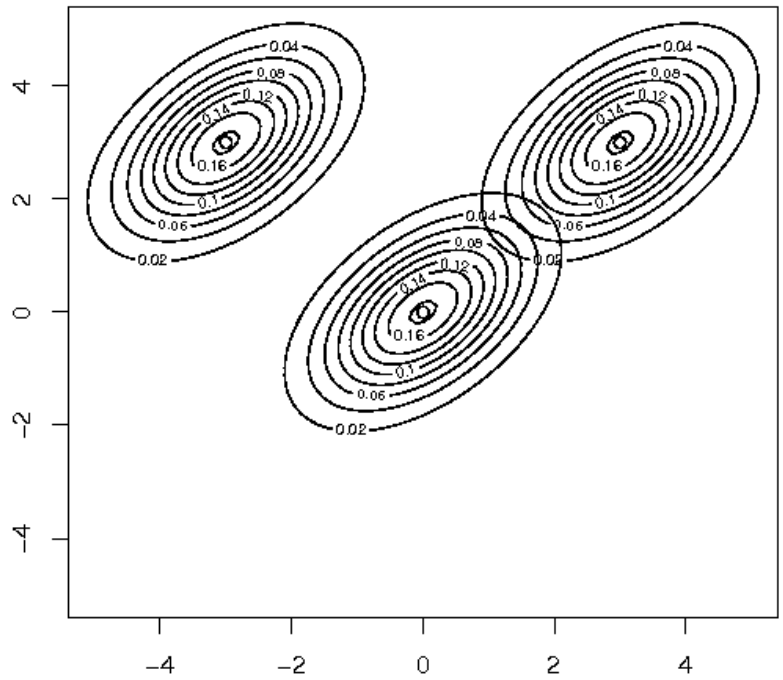
For a graphical illustration, we consider the special cases where

1. $p = 1$, $g = 3$ (three Gaussian densities of equal height, but shifted means)



One-way MANOVA

1. $p = 2$, $g = 3$ (Elliptical contours of equal orientation, but shifted means)



We may rewrite the model (M) of one-way MANOVA as follows

Model of one-way MANOVA

$$\underline{X}_{ik} = \begin{pmatrix} X_{i1k} \\ X_{i2k} \\ \vdots \\ X_{ipk} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \mu_{1k} - \mu_1 \\ \mu_{2k} - \mu_2 \\ \vdots \\ \mu_{pk} - \mu_p \end{pmatrix} + \begin{pmatrix} X_{i1k} - \mu_{1k} \\ X_{i2k} - \mu_{2k} \\ \vdots \\ X_{ipk} - \mu_{pk} \end{pmatrix} = \underline{\mu} + \underline{\alpha}_k + \underline{\epsilon}_{ik}$$

The moment estimators for the group means, grand mean and group effects, respectively, are thus given by

$$\hat{\underline{\mu}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \underline{X}_{ik} \; ; \quad \hat{\underline{\mu}} = \frac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{n_k} \underline{X}_{ik} = \frac{1}{n} \sum_{k=1}^{g} n_k \hat{\underline{\mu}}_k \; ; \quad \hat{\underline{\alpha}}_k = \hat{\underline{\mu}}_k - \hat{\underline{\mu}} \; .$$

To guarantee identifiability, we need a **reparameterization** condition

$$\sum_{k=1}^{g} n_k \underline{\alpha}_k = \underline{0} \; .$$

Least squares estimates

Theorem (MANOVA LSE's and MLE's)

Under the reparameterization condition, the moment estimators given above are least squares estimators for $\underline{\mu}, \underline{\alpha}_1, ..., \underline{\alpha}_g$, *i.e.*

$$\min \sum_{k=1}^{g} \sum_{i=1}^{n_k} \underline{\epsilon}_{ik}^T \underline{\epsilon}_{ik} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\underline{x}_{ik} - \hat{\underline{\mu}} - \hat{\underline{\alpha}}_k)^T (\underline{x}_{ik} - \hat{\underline{\mu}} - \hat{\underline{\alpha}}_k)$$

Under the additional assumption $\underline{\epsilon}_{ik} \sim N(\underline{0}, \Sigma)$, *they are also maximum likelihood estimators for* $\underline{\mu}, \underline{\alpha}_1, ..., \underline{\alpha}_g$.

The sample covariance matrix of the observation vectors of the k-th group is given by

$$\widehat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\underline{x}_{ik} - \hat{\underline{\mu}}_k)(\underline{x}_{ik} - \hat{\underline{\mu}}_k)^T \; ; \; k = 1, ..., g.$$

Least squares estimates

The overall covariance matrix Σ is then estimated as a pooled sample covariance matrix:

$$\widehat{\Sigma} = \frac{(n_1-1)\widehat{\Sigma}_1 + (n_2-1)\widehat{\Sigma}_2 + \ldots + (n_g-1)\widehat{\Sigma}_g}{(n_1-1)+(n_2-1)+\ldots+(n_g-1)} = \frac{1}{n-g}\sum_{k=1}^{g}(n_k-1)\widehat{\Sigma}_k$$

Testing the null hypothesis H_0 is then based on the decomposition of the total sum of squares matrix into within and between group of squares:

$$T = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{x}_{ik}-\underline{\hat{\mu}})(\underline{x}_{ik}-\underline{\hat{\mu}})^T \qquad \text{total sum of squares matrix}$$

$$W = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{x}_{ik}-\underline{\hat{\mu}}_k)(\underline{x}_{ik}-\underline{\hat{\mu}}_k)^T \quad \text{within (group) sum of squares matrix}$$

$$B = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{\hat{\mu}}_k-\underline{\hat{\mu}})(\underline{\hat{\mu}}_k-\underline{\hat{\mu}})^T \quad \text{between (group) sum of squares matrix}$$

Between and within group decomposition

Theorem (Orthogonal Decomposition of T)

The total sum of squares can be orthogonally decomposed as:

$$T = W + B.$$

Proof: Zero addition by $\underline{\hat{\mu}}_k$ yields

$$\begin{aligned} T &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{x}_{ik} - \underline{\hat{\mu}})(\underline{x}_{ik} - \underline{\hat{\mu}})^T \\ &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}\left[(\underline{x}_{ik} - \underline{\hat{\mu}}_k) + (\underline{\hat{\mu}}_k - \underline{\hat{\mu}})\right]\left[(\underline{x}_{ik} - \underline{\hat{\mu}}_k) + (\underline{\hat{\mu}}_k - \underline{\hat{\mu}})\right]^T \\ &= W + B + \underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{x}_{ik} - \underline{\hat{\mu}}_k)(\underline{\hat{\mu}}_k - \underline{\hat{\mu}})^T}_{=:C} + \underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{\hat{\mu}}_k - \underline{\hat{\mu}})(\underline{x}_{ik} - \underline{\hat{\mu}}_k)^T}_{=:C^T} \end{aligned}$$

Orthogonal decomposition

We show that C is a matrix of zeros, which concludes the proof:

$$C = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{x}_{ik} - \hat{\underline{\mu}}_k)(\hat{\underline{\mu}}_k - \hat{\underline{\mu}})^T = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\underline{x}_{ik} - \hat{\underline{\mu}}_k)\hat{\underline{\alpha}}_k^T$$

$$= \sum_{k=1}^{g}\sum_{i=1}^{n_k}\underline{x}_{ik}\hat{\underline{\alpha}}_k^T - \sum_{k=1}^{g}\sum_{i=1}^{n_k}\hat{\underline{\mu}}_k\hat{\underline{\alpha}}_k^T$$

$$= \sum_{k=1}^{g} n_k\hat{\underline{\mu}}_k\hat{\underline{\alpha}}_k^T - \sum_{k=1}^{g} n_k\hat{\underline{\mu}}_k\hat{\underline{\alpha}}_k^T = \mathbb{O}.$$

Relationship between $\widehat{\Sigma}_k$, $\widehat{\Sigma}$ and W:

$$\widehat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(\underline{x}_{ik} - \hat{\underline{\mu}}_k)(\underline{x}_{ik} - \hat{\underline{\mu}}_k)^T$$

$$\widehat{\Sigma} = \frac{(n_1 - 1)\widehat{\Sigma}_1 + \ldots + (n_g - 1)\widehat{\Sigma}_g}{(n_1 - 1) + \ldots + (n_g - 1)} = \frac{1}{n - g}\sum_{k=1}^{g}(n_k - 1)\widehat{\Sigma}_k = \frac{1}{n - g}W$$

Two-sample-test

We have $n-g$ independent summands in W, $g-1$ independent summands in B
$\Longrightarrow df_W + df_B = df_T = n-1$ independent summands in T.

Before we proceed, let us briefly recall the siuation in the

Special case: $g = 2,\ p = 1$ (univariate two-sample-test problem)

$$X_{i1} \sim N(\mu_1, \sigma^2); i = 1, \ldots, n_1 \text{ and } X_{j2} \sim N(\mu_2, \sigma^2); j = 1, \ldots, n_2$$

where we test $H_0 : \mu_1 = \mu_2$ on the basis of

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{i1} \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) \text{ and } \hat{\mu}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{j2} \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right).$$

Starting from the estimated difference

$$\hat{\mu}_1 - \hat{\mu}_2 \overset{H_0}{\sim} N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right),$$

Two-sample-test

it follows that

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \overset{H_0}{\sim} N(0,1) \; .$$

Replacing σ^2 by its (pooled) estimate

$$\hat{\sigma}^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n - 2} \quad \text{where } n = n_1 + n_2 \; ,$$

T then becomes Student-t-distributed:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \Longrightarrow T = \frac{\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} / (n-2)}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n-2}$$

The null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected whenever $|T| > t_{n-2;1-\frac{\alpha}{2}}$.

4.2 Test Statistics

a) Wilks-Λ-distribution: Recall that in case of

$$\underline{Y}_i \overset{iid}{\sim} N_p(\underline{0}, \Sigma);\ i = 1, \ldots, n$$

we have

$$Y = \underline{Y}_1\underline{Y}_1^T + \underline{Y}_2\underline{Y}_2^T + \ldots + \underline{Y}_n\underline{Y}_n^T \sim W_p(n, \Sigma) \text{ and } \mathbb{E}(Y) = n\Sigma.$$

Definition (Wilks-Λ-Distribution)

The random variable Z is said to be Wilks-Λ_p-distributed with parameters m and k, if Z can be represented in the form

$$Z = \frac{det(Q_1)}{det(Q_1+Q_2)} \text{ where } Q_1 \sim W_p(m, I_p),\ Q_2 \sim W_p(k, I_p)$$

and Q_1 is stochastically independent from Q_2.

Briefly, we then write $Z \sim \Lambda_p(m, k)$.

Wilks- Λ

Illustration for p=1:

$Z = \frac{Q_1}{Q_1+Q_2}$ where $Q_1 \sim X^2_m$ and $Q_2 \sim X^2_k$ are stoch. independent.

$$\Longrightarrow \frac{Q_1/m}{(Q_1+Q_2)/(m+k)} = \frac{m+k}{m}\frac{Q_1}{Q_1+Q_2} \sim F_{m,m+k}\ .$$

Wilks-Lambda-distribution can thus be considered as a multivariate generalization of the Fisher-F-distribution.

b) Wilks-Λ-Statistic: Assume, again, that

$$\underline{X}_{ik} \sim \mathrm{N}(\underline{\mu}_k, \Sigma);\ i = 1, \ldots, n_k;\ k = 1, \ldots, g.$$

The test of

$$H_0 : \underline{\mu}_1 = \ldots = \underline{\mu}_g = \underline{\mu}$$

is based on the Liklihood-Ratio-Test statistics:

Wilks- Λ

$$T_{LR} = \frac{\sup\limits_{\underline{\mu}_1 = \ldots = \underline{\mu}_g, \Sigma} L(\underline{\mu}_1, \ldots, \underline{\mu}_g, \Sigma; x_{11}, \ldots, x_{n_g g})}{\sup\limits_{\underline{\mu}_1, \ldots, \underline{\mu}_g, \Sigma} L(\underline{\mu}_1, \ldots, \underline{\mu}_g, \Sigma; x_{11}, \ldots, x_{n_g g})} = \ldots = \frac{|W|}{|W+B|} = \frac{|W|}{|T|}$$

where the maximum is over all $\underline{\mu}_1, \ldots, \underline{\mu}_g \in \mathbb{R}^p, \Sigma \in PSD(p)$.

From the orthogonality of the decomposition $T = W + B$ we may then conclude that the LR-statistics is Wilks-Λ-distributed.

Corollary 4.1: $T_{LR} = \frac{|W|}{|W+B|} \sim \Lambda_p(n-g, g-1)$.

c) Rao's-F-statistic: For large n, the Λ-distribution can be approximated by the Fisher-F-distribution:

$$F = \frac{m_2}{m_1}\left(T_{LR}^{-1/s} - 1\right) \overset{as}{\sim} F_{m_1, m_2} \quad \text{where} \quad s = \frac{\sqrt{p^2(g-1)^2 - 4}}{p^2 + (g-1)^2 - 5}$$

Rao's- F

and, further,

$$m_1 = p(g-1),\ m_2 = 1 + \frac{s}{2}(2n-p-g-2) - \frac{m_1}{2},\ n = n_1 + \ldots + n_g.$$

For the special case $p = 1$ we have: $s = 1, m_1 = g-1, m_2 = n-g$, i.e.

$$F = \frac{n-g}{g-1}\left(\frac{W+B}{W} - 1\right) = \frac{B/(g-1)}{W/(n-g)} \sim F_{g-1,n-g},$$

an exact F-distribution. Rao's-F-approximation results in an exact F-distribution in further situations, too:

1. $p = 2$ (assuming H_0 is true), $\forall g \geqslant 1$
2. $\forall p \geqslant 1$ (assuming H_0 is true), if $g = 2 \vee g = 3$

In all other cases, we have an approximate F-distribution.

Decision: H_0 is rejected whenever $F > F_{m_1,m_2;1-\alpha}$.

Comparing two groups

In case that H_0 is rejected, we have to proceed with pairwise comparisons. Then, however, α must be modified, e.g. we replace α by $\alpha/\binom{g}{2}$ (Bonferroni-correction).

4.3 Comparing two groups

ANOVA-case, $p = 1$: $X_{i1} \sim N(\mu_1, \sigma^2);\ i = 1, \ldots, n_1$

$$X_{j2} \sim N(\mu_2, \sigma^2);\ j = 1, \ldots, n_2.$$

We test $H_0 : \mu_1 = \mu_2$ using the Student-t-statistic

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\ \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}} \ \sim\ t_{n-2}$$

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}}\right)^2 \ \sim\ F_{1,n-2}$$

T^2 coincides with Rao's-F-statistic for $p = 1,\ g = 2$.

Comparing two groups

MANOVA-case, $p > 1$: $\underline{X}_{i1} \sim N_p(\underline{\mu}_1, \Sigma); \quad i = 1, \ldots, n_1$

$$\underline{X}_{j2} \sim N_p(\underline{\mu}_2, \Sigma); \quad j = 1, \ldots, n_2.$$

$H_0 : \underline{\mu}_1 = \underline{\mu}_2$ is tested on the basis of

$$\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2 \overset{H_0}{\sim} N_p\left(\underline{0}, (\frac{1}{n_1} + \frac{1}{n_2})\Sigma\right).$$

$$\Rightarrow (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right]^{-1} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2) = d_M^2(\hat{\underline{\mu}}_1, \hat{\underline{\mu}}_2) \overset{H_0}{\sim} \chi_p^2$$

Replacing Σ by its estimate $\widehat{\Sigma} = \frac{(n_1-1)\widehat{\Sigma}_1 + (n_2-1)\widehat{\Sigma}_2}{n-2}$, leads us to Hotelling's

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)^T \widehat{\Sigma}^{-1} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)$$

for testing $H_0 : \underline{\mu}_1 = \underline{\mu}_2$.

Corollary 4.2: In the special case of $g = 2$ groups, the hypothesis of the equality of the group means $H_0 : \underline{\mu}_1 = \underline{\mu}_2$ is tested by means of Hotelling's T^2-statistic, for which it holds:

$$\frac{n-p-1}{p(n-2)}\, T^2 \;\sim\; F_{p,n-p-1} \quad \forall p \geqslant 1 \quad \forall n > p+1 \;.$$

4.4 Implementation in R

For testing the equality of the group means we use the R-function "manova". As an example, consider, again, the iris data:

> data(iris)

> dimnames(iris)[[2]]

[1]" *Sepal.Length*"" *Sepal.Width*"" *Petal.Length*"" *Petal.Width*"" *Species*"

> model = manova(cbind(Sepal.Length, Sepal.Width,

+ Petal.Length, Petal.Width) ∼ Species, iris)

R- Implementation

> summary(model)

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
Species	2	1.1919	53.466	8	290	< 2.2e-16 ***
Residuals	147					

> summary(model, test="Wilks")

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
Species	2	0.0234	199.15	8	288	< 2.2e-16 ***
Residuals	147					

The default test is **Pillai**'s test, which uses the test statistic $T_P = tr(T^{-1}B)$, where $T = W + B$ and T, W, B are the total, within and between sum of squares matrices as defined before. There are two further test options:

Hotelling-Lawley test statistic: $T_{HL} = tr(W^{-1}B)$ and

Roy's test statistic: $T_{Roy} = \lambda_{max}(W^{-1}B)$.

R- Implementation

Wilk's Λ, T_P, T_{HL} *and* T_{Roy} are equivalent test statistics.

Univariate ANOVA tables for the single components can be obtained as follows:

$>$ summary.aov(model)

Response Sepal.Length :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	63.212	31.606	119.26	< 2.2e-16 ***
Residuals	147	38.956	0.265		

Response Sepal.Width :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	11.345	5.6725	49.16	< 2.2e-16 ***
Residuals	147	16.962	0.1154		

Univariate ANOVA tables

Response Petal.Length :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	437.10	218.551	1180.2	< 2.2e-16 ***
Residuals	147	27.22	0.185		

Response Petal.Width :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species	2	80.413	40.207	960.01	< 2.2e-16 ***
Residuals	147	6.157	0.042		

5. Cluster analysis

Goal: Classification/Grouping of objects (individuals) into homogeneous units (clusters) such that objects within a group are similar to each other and objects belonging to different groups exhibit significant differences.

Applications: are wide-ranging and include, for example

- classification tasks in biology, ecology, phytology, zoology, anthropology, . . .
- pattern recognition in linguistics, psychology, marketing, computer vision, artificial intelligence, ...

Classification types:

- supervised learning: object classes are known (labeled objects), learning on the basis of known objects
- unsupervised learning: classification of unlabeled objects (no prior knowledge available).

Mixed type: semi-supervised learning, where labeled and unlabeled objects occur jointly.

The key to classification is to define a measure for the similarity/dissimilarity of objects.

5.1 Distance and similarity of objects

We start again from the data matrix

$$\underset{n\times p}{X} = \begin{pmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Denote $\mathscr{X} = \{\underline{x}_1, \dots, \underline{x}_n\}$. The mapping $d : \mathscr{X} \times \mathscr{X} \to [0, \infty)$ is called a **distance measure** for observation vectors if it holds

1) $d(\underline{x}_i, \underline{x}_j) \geqslant 0 \quad \wedge \quad d(\underline{x}_i, \underline{x}_i) = 0 \qquad \forall i, j = 1, \dots, n$
2) $d(\underline{x}_i, \underline{x}_j) = d(\underline{x}_j, \underline{x}_i) \qquad \forall i, j = 1, \dots, n.$

Similarity measure

If it holds, additionally:

3) $d(\underline{x}_i, \underline{x}_j) = 0 \Longrightarrow i = j$

4) $d(\underline{x}_i, \underline{x}_k) \leqslant d(\underline{x}_i, \underline{x}_j) + d(\underline{x}_j, \underline{x}_k) \qquad \forall i, j, k = 1, \ldots, n$

then $d(\cdot, \cdot)$ defines a metric.

Accordingly, a mapping $s : \mathscr{X} \times \mathscr{X} \to [0, 1]$ is called a **similarity measure** for objects if it holds

1. $0 \leqslant s(\underline{x}_i, \underline{x}_j) \leqslant 1$
2. $s(\underline{x}_i, \underline{x}_j) = s(\underline{x}_j, \underline{x}_i)$
3. $s(\underline{x}_i, \underline{x}_i) = 1 \qquad \forall i, j = 1, \ldots, n.$

Remark: Distance and similarity are inversely related, a small distance between objects means means a high similarity. A distance measure can be easily transformed into a similarity measure, e.g.:

- $s(\cdot, \cdot) = 1 - d(\cdot, \cdot)/d_0;\ d_0 = max\left\{d(\underline{x}_i, \underline{x}_j) : \underline{x}_i, \underline{x}_j \in \mathscr{X}\right\}$
- $s(\cdot, \cdot) = exp(-d(\cdot, \cdot))$.

Distance matrix

The basis of any cluster analysis is the so-called **distance matrix**. For a given distance measure, this matrix is defined element-wise by the distances between all observed objects:

$$\underset{n\times n}{D} = (d_{ij})_{i,j=1,\ldots,n} = \begin{pmatrix} 0 & d_{12} & \ldots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \ldots & 0 \end{pmatrix}$$

Clearly, D is symmetric, i.e. we need to store only the upper off-diagonal elements $d_{ij} = d(\underline{x}_i, \underline{x}_j), i < j = 1, \ldots, n$.

Example (Distance measures for metric variables)

- Euklidean distance:

$$d_E(\underline{x}_i, \underline{x}_j) = \sqrt{(\underline{x}_i - \underline{x}_j)^T(\underline{x}_i - \underline{x}_j)} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Distance measures

Example (cont'd)

- Minkowski distance: $d_r(\underline{x}_i, \underline{x}_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}}; \quad r \geqslant 1$

  for $r = 1$: Manhattan distance (l_1-distance)

  for $r = 2$: Eukidean distance

  for $r = \infty$: maximum norm

  $$d_\infty(\underline{x}_i, \underline{x}_j) = \max\left\{ |x_{ik} - x_{jk}| : k = 1, ..., p \right\}$$

- Mahalanobis-distance: $d_M^2(\underline{x}_i, \underline{x}_j) = (\underline{x}_i - \underline{x}_j)^T \widehat{\Sigma}^{-1} (\underline{x}_i - \underline{x}_j)$

  Advantage: it takes account of (weighted) variances and correlations, and it is invariant w.r.t. scale transformations.

  Disadvantage: homogeneity assumption (equal mean vectors and covariance matrices).

5.2 Distance measures for discrete variables

We start with binary variables, which serve as indicators for the presence/non-presence of features. For a **binary variable** X we have:

$$X = \begin{cases} 1 & \text{, if the feature is present} \\ 0 & \text{, if the feature is not present} \end{cases}$$

As an example, consider the variable

$$X = \text{smoking habit} = \begin{cases} 1 & \text{, person is a smoker} \\ 0 & \text{, person does not smoke} \end{cases}$$

Observation vectors whose components are all binary may look like:
$\underline{x}_i = (1, 0, 1, 1, 1, 1, 0, 0)$
$\underline{x}_j = (0, 1, 1, 0, 0, 1, 1, 0)$
How to measure their distance?

Discrete distance measures

Denotations:

n_i^1 := number of features only occuring in the ith object;

n_{ij}^1 := number of features occuring both in objects i and j;

n_{ij}^0 := number of features occuring in none of the objects i and j.

Common distance measures in case of $0/1$ coding include:

- Hamming-distance : $d_H(\underline{x}_i, \underline{x}_j) = d_1(\underline{x}_i, \underline{x}_j) = n_i^1 + n_j^1$
- Tanimoto-distance : $d_T(\underline{x}_i, \underline{x}_j) = d_1(\underline{x}_i, \underline{x}_j)/(n_i^1 + n_j^1 + n_{ij}^1)$
- simple-matching : $d_{sm}(\underline{x}_i, \underline{x}_j) = d_1(\underline{x}_i, \underline{x}_j)/p$,
  where $p = \dim(\underline{x}_i) = \dim(\underline{x}_j)$.

In the above example, we have: $n_i^1 = 3, n_j^1 = 2, n_{ij}^1 = 2, n_{ij}^0 = 1$,
and thus: $d_1(\underline{x}_i, \underline{x}_j) = 5, d_T(\underline{x}_i, \underline{x}_j) = 5/7, d_{sm}(\underline{x}_i, \underline{x}_j) = 5/8$.

Nominal variables

We now consider discrete (categorial) variables having more than two categories. The variable is called an **ordered variable** if these categories can be ordered in a natural way (e.g. gradings, education level, ...), otherwise it is called a **nominal variable** (e.g. hair color, gender, profession,...).

Let $\underline{X} = (X_1, X_2, ..., X_p)^T$ be a vector of categorial variables X_i with k_i categories each; $i = 1, ..., p$. Then $\underline{X}$ is coded as a binary vector of length $k = k_1 + k_2 + \ldots + k_p$ as follows.

- Nominal variables: Put a "1" exactly at the position at which the category is present, all other positions are filled with "0".
  For example, let $\underline{X} = (X_1, X_2, X_3)$; $p = 3$, where
  X_1 has 3 categories: $k_1 = 3$,
  X_2 has 6 categories: $k_2 = 6$,
  X_3 has 4 categories: $k_3 = 4$.

  Then the observation $\underline{x} = (3, 4, 1)$ is coded as:
  $\underline{x}_{\text{bin}} = (0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0)$.

Ordinal variables

- Ordinal variables: For ordered categories, we do not only put a "1"at the position at which the category is present, but at all preceding positions as well.

  For example, let, again, $\underline{X} = (X_1, X_2, X_3);\ p = 3$, where now
  X_1 has 3 ordered categories: $k_1 = 3$,
  X_2 has 6 ordered categories: $k_2 = 6$,
  X_3 has 4 ordered categories: $k_3 = 4$.

  Then the observation $\underline{x} = (3, 4, 1)$ is coded as:

  $\underline{x}_{\text{bin}} = (1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0)$.

R-Terminology:

metric variable:	numeric	(as.numeric)
nominal (non-ordered) variable:	factor	(as.factor)
ordinal (odered) variable:	ordered	(as.ordered)

5.3 Distance matrices for mixed-type variables

In practical applications, we often have a mix of variables (metric+nominal+ordinal varibles).
In these cases, the distance matrix is computed in the following way:

- Sorting the columns of the data matrix: $\underset{(n,p)}{X} = (\underset{(n,p_N)}{X_N} \mid \underset{(n,p_O)}{X_O} \mid \underset{(n,p_M)}{X_M})$

  where $p_N + p_O + p_M = p$.
  Thus, in the sorted data matrix we first see the realizations of the nominal variables, then the realizations of the ordinal variables and, finally, the realizations of the metric variables.

- Separate computation of the distance matrices $\underset{(n,n)}{D_N}, \underset{(n,n)}{D_O}, \underset{(n,n)}{D_M}$

- Weighted composition of the separate distance matrices to an overall distance matrix:

$$\underset{(n,n)}{D} = \frac{p_N}{p} D_N + \frac{p_O}{p} D_O + \frac{p_M}{p} D_M.$$

Distance between clusters

Computation of distance matrices in *R*:

- Distance matrix for purely metric variables ($p_M = p, p_N = p_O = 0$):
  $>$ dist(X, method="...", ...)
- Distance matrix for mixed variables:
  $>$ daisy(X, ...)

5.4 Cluster algorithms and distances between clusters

The formation of groups proceeds on the basis of the distance matrix *D*. We distinguish between hierarchical and non-hierarchical clustering methods. Hierarchical methods are structure-preserving, whereas non-hierarchical methods allow the creation of new structures during the clustering process. Here, "structure-preserving" means that objects which have been assigned to a common cluster at some stage, cannot be separated at later stages of the clustering process.
In case of hierarchical clustering methods, in turn, we can distinguish between agglomerative and divisive methods.

Agglomerative clustering

For agglomerative methods, the formation of clusters starts with single objects which are then fused to clusters, whereas divisive methods start with a single cluster containing all objects, which are then splitted into different (homogeneous) groups.

Algorithm for agglomerative clustering:

Start: each object i ($i = 1, \ldots, n$) is considered as a single cluster, i.e. $C_1 = \{\underline{x}_1\}, C_2 = \{\underline{x}_2\}, \ldots, C_n = \{\underline{x}_n\}$. Then

1. Determine pair(s) (i_0, j_0) such that: $\min\limits_{\substack{i,j=1,\ldots,n \\ i \neq j}} (d_{ij}) = d_{i_0 j_0}$
2. Update the distance matrix: objects i_0, j_0 are arranged in a new, joint cluster $\{\underline{x}_{i_0}, \underline{x}_{j_0}\}$ and then the distances of the new cluster(s) to the remaining clusters are computed
   $\Longrightarrow$ new distance matrix of dimension $k \leq n - 1$.
3. Stop if $k = 1$; otherwise continue with step 1, replacing n by k.

Cluster distance

How to compute the distances between clusters in step 2?

Denotation: $d_{AB}(\underline{x}_i, \underline{x}_j) :=$ distance between the objects $\underline{x}_i \in$ cluster A and $\underline{x}_j \in$ cluster B.

- SL : single linkage (nearest neighbour)

$$d_{AB}^{(SL)} = \min_{\underline{x}_i \in A, \underline{x}_j \in B} d_{AB}(\underline{x}_i, \underline{x}_j)$$

- CL : complete linkage (farthest neighbour)

$$d_{AB}^{(CL)} = \max_{\underline{x}_i \in A, \underline{x}_j \in B} d_{AB}(\underline{x}_i, \underline{x}_j)$$

- AL : average linkage

$$d_{AB}^{(AL)} = \frac{1}{|A| \cdot |B|} \sum_{\underline{x}_i \in A, \underline{x}_j \in B} d_{AB}(\underline{x}_i, \underline{x}_j).$$

Numerical illustration

Average Linkage clustering; $n = 5$ objects

Assume a distance matrix $D = \begin{pmatrix} 0 & 2 & 6 & 10 & 9 \\ 2 & 0 & 5 & 9 & 8 \\ 6 & 5 & 0 & 4 & 5 \\ 10 & 9 & 4 & 0 & 3 \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix}$

(1) $\min_{i>j \in \{1,\ldots,5\}} (d_{ij}) = d_{12} = 2 \Longrightarrow i_0 = 1,\ j_0 = 2$

$\Longrightarrow A = \{1,2\}, \{3\}, \{4\}, \{5\}$

(2) $d_{A,\{3\}} = \frac{1}{2}(6+5) = 5.5,\ d_{A,\{4\}} = \frac{1}{2}(10+9) = 9.5,$

$d_{A,\{5\}} = \frac{1}{2}(9+8) = 8.5$

$\Longrightarrow D^{(1)} := \begin{pmatrix} 0 & 5.5 & 9.5 & 8.5 \\ 5.5 & 0 & 4 & 5 \\ 9.5 & 4 & 0 & 3 \\ 8.5 & 5 & 3 & 0 \end{pmatrix}$

Numerical illustration

(3) $\min(d_{ij}^{(1)}) = d_{45} = 3 \Longrightarrow i_1 = 4,\ j_1 = 5$

$\Longrightarrow A = \{1,2\}, \{3\}, C = \{4,5\}$

(4) $d_{A,\{3\}} = \frac{1}{2}(6+5) = 5.5,\ d_{AC} = \frac{1}{4}(10+9+9+8) = 9,$
$d_{C,\{3\}} = \frac{1}{2}(4+5) = 4.5$

$$\Longrightarrow D^{(2)} := \begin{pmatrix} 0 & 5.5 & 9 \\ 5.5 & 0 & 4.5 \\ 9 & 4.5 & 0 \end{pmatrix}$$

(5) $\min(d_{ij}^{(2)}) = d_{345} = 4.5 \Longrightarrow i_2 = 3,\ j_2 = 45$

$\Longrightarrow A = \{1,2\}, B = \{3,4,5\}$

(6) $d_{AB} = \frac{1}{6}(6+10+9+5+9+8) = 7.8\dot{3}$

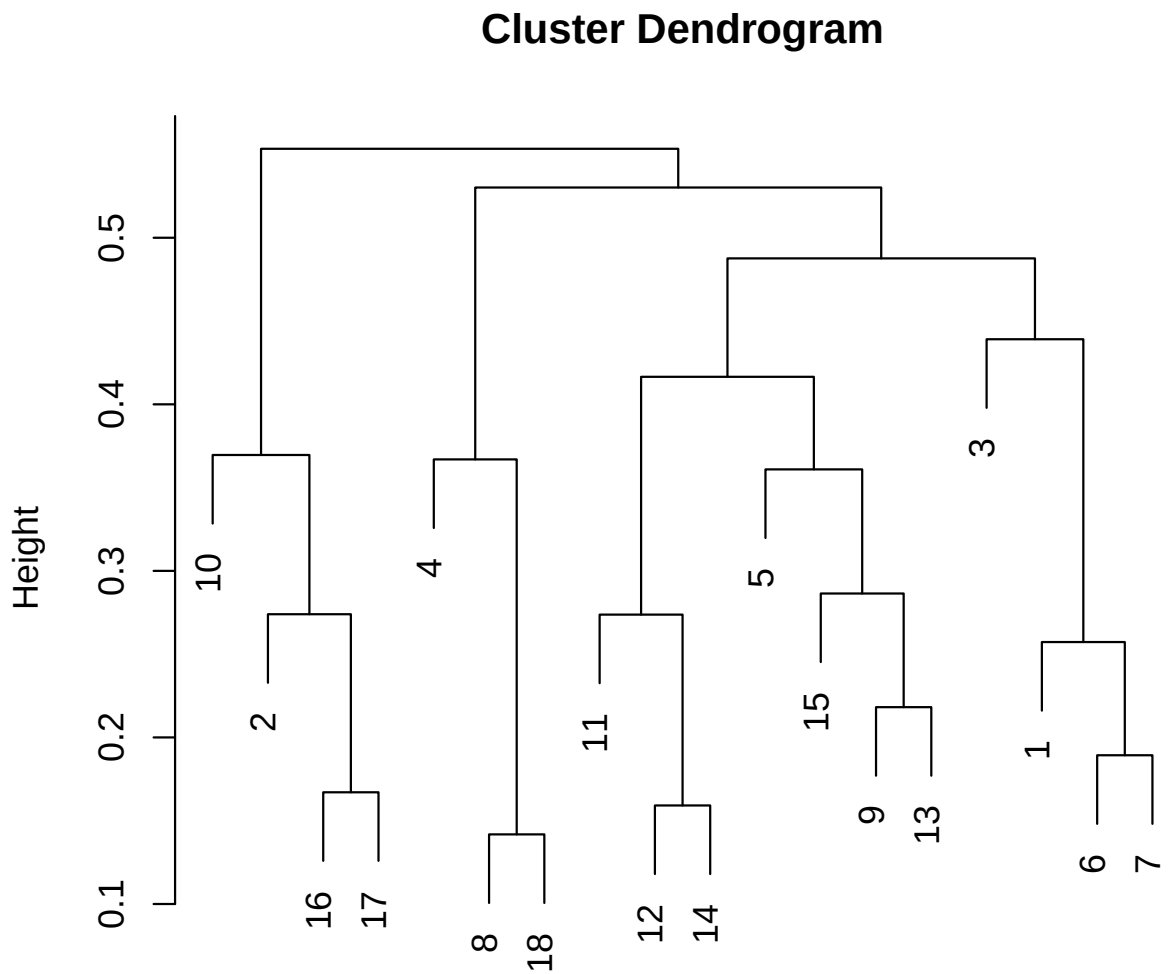
$$\Longrightarrow D^{(3)} := \begin{pmatrix} 0 & 7.8\dot{3} \\ 7.8\dot{3} & 0 \end{pmatrix}$$

Clustering of the "flower" data

```
> library(cluster)
> data(flower); > str(flower)
'data.frame': 18 obs. of 8 variables:
$ V1: Factor w/ 2 levels "0", "1": 1  2  1  1  1  1  1  1  2  2 ...
$ V2: Factor w/ 2 levels "0", "1": 2  1  2  1  2  2  1  1  2  2 ...
$ V3: Factor w/ 2 levels "0", "1": 2  1  1  2  1  1  1  2  1  1 ...
$ V4: Factor w/ 5 levels "1", "2", "3", "4",..: 4  2  3  4  5  4  4  2  3 ...
$ V5: Ord.factor w/ 3 levels "1"<"2"<"3": 3  1  3  2  2  3  3  2  1  2 ...
$ V6: Ord.factor w/ 18 levels "1"<"2"<"3"<"4"<..: 15  3  1  16  2  12 ...
$ V7: num 25  150  150  125  20  50  40  100  25  100 ...
$ V8: num 15  50  50  50  15  40  20  15  15  60 ...
> dflower=daisy(flower)          # distance matrix
> hdf=hclust(dflower, "ave")     # average linkage clustering
> plot(hdf)
```

Numerical illustration

The **dendrogram** provides a graphical visualization of the clustering process.



Specific clustering methods

k-means-clustering:

The number of clusters k that we have in mind must be prespecified. The algorithm then provides us with a clustering of objects into k clusters such that

$$\sum_{i=1}^{k} \sum_{\underline{x}_j \in C_i} ||\underline{x}_j - \underline{\hat{\mu}}_i||^2 \to \min ,$$

where C_i denotes the i-th cluster ($i = 1, \ldots, k$) with mean

$$\underline{\hat{\mu}}_i = \frac{1}{|C_i|} \sum_{\underline{x}_i \in C_i} \underline{x}_i .$$

fuzzy-c-means-clustering:

Again, we start with a prespecified number of clusters k. However, the assignment of the n given objects to a cluster is non-crisp (in the fuzzy sense). The degree of membership u_{ij} of object $\underline{x}_i$ to the cluster C_j is stored in the so-called membership matrix

fuzzy c-means

$$\underset{(n,k)}{U} = (u_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} .$$

The matrix U must satisfy the following conditions:

1. $0 \leqslant u_{ij} \leqslant 1, \quad i = 1, \ldots, n \,;\; j = 1, \ldots, k$
2. $0 < \sum_{i=1}^{n} u_{ij} < n, \quad \forall j = 1, \ldots, k$ (no "empty" clusters)
3. $\sum_{j=1}^{k} u_{ij} = 1, \quad \forall i = 1, \ldots, n$ (normalization).

The memberships are updated via an iterative procedure:

(1) Computing cluster centers:

$$\underline{c}_j = \sum_{i=1}^{n} u_{ij}^2 \underline{x}_i / \sum_{i=1}^{n} u_{ij}^2 \qquad j = 1, \ldots, k$$

(2) Computing object distances $\underline{x}_i$ to the cluster centers:

$$d(\underline{x}_i, \underline{c}_j) = ||\underline{x}_i - \underline{c}_j|| \qquad j = 1, \ldots, k$$

(3) Updating the memberships: inverse distance weighting

$$u_{ij}^{(new)} = d(\underline{x}_i, \underline{c}_j)^{-1} / \sum_{l=1}^{k} d(\underline{x}_i, \underline{c}_l)^{-1} \quad i = 1, \ldots, n; \; j = 1, \ldots, k$$

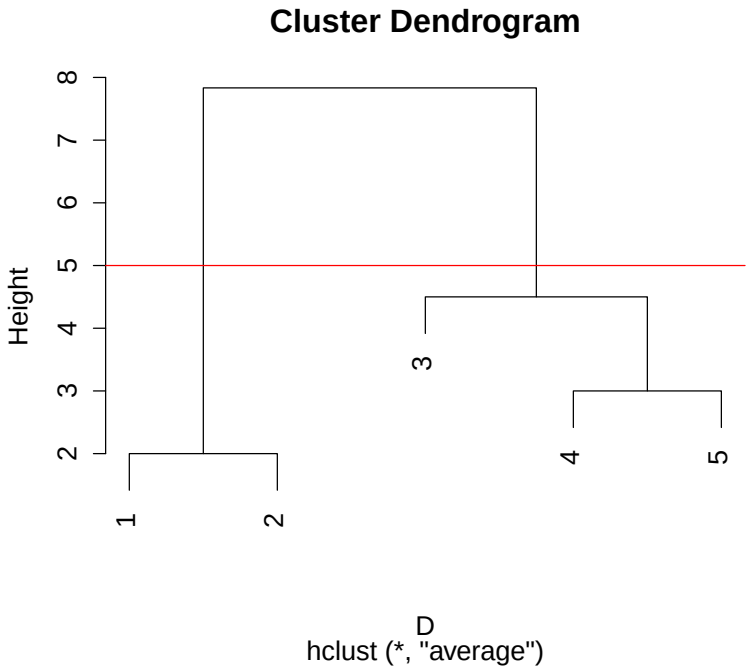
(4) Checking the convergence of U:

Stop if $||\Delta U|| < \epsilon$; otherwise return to (1) and continue.

5.5 Remarks on the goodness of clustering:

How to choose the number of clusters?
Commonly used cluster algorithms require that the number of clusters is determined in advance. From the dendrogram we can see at what distances the objects are combined to clusters.
This graphical visualization can be used for finding a suitable number of clusters, just by choosing an appropriate cut-off level.
An important tool that aids in the choice of such a level is the so-called cophenetic correlation.
For example, looking at the previous (toy) distance matrix

Procrustes distances



and cutting off at level $h = 5$ we obtain $k = 2$ clusters and "procrustes" distances

$$\hat{D} = \begin{pmatrix} 0 & 2 & 5 & 5 & 5 \\ 2 & 0 & 5 & 5 & 5 \\ 5 & 5 & 0 & 4.5 & 4.5 \\ 5 & 5 & 4.5 & 0 & 3 \\ 5 & 5 & 4.5 & 3 & 0 \end{pmatrix}$$

Cophenetic correlation

The cophenetic correlation is defined as

$$\phi = \text{Corr}(\text{vec}_{i<j}(D), \text{vec}_{i<j}(\hat{D})),$$

where $\hat{D} = (\hat{d}_{ij})$ = matrix of procrustes distances and $D = (d_{ij})$ = matrix of original distances.

In our example above, we have $\text{vec}_{i<j}(D) = (2, 6, 10, 9, 5, 9, 8, 4, 5, 3)$, $\text{vec}_{i<j}(\hat{D}) = (2, 5, 5, 5, 5, 5, 5, 4.5, 4.5, 3)$ and thus obtain $\phi \approx 0.77$.

It is recommended to use the so-called cophenetic distances instead of deliberately chosen procrustes distances. The **cophenetic distance** between two observations that have been clustered is defined to be the intergroup distance at which the two observations are first combined into a single cluster. Note that this distance has many ties.

Implementation in R:

```
> library(cluster)
> data(flower)
```

```
> dflower=daisy(flower)              # distance matrix
> hdf=hclust(dflower, "ave")
> dfl.co=cophenetic(hdf)      # cophenetic distances
> cor(dflower, dfl.co)
[1] 0.6596862                 # cophenetic correlation
```

Further measures for the evaluation of the appropriateness of clustering outputs include the

- stress measure S

$$S^2 = \sum_{i=1}^{n}\sum_{j>i}^{n}(d_{ij} - \hat{d}_{ij})^2 \Big/ \sum_{i=1}^{n}\sum_{j>i}^{n}\hat{d}_{ij}^2,$$

  Positive real numbers $\hat{d}_{ij}$ are sought to minimize the stress S, usually by genetic optimization.

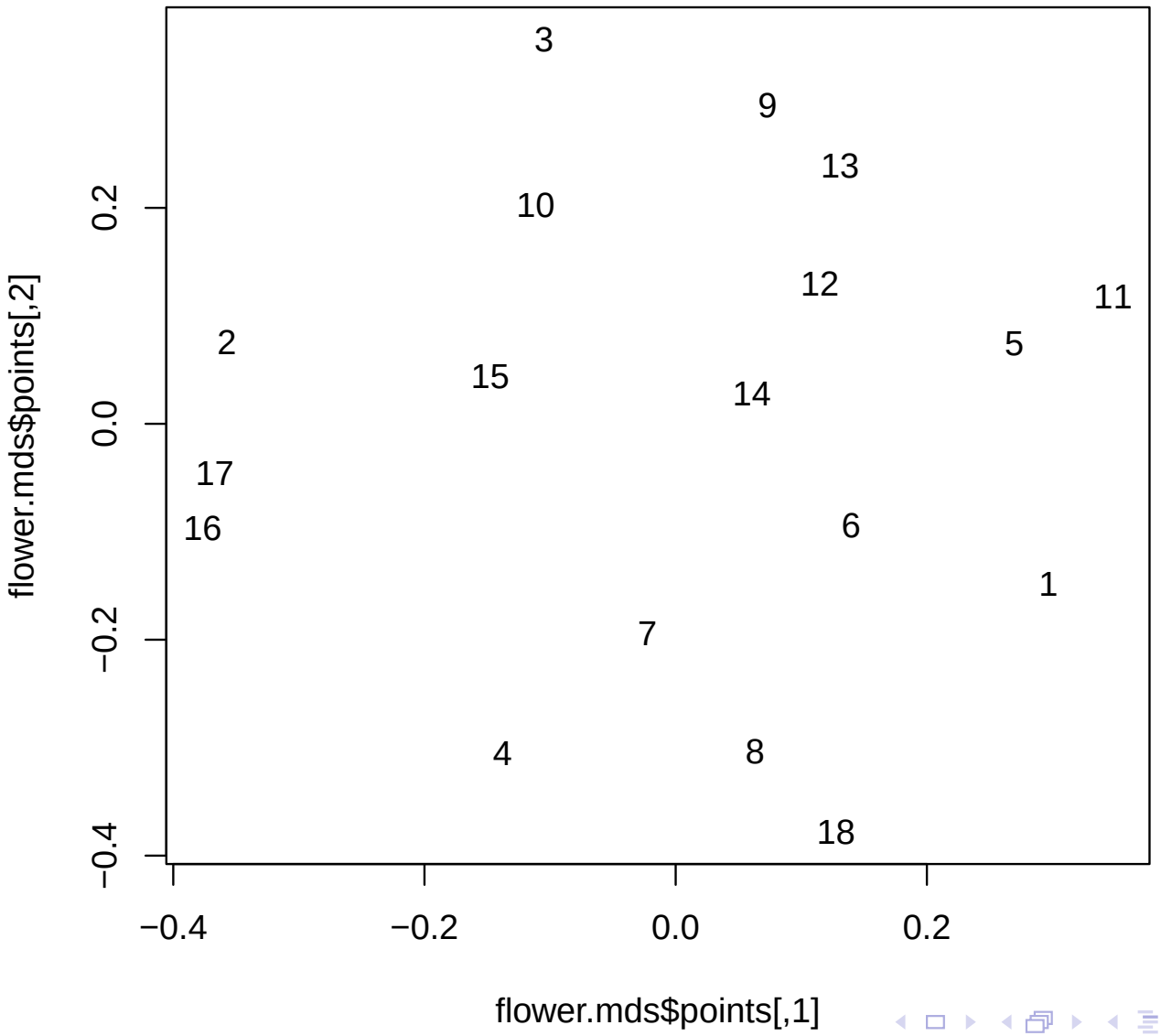
Multidimensional Scaling

- Multidimensional scaling (MDS)
  This chooses a k-dimensional (default k = 2) configuration to minimize the stress. Most commonly, Kruskal's (non-metric) MDS method is used, implemented by the R-function "isoMDS". This method is based on an iterative (steepest descent) algorithm, which will usually converge in a few iterations. The configuration is only determined up to rotations and reflections (by convention the centroid is at the origin).

```
> flower.mds= isoMDS(dflower)
initial value 28.751482
iter 5 value 24.247183
final value 23.546244
converged
```

```
> plot(flower.mds$points, type = "n")
> text(flower.mds$points, labels = as.character(1:nrow(flower)))
```

Kruskal's non-metric MDS



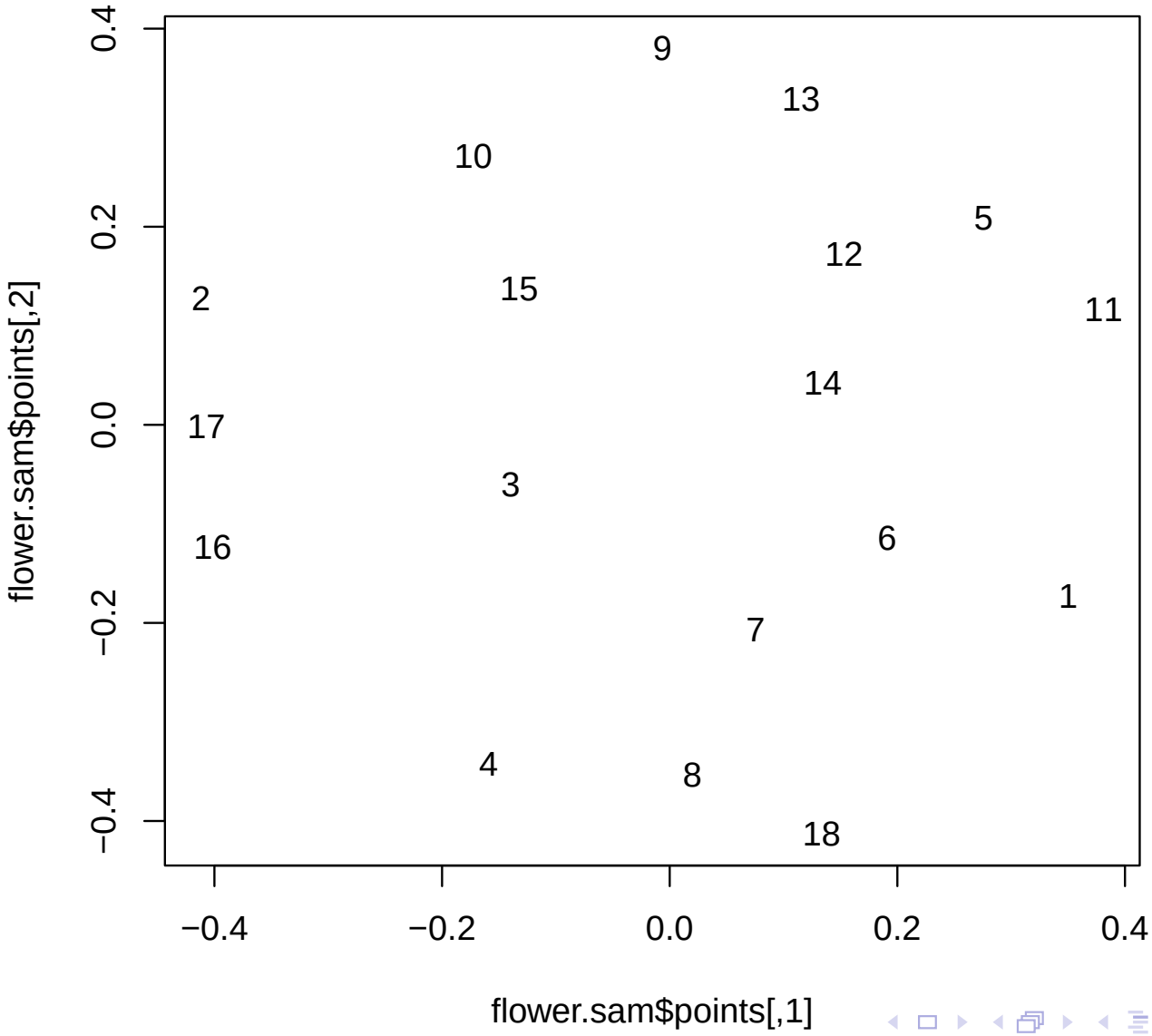
Sammon mapping

- Sammon mapping
  Another popular method of non-metric MDS to aid in a graphical visualization of the clustering process is Sammon's non-linear mapping of distances.
  This method, again, chooses a two-dimensional configuration with interpoint distances $\hat{d}_{ij}$ to minimize the (scale-free) mapping error

$$E^2 = \sum_{i=1}^{n} \sum_{j>i}^{n} d_{ij}^{-1} (d_{ij} - \hat{d}_{ij})^2 \Big/ \sum_{i=1}^{n} \sum_{j>i}^{n} d_{ij},$$

```
> flower.sam= sammon(dflower)
Initial stress : 0.11681
stress after 10 iters: 0.07863, magic = 0.092
stress after 20 iters: 0.07542, magic = 0.500
stress after 30 iters: 0.07533, magic = 0.500
> plot(flower.sam$points, type = "n")
> text(flower.sam$points, labels = as.character(1:nrow(flower)))
```

Sammon mapping



Silhouette

- Mean Silhouette-Width:
  $\overline{s} := \frac{1}{n} \sum_{i=1}^{n} s(i)$, where the silhouette-width $s(i)$ of the i-th object $\underline{x}_i = (x_{i1}, \ldots, x_{ip})$; $i \in \{1, \ldots, n\}$; is defined as follows:
  - Let $a(i)$ denote the average distance of the i-th object to all other objects belonging to the same cluster (for a single object we define $s(i) := 0$)
  - For all other clusters C with $i \notin C$ denote $d(i, C) :=$ the average distance of object i to all objects in C
  - Let $b(i) := \min_C d(i, C)$ be the distance to the "neighboring group"
  - Finally, define $s(i) := \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$.

    Then it holds: $-1 \leq s(i) \leq 1$.

    Samples with $s(i) \approx 0$ are located between groups, whereas samples with $s(i)$ close to one are well grouped and samples with $s(i) < 0$ may be grouped wrongly.

Agglomeration Quality

- Agglomerative coefficient:

  Define

  $$m(i) = d_{first}(i)/d_{last}(i)\ ;\ i = 1, ..., n$$

  where $d_{first}(i)$ and $d_{last}(i)$ denote the distances of object i from the first and last cluster, repectively, with which i is merged.

  The agglomerative coefficient AC is then defined as the average

  $$AC = \frac{1}{n}\sum_{i=1}^{n}(1 - m(i)) = 1 - \frac{1}{n}\sum_{i=1}^{n} m(i)$$

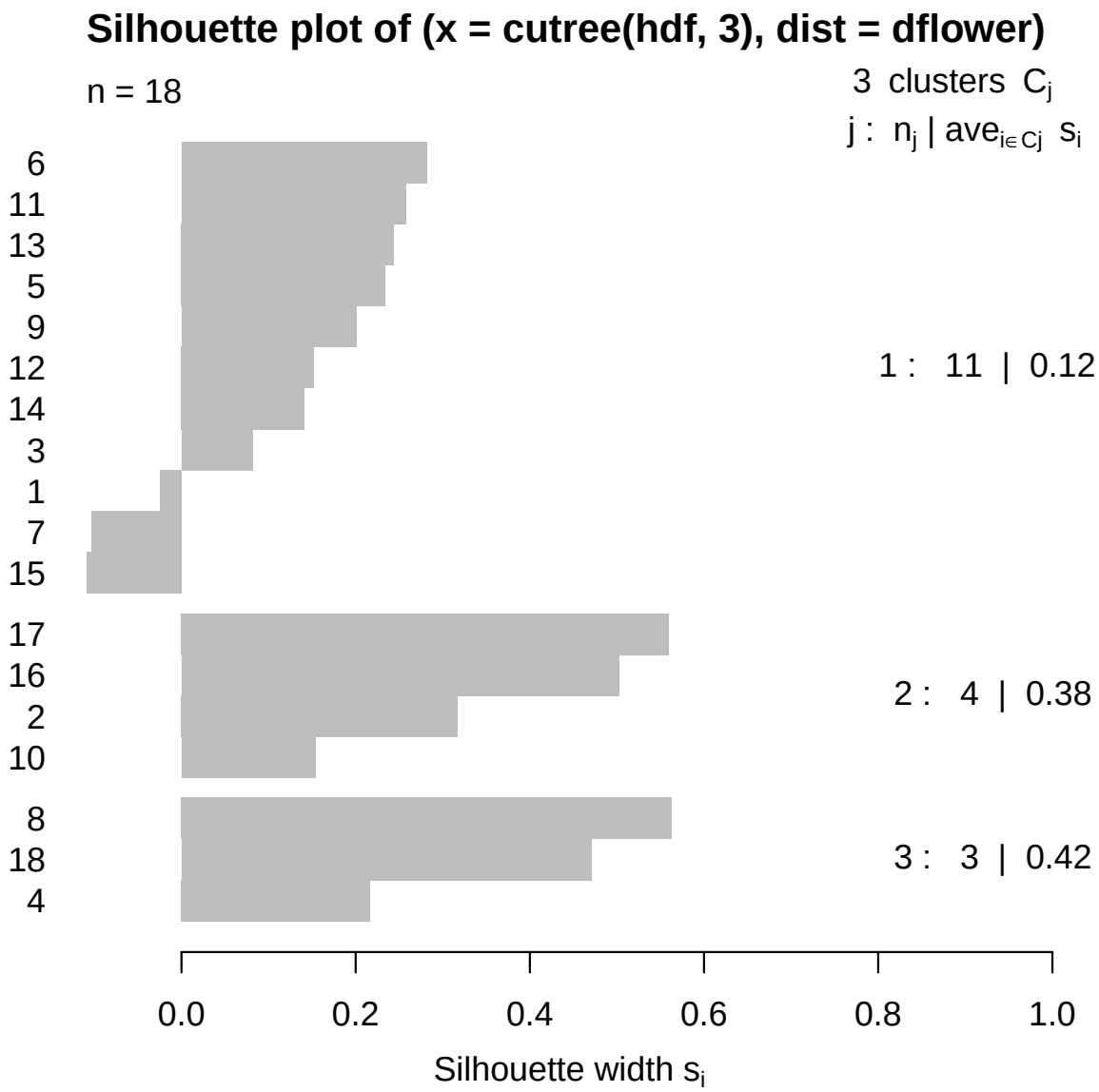
  Clearly, $0 \leq AC \leq 1$.

  AC values near 1 correspond to good clustering quality.

Computing silhouette widths and AC in R:

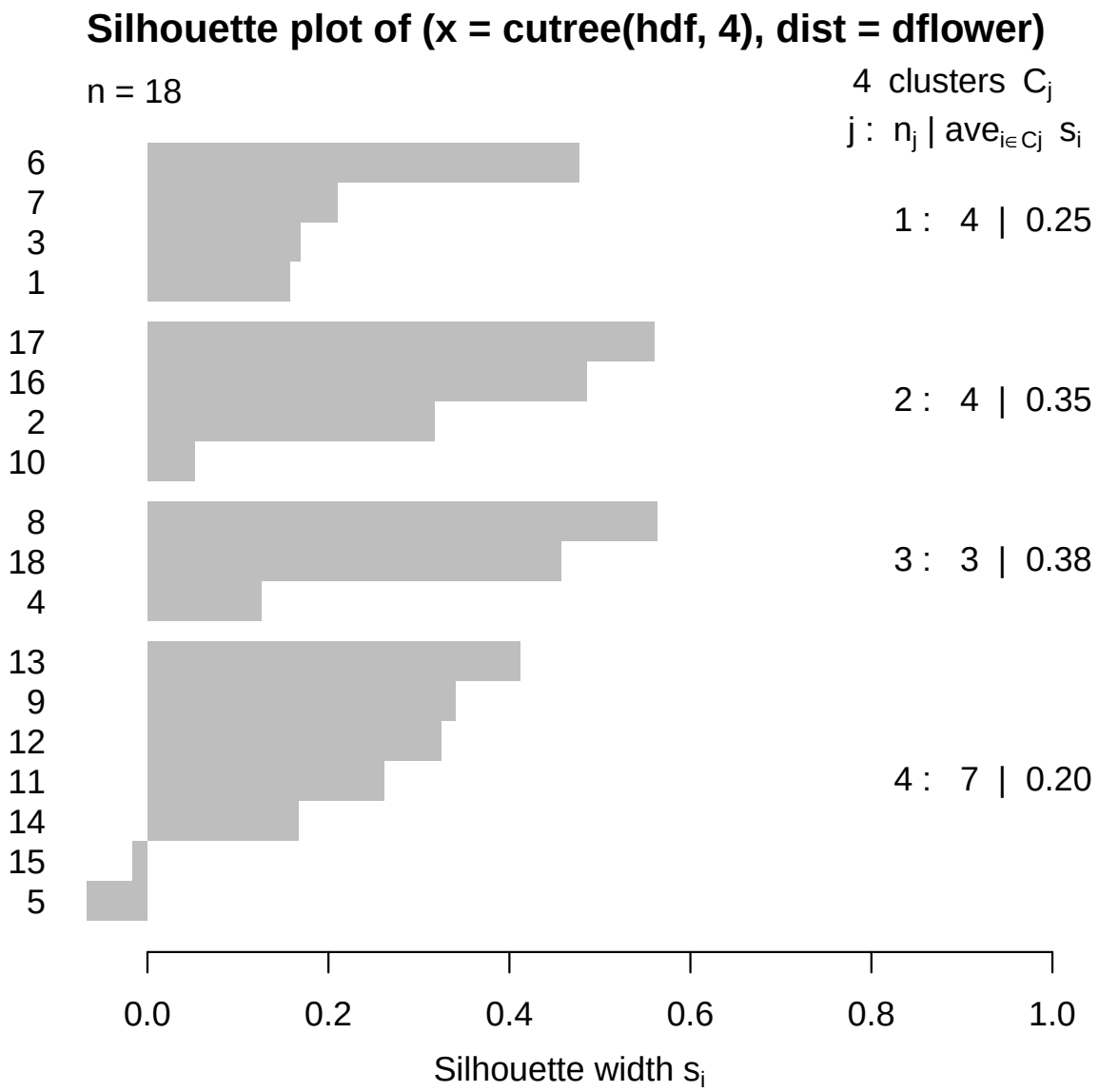
```
> plot(silhouette(cutree(hdf, 3), dflower))    # cut using three clusters
```

Agglomeration Quality: flower, 3 clusters



Average silhouette width : 0.23

Agglomeration Quality: flower, 4 clusters



Average silhouette width : 0.28

Final remarks on clustering

- Model-based clustering: **Gaussian mixture** modeling (GMM)
  This assumes a mixture of normals, i.e. the pdf of an observation reads

$$f(\underline{x}) = \sum_{i=1}^{g} \alpha_i \, \phi(\underline{\mu}_i, \Sigma_i), \ \textit{where}$$

$$\phi(\underline{\mu}_i, \Sigma_i) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp(-\tfrac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i));$$

$$0 \leqslant \alpha_i \leqslant 1, \ \ \sum_{i=1}^{g} \alpha_i = 1.$$

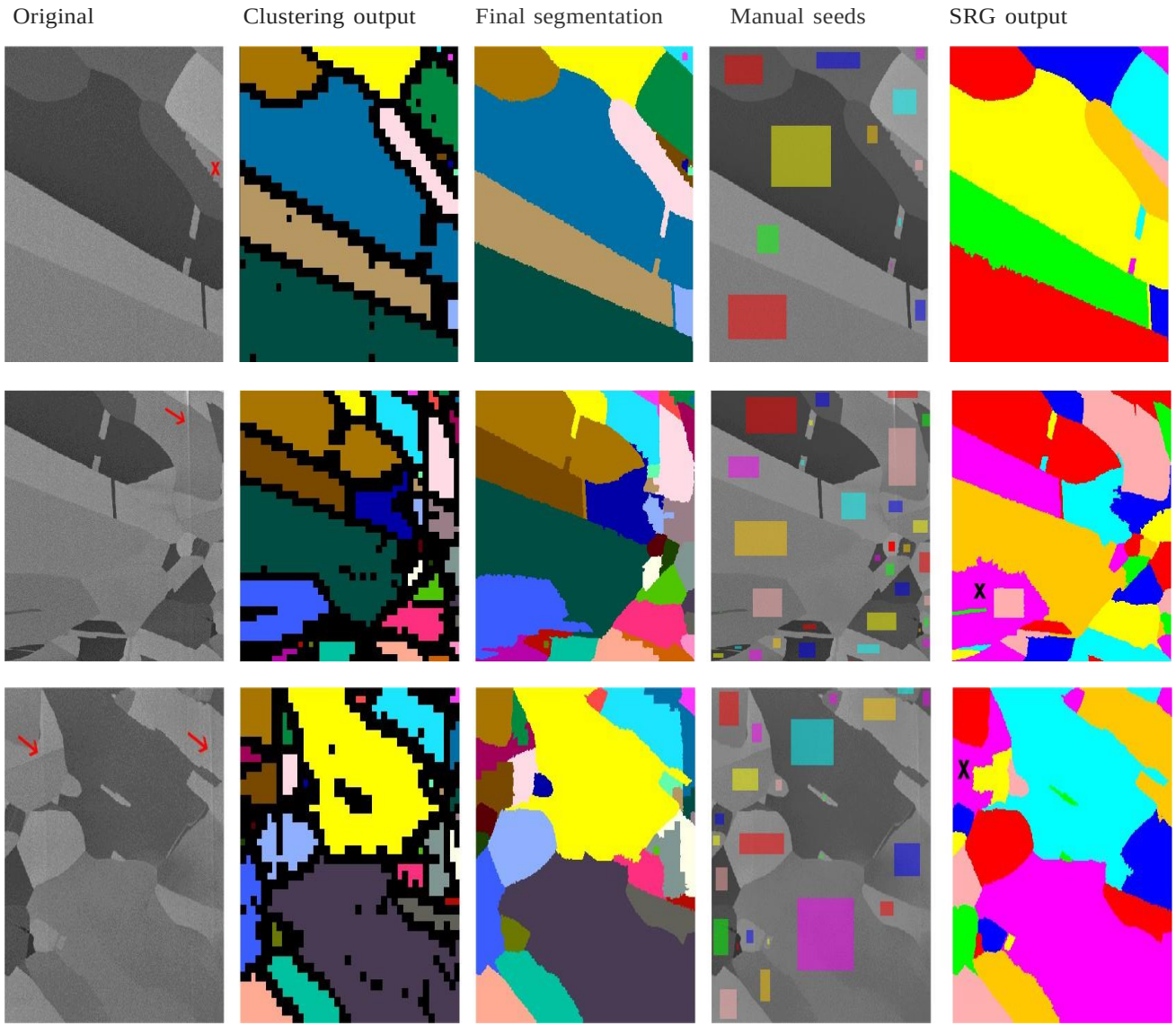
  The parameters μ_i, Σ_i and the mixture weights $\alpha_i; i = 1, ..., g;$ are estimated via an EM-algorithm.

  Implementation in R: $>$ library(mclust)

  For a recent application of this and other clustering methods see

  D. Alagic: Application of Unsupervised and Fusion Methods to Characterize the Microstructure of Polycrystalline Materials. PhD thesis, Dept. of Statistics, AAU Klagenfurt 2021.

GMM in material sciences



Final remarks on clustering

- Further remarks on implementation

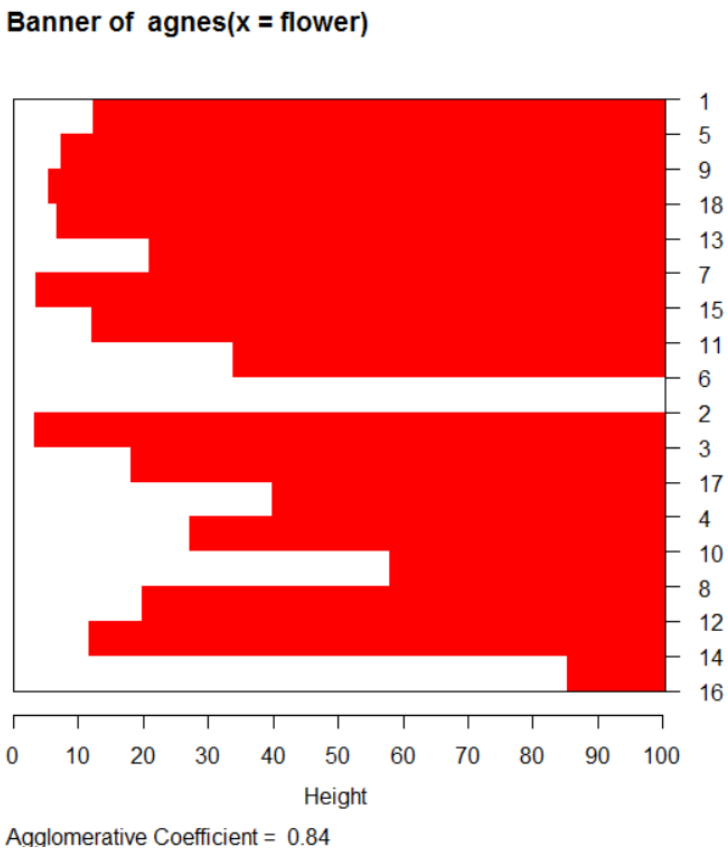
```
> agf= agnes(flower)      # agglomerative nested
> summary(agf)
Agglomerative coefficient: 0.8352937
Order of objects:
[1] 1 5 9 18 13 7 15 11 6 2 3 17 4 10 8 12 14 16
Height:
[1]   12.2209   7.3193    5.4772    6.6245  20.9567   3.4641
[7]   12.0358  33.8262  100.3639    3.3166  18.1037  39.9529
[13] 27.0740  58.0439   19.9118   11.7473  85.3493
153 dissimilarities, summarized :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.317  28.302  62.522  70.215 108.290 187.230
Metric : euclidean
Number of objects : 18
```

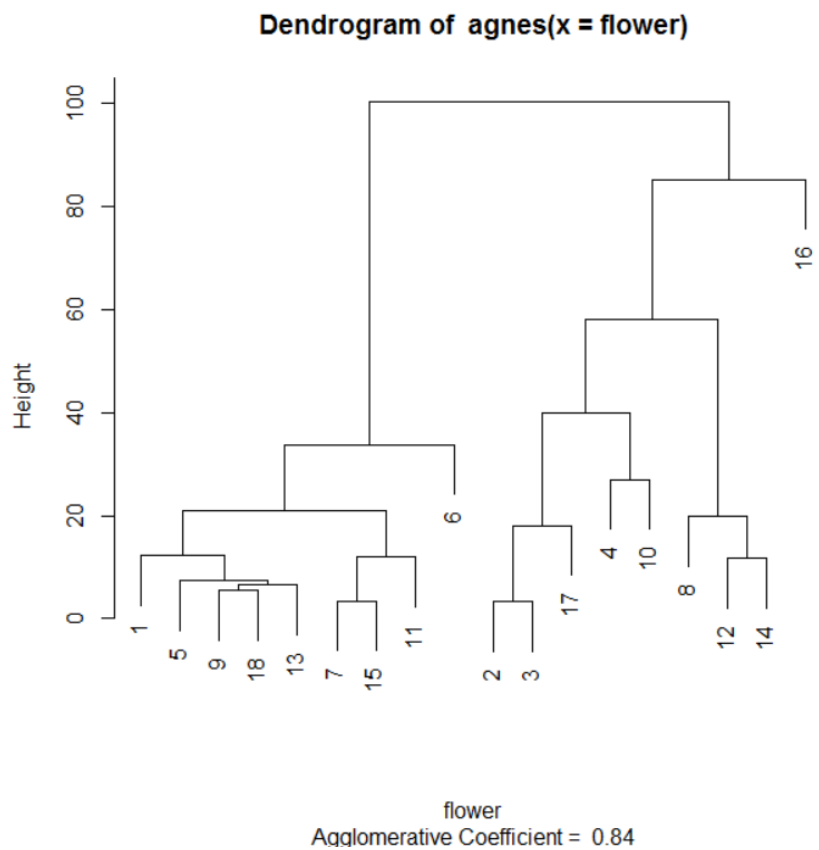
Summary and plot of agnes(flower)

Available components:

[1] "*order*" "*height*" "*ac*" "*merge*" "*diss*" "*call*" "*method*" "*data*"

> plot(agf)





Alternative cluster methods

Instead of "agnes" we might also use "diana" (brief for divisive analysis) and "fanny" (fuzzy analysis), respectively, in the same vein as above.

Correspondingly, in R we the use the commands

> dif=diana(flower); summary(dif); plot(dif) and

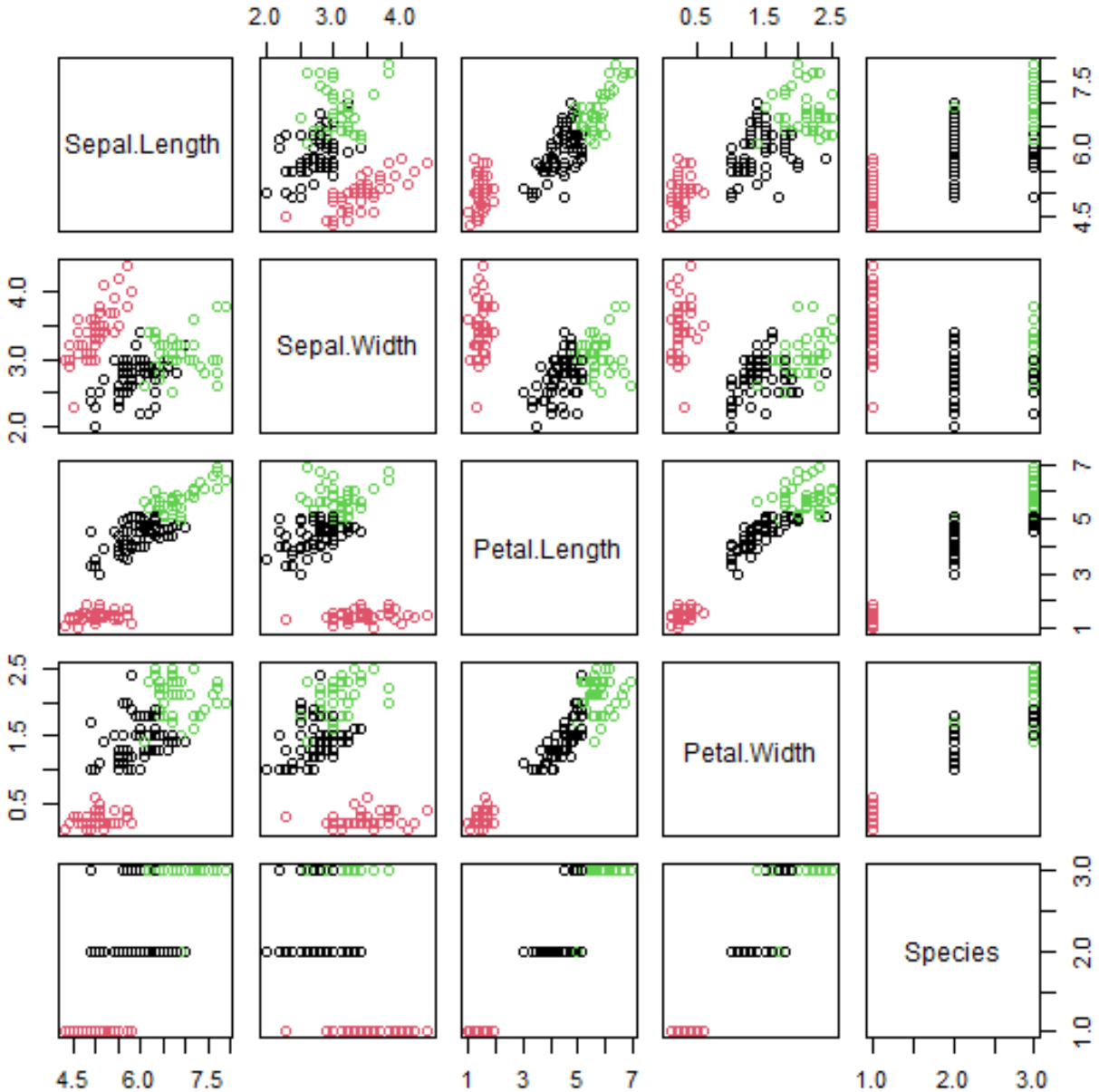
> faf=fanny(flower); summary(faf); plot(faf).

For purely numeric variables, we might also use the "kmeans" clustering method as follows:

```
> kiris=kmeans(iris[, 1:4], 3)
> plot(iris, col= kiris$cluster)     # yields the plot on the next slide
```

Note: Cluster methods are of descriptive or exploratory nature. The cluster groupings obtained must be validated by methods of statistical inference, such as MANOVA and discriminant analysis.

kmeans clustering of iris data



6. Discriminant analysis (DA)

DA can be seen as quantitative completion of MANOVA

Objectives of DA:

1. Description of the differences between groups
2. Classification of new unknown objects

6.1 Fisher's discriminant criterion

Fisher's Idea: separate objects through hyperplanes:
$Y = \underline{c}^T \underline{X}$ with $\underline{X} := (X_1, \ldots, X_p)^T$ and $\underline{c}^T = (c_1, \ldots, c_p)$.

Separation proceeds on the basis of Rao's-F-statistics, which then follows an exact F-distribution (since $dim(Y) = 1$), i.e.

$$F_Y = \frac{SSB_Y/(g-1)}{SSW_Y/(n-g)} = \frac{n-g}{g-1}\frac{SSB_Y}{SSW_Y} \sim F_{g-1,n-g}$$

where

$$SSB_Y = \underline{c}^T B \underline{c} \text{ and } SSW_Y = \underline{c}^T W \underline{c}.$$

Fisher's discriminant criterion

How to choose the coordinates of vector $\underline{c}$?

Opting for maximum possible separation, Fisher's discriminant problem reads:

Find $\underline{c}^* \in \mathbb{R}^p$ such that:

$$\max_{\underline{c} \in \mathbb{R}^p} \frac{\underline{c}^T B \underline{c}}{\underline{c}^T W \underline{c}} = \frac{\underline{c}^{*T} B \underline{c}^*}{\underline{c}^{*T} W \underline{c}^*} . \quad (FD)$$

In the special case where $W = I_p$, the ratio is just the Rayleigh quotient and the solution to problem (FD) can be found easily: $\underline{c}^*$ is the eigenvector corresponding to the maximum eigenvalue of the

$$\text{EV-problem:} \quad \det(B - \lambda I) = 0.$$

In case of $W \neq I_p$ we are led to the

$$\text{generalized EV-problem:} \quad \det(B - \lambda W) = 0. \quad (*)$$

Generalized EV-problem

$(*)$ is equivalent to the classical EV-problem

$$|W^{-1}B - \lambda I| = 0 \Leftrightarrow |W^{-1}||B - \lambda W| = 0 \Leftrightarrow \left|W^{-1/2}BW^{-1/2} - \lambda I\right| = 0$$

Since W is positive definite and B is (at least) positive semidefinite, all the eigenvalues of $(^*)$ are real and non-negative. Let the solutions be ordered such that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0; \lambda_{r+1} = \ldots = \lambda_p = 0.$$

Further, denote the eigenvectors corresponding to non-zero eigenvalues as $\underline{c}_1, \ldots, \underline{c}_r$, where

$$B\underline{c}_i = \lambda W\underline{c}_i \; ; i = 1, \ldots, r; \;\; r = \min\left(rank(W^{-1/2}BW^{-1/2}), g - 1\right).$$

W.l.o.g. we may assume that these eigenvectors are orthonormal w.r.t. the inner product $(\cdot, \cdot)_W$, i.e. we have

$$\underline{c}_i^T W\underline{c}_j = \delta_{ij} \text{ and } \underline{c}_j^T B\underline{c}_j = \lambda_j \; ; i, j = 1, \ldots, r$$

Generalized EV-problem

The new features $Y_j = \underline{c}_j^T \underline{X}$; $i = 1, \ldots, r$;
with $\underline{c}_i$ as the i-th eigenvector corresponding to the solution λ_i
of the generalized EV-problem $\det(B - \lambda W) = 0$, are called
(Fisher-) **discriminant features**.

Corollary 6.1: The Fisher-discriminant features are uncorrelated.

Proof:

$$\widehat{\mathrm{Cov}}(Y_i, Y_j) = \widehat{\mathrm{Cov}}(\underline{c}_i^T \underline{X}, \underline{c}_j^T \underline{X}) = \underline{c}_i^T \widehat{\mathrm{Cov}(\underline{X})} \underline{c}_j = \underline{c}_i^T \widehat{\Sigma} \underline{c}_j$$
$$= \frac{1}{n-g} \underline{c}_i^T W \underline{c}_j = \frac{1}{n-g} \delta_{ij} = 0 \text{ for } i \neq j = 1, ..., r.$$

We are collecting all these (significant) discriminant features
$Y_1, \ldots, Y_r$ in a vector $\underline{Y}$ of dimension $dim(\underline{Y}) = r$:

$$\underline{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} = \begin{pmatrix} \underline{c}_1^T \underline{X} \\ \vdots \\ \underline{c}_r^T \underline{X} \end{pmatrix}$$

Discriminant space

We call $\underline{Y}$ the **discriminant vector**, in matrix form it reads

$$\underline{Y} = C^T \underline{X} \text{ with } \underset{(p,r)}{C} = [\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_r].$$

The image generated by $\underline{Y}$ is called the **discriminant space**, it has dimension $r \leqslant p$. In this space, we are next forming the Within-Group and Between-Group sum of squares matrices:

- $B_{\underline{Y}} = C^T B C = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{pmatrix}$
- $W_{\underline{Y}} = C^T W C = I_r$
- $T_{\underline{Y}} = B_{\underline{Y}} + W_{\underline{Y}} = \begin{pmatrix} 1+\lambda_1 & & \\ & \ddots & \\ & & 1+\lambda_r \end{pmatrix}.$

6.2 Significance tests

Starting from Wilks-Λ-statistic in the discriminant space

$$\Lambda = \frac{\det(W_{\underline{Y}})}{\det(T_{\underline{Y}})} = \frac{1}{\prod_{j=1}^{r}(1+\lambda_j)}$$

and using the following asymptotic result

$$\Lambda \sim \Lambda_p(m,k) \Longrightarrow -\left[m - \frac{1}{2}(p-k+1)\right] \log(\Lambda) \overset{as.}{\sim} \chi^2_{pk}$$

we obtain in our case of $m = n-g$, $k = g-1$:

Corollary 6.2:

$$\left[n-1-\frac{1}{2}(p+g)\right] \sum_{j=1}^{r} log(1+\lambda_j) \overset{as.}{\sim} \chi^2_{p(g-1)} \text{ as } n \to \infty$$

SLRT

This result allows us to develop a sequential likelihood-ratio-test (SLRT) for checking the hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g \Leftrightarrow \tilde{H}_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_r = 0$$

that there are no group differences vs. the alternative that there are at least two different groups.

The test proceeds as follows:

Step 1: Test of H_0. Stop when it is accepted, then there are no group differences. If H_0 is rejected (i.e. $\lambda_1 > 0$), then go to

Step 2: Test the new hypothesis $H_{0_1} : \lambda_2 = \ldots = \lambda_r = 0$ using the modified test statistic

$$\left[n - 1 - (\frac{p+g}{2})\right] \sum_{j=2}^{r} \log(1 + \lambda_j) \overset{as}{\sim} \chi^2_{(p-1)(g-2)}$$

In case of rejection of H_{0_1} continue with testing.

Step (k+1): Test $H_{0_k} : \lambda_{k+1} = \ldots = \lambda_r = 0$ using

$$\left[n - 1 - (\frac{p+g}{2})\right] \sum_{j=k+1}^{r} log(1+\lambda_j) \overset{as.}{\sim} \chi^2_{(p-k)(g-k-1)}$$

In case of rejection, replace k by $k+1$ and continue with testing.

Stop, whenever $H_{0_k} : \lambda_{k+1} = \ldots = \lambda_r = 0$ is accepted for the first time. Then we have at least $k+1$ different groups; $1 \leq k \leq r-1$.

6.3 Discriminant analysis and classification

Assumption: $\Sigma_1 = \ldots = \Sigma_g = \Sigma$ (Variance homogeneity).
We start with the special case of $g = 2$ groups. This implies that

$$r = min(p, g-1) = 1 \implies \lambda_1 > 0.$$

Q: Can we determine the eigenvalue λ_1 explicitly?

A: Yes!

Finding the maximum eigenvalue

Let be given two groups of normally distributed observations: $N_p(\underline{\mu}_1, \Sigma)$, $N_p(\underline{\mu}_2, \Sigma)$ and assume, for the moment, that $\underline{\mu}_1, \underline{\mu}_2, \Sigma$ are known. Then we have

$$\max_{\underline{c}} \frac{\underline{c}^T B \underline{c}}{\underline{c}^T W \underline{c}} = \max_{\underline{c}} \frac{\underline{c}^T (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{c}}{\underline{c}^T \Sigma \underline{c}} = \lambda_1 \ ,$$

since the ratio is just the Rayleigh quotient. Hence, it follows

$$\lambda_1 = \text{tr}\left(\Sigma^{-1}\left(\underline{\mu}_1 - \underline{\mu}_2\right)\left(\underline{\mu}_1 - \underline{\mu}_2\right)^T\right) = \left(\underline{\mu}_1 - \underline{\mu}_2\right)^T \Sigma^{-1}\left(\underline{\mu}_1 - \underline{\mu}_2\right)$$

i.e. λ_1 coincides with the Mahalanobis distance between $\underline{\mu}_1$ and $\underline{\mu}_2$: $\lambda_1 = d_M^2\left(\underline{\mu}_1, \underline{\mu}_2\right)$. The corresponding eigenvector then takes the form

$$\underline{c}_1 = k \cdot \Sigma^{-1}\left(\underline{\mu}_1 - \underline{\mu}_2\right)$$

where k is a constant satisfying the normalization condition:

Fisher discrimination

$$\underline{c}_1^T \Sigma \underline{c}_1 = 1 \Longleftrightarrow k^2 \left(\underline{\mu}_1 - \underline{\mu}_2\right)^T \Sigma^{-1} \Sigma \Sigma^{-1} \left(\underline{\mu}_1 - \underline{\mu}_2\right) = 1.$$

This, however, implies that

$$k = 1/\sqrt{d_M^2(\underline{\mu}_1, \underline{\mu}_2)}\,.$$

Using this eigenvector, we find the separating hyperplane between group 1 and group 2.

Theorem (Fisher's discriminant rule)

The object $\underline{x} = (x_1, \ldots, x_p)^T$ is allocated to group 1 if it holds:

$$\underline{c}_1^T \underline{x} > \frac{1}{2}\left(\underline{c}_1^T \underline{\mu}_1 + \underline{c}_1^T \underline{\mu}_2\right) = m\,.$$

This is equivalent to the allocation rule

$$\underline{x} \longmapsto \textit{group 1} \Longleftrightarrow d_M(\underline{x}, \underline{\mu}_1) < d_M(\underline{x}, \underline{\mu}_2)\,.$$

Fisher discrimination

Proof: $d_M^2(\underline{x}, \underline{\mu}_1) < d_M^2(\underline{x}, \underline{\mu}_2)$

$$\iff (\underline{x} - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_1) < (\underline{x} - \underline{\mu}_2)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_2)$$

$$\iff \underline{x}^T \Sigma^{-1} \underline{x} + \underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 - 2\underline{\mu}_1^T \Sigma^{-1} \underline{x} < \underline{x}^T \Sigma^{-1} \underline{x} + \underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2 - 2\underline{\mu}_2^T \Sigma^{-1} \underline{x}$$

$$\iff \underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2 < 2(\underline{\mu}_1 \Sigma^{-1} \underline{x} - \underline{\mu}_2 \Sigma^{-1} \underline{x})$$

$$\iff \tfrac{1}{2}(\underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2) < \underline{x}^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \qquad / \cdot k$$

$$\iff \tfrac{1}{2} k \cdot (\underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2) < k \cdot \underline{x}^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

$$\iff \tfrac{k}{2} (\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2) < \underline{c}_1^T \underline{x}$$

$$\iff \tfrac{1}{2} \underline{c}_1^T (\underline{\mu}_1 + \underline{\mu}_2) < \underline{c}_1^T \underline{x} \, .$$

Numerical Example: $p = 2$, $g = 2$;

$$\underline{\mu}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \quad \underline{\mu}_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

Example

The eigenvector then becomes

$$\underline{c}_1 = k \cdot \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) = k \cdot \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = k \cdot \begin{pmatrix} 3 \\ -2 \end{pmatrix},$$

where k derives from the Mahalanobis distance

$$d_M^2(\underline{\mu}_1, \underline{\mu}_2) = (1, -1) \begin{pmatrix} 3 \\ -2 \end{pmatrix} = 5 \Longrightarrow k = 1/\sqrt{5}.$$

Further, $m = \frac{1}{2}\underline{c}_1^T \left(\underline{\mu}_1 + \underline{\mu}_2\right) = \frac{k}{2}(3, -2) \begin{pmatrix} 3 \\ 5 \end{pmatrix} = -\frac{k}{2} = -\frac{1}{2\sqrt{5}}$.

Thus, the allocation proceeds as follows:

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \text{ group 1 } \Longleftrightarrow \underline{c}_1^T \underline{x} > m$$

$$\Longleftrightarrow k(3x_1 - 2x_2) > -\frac{k}{2} \Longleftrightarrow x_2 < \frac{3}{2}x_1 + \frac{1}{4}\ .$$

All objects $\underline{x}$ falling below the separating hyperplane $x_2 = \frac{3}{2}x_1 + \frac{1}{4}$ are allocated to group 1, those above this line will go to group 2.

Further discrimination rules

Note: For the group centers it holds that $\underline{c}_1^T \underline{\mu}_1 > m > \underline{c}_1^T \underline{\mu}_2$.

To see this observe that $\underline{c}_1^T \underline{\mu}_1 - m = m - \underline{c}_1^T \underline{\mu}_2 = \frac{k}{2}\, d_M^2(\underline{\mu}_1, \underline{\mu}_2) > 0$.

6.4 ML- and Bayes discrimination for $g = 2$ groups

a) Maximum-Likelihood criterion

This allocates the object to the group which has larger likelihood:

$$\underline{x} \longmapsto \text{group 1}$$

$$\Longleftrightarrow L\left(\underline{\mu}_1, \Sigma; \underline{x}\right) > L\left(\underline{\mu}_2, \Sigma; \underline{x}\right) \Longleftrightarrow f\left(\underline{x}|\text{group 1}\right) > f\left(\underline{x}|\text{group 2}\right).$$

In case of multivariate normality of $\underline{X}$ this means:

$$\underline{x} \longmapsto \text{group 1}$$

$$\Longleftrightarrow a\, \exp\left(-\tfrac{1}{2} d_M^2(\underline{x}, \underline{\mu}_1)\right) > a\, \exp\left(-\tfrac{1}{2} d_M^2(\underline{x}, \underline{\mu}_2\right)$$

$$\Longleftrightarrow d_M^2\left(\underline{x}, \underline{\mu}_1\right) < d_M^2\left(\underline{x}, \underline{\mu}_2\right), \text{where } a = (2\pi)^{-p/2} \det(\Sigma)^{-1/2}.$$

This is, however, equivalent to Fisher's discriminant rule.

Further discrimination rules

b) Bayes criterion

This allocates the object to the group which has the larger posterior probability. To formalize this, let denote
p_i the prior probability for objects to be part of group i; $i = 1, 2$;
$p_1, p_2 > 0$; $p_1 + p_2 = 1$.

The posterior probabilities for the group memberships then read:

$$\begin{aligned} \underline{x} \longmapsto \text{ group 1} &\iff P(\text{group 1}|\underline{x}) > P(\text{group 2}|\underline{x}) \\ &\iff \frac{p_1 f(\underline{x}|\text{group 1})}{f(\underline{x})} > \frac{p_2 f(\underline{x}|\text{group 2})}{f(\underline{x})} \\ &\iff p_1 f(\underline{x}|\text{group 1}) > p_2 f(\underline{x}|\text{group 2}), \end{aligned}$$

where $f(\underline{x}) = p_1 f(\underline{x}|\text{group 1}) + p_2 f(\underline{x}|\text{group 2})$, according to the law of total probability.

In case of equal prior probabilities $p_1 = p_2 = \frac{1}{2}$, the Bayes criterion coincides with the ML-criterion and Fisher's discriminant rule.

6.5 Classification with unequal covariance matrices

Consider $g = 2$ groups with different means and covariance matrices

$$\text{Group 1: } N(\underline{\mu}_1, \Sigma_1)\ , \quad \text{Group 2: } N(\underline{\mu}_2, \Sigma_2);$$
$$\text{where } \ \underline{\mu}_1 \neq \underline{\mu}_2\ ;\ \Sigma_1 \neq \Sigma_2\ .$$

Then the ML-allocation rule says that

$$\underline{x} \longmapsto \text{group 1} \iff f(\underline{x}|\text{group 1}) > f(\underline{x}|\text{group 2})$$

$$\iff |\Sigma_1|^{-1/2} \exp\left(-\tfrac{1}{2} d_M^2(\underline{x}, \underline{\mu}_1)\right) > |\Sigma_2|^{-1/2} \exp\left(-\tfrac{1}{2} d_M^2(\underline{x}, \underline{\mu}_2)\right)$$

$$\iff (\underline{x} - \underline{\mu}_1)^T \Sigma_1^{-1} (\underline{x} - \underline{\mu}_1) < (\underline{x} - \underline{\mu}_2)^T \Sigma_2^{-1} (\underline{x} - \underline{\mu}_2) + \ln(|\Sigma_2|/|\Sigma_1|).$$

Obviously, this is a quadratic discriminant analysis **(QDA)**-type rule

$\underline{x} \longmapsto \text{group 1} \iff \underline{x}^T A \underline{x} - 2\underline{b}^T \underline{x} + c < 0$, where $A = \Sigma_1^{-1} - \Sigma_2^{-1}$,

$\underline{b} = \Sigma_1^{-1} \underline{\mu}_1 - \Sigma_2^{-1} \underline{\mu}_2$ and $c = \mu_1^T \Sigma_1^{-1} \underline{\mu}_1 - \mu_2^T \Sigma_2^{-1} \underline{\mu}_2 - \ln(|\Sigma_2|/|\Sigma_1|)$.

Testing variance homogeneity

In applications, $\underline{\mu}_1$, $\underline{\mu}_2$, Σ_1 and Σ_2 have to be replaced by the respective estimates : $\hat{\underline{\mu}}_1$, $\hat{\underline{\mu}}_2$, $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$.

If the two covariance matrices do not differ significantly, then it is recommended to use Fisher's linear discriminant analysis **(LDA)** rule with the pooled covariance matrix

$$\widehat{\Sigma} = \frac{1}{n-2}\left((n_1-1)\widehat{\Sigma}_1 + (n_2-1)\widehat{\Sigma}_2\right).$$

The equality of covariance matrices of $g \geq 2$ different groups can be tested on the basis of Box's LRT statistics

$$M = (n-g)\ln|\widehat{\Sigma}| - \sum_{k=1}^{g}(n_k-1)\ln|\widehat{\Sigma}_k|$$

In practice, there are various transformations of the value of M to yield a test statistic with an approximately known distribution. The χ^2-approximation due to Box (1950) reads

Testing variance homogeneity

$$B_M = -2(1-b)\ln(M) \sim \chi^2_q\ ,\ q = \tfrac{1}{2}p(p+1)(g-1)$$

with q degrees of freedom and bias correction constant

$$b = \left(\sum_{k=1}^{g} \frac{1}{n_k-1} - \frac{1}{n-g}\right)\ \frac{2p^2+3p-1}{6(p+1)(g-1)}\ .$$

In R, the test can be executed as follows:

$>$ X=iris[, -5]; X=as.matrix(X) # Data matrix (150x4) of iris data

$>$ boxM(X $\sim$ Species, data= iris)

Box's M-test for Homogeneity of Covariance Matrices

Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16 .

Additional information can be obtained using *summary*(*boxM*(,)).

The generalization of the above discriminant rules to $g \geqslant 2$ groups with significantly different covariance matrices is straight forward. For example, the ML-QDA rule becomes

$$\underline{x} \longmapsto \text{group } i_0 \iff i_0 = \underset{i=1,\ldots,g}{\arg\min} \left\{ \ln\left(|\Sigma_i|\right) + \left(\underline{x} - \underline{\hat{\mu}}_i\right)^T \hat{\Sigma}_i^{-1} \left(\underline{x} - \underline{\hat{\mu}}_i\right) \right\}$$

6.6 Discrimination and class prediction in R

```
> data(iris); dim(iris)
[1]  150  5
> train= sample(1:150, 100)      # training size=100
> table(iris$Species[train])
    setosa versicolor  virginica
        35         32         33
> z= lda(Species ~ ., iris, prior= c(1,1,1)/3, subset= train)
> predict(z, iris[-train, ])$class
 [1] seto seto seto seto seto seto seto seto seto seto seto seto seto seto
[14] seto seto vers vers vers vers vers vers vers vers vers vers vers vers
[27] vers vers vers vers vers vers vers virg virg virg virg virg virg
[40] virg virg virg virg vers virg virg virg virg virg virg
```

Classification in R

```
> z1= update(z, .~ . - Petal.Width)
> predict(z1, iris[-train, ])$class
 [1] seto seto seto seto seto seto seto seto seto seto seto seto seto seto
[14] seto seto vers vers vers vers vers vers vers vers vers vers vers vers
[27] vers vers vers vers vers vers vers virg virg virg virg virg virg
[40] virg vers virg vers virg virg virg virg vers virg virg
```

```
> predpost= predict(z, iris[-train, ])$posterior
> round(predpost, 3)
```

	setosa	*versicolor*	*virginica*
2	1.000	0.000	0.000
4	1.000	0.000	0.000
.....			
44	1.000	0.000	0.000
53	0.000	0.997	0.003
.....			
73	0.000	0.870	0.130

96	0.000	1.000	0.000
104	0.000	0.002	0.998
.....			
127	0.000	0.184	0.816
134	0.000	0.849	0.151
135	0.000	0.209	0.791
.....			
145	0.000	0.000	1.000
147	0.000	0.004	0.996.

Next we give the coordinates of the predicted objects in the discriminant space of dimension $r = 2$:

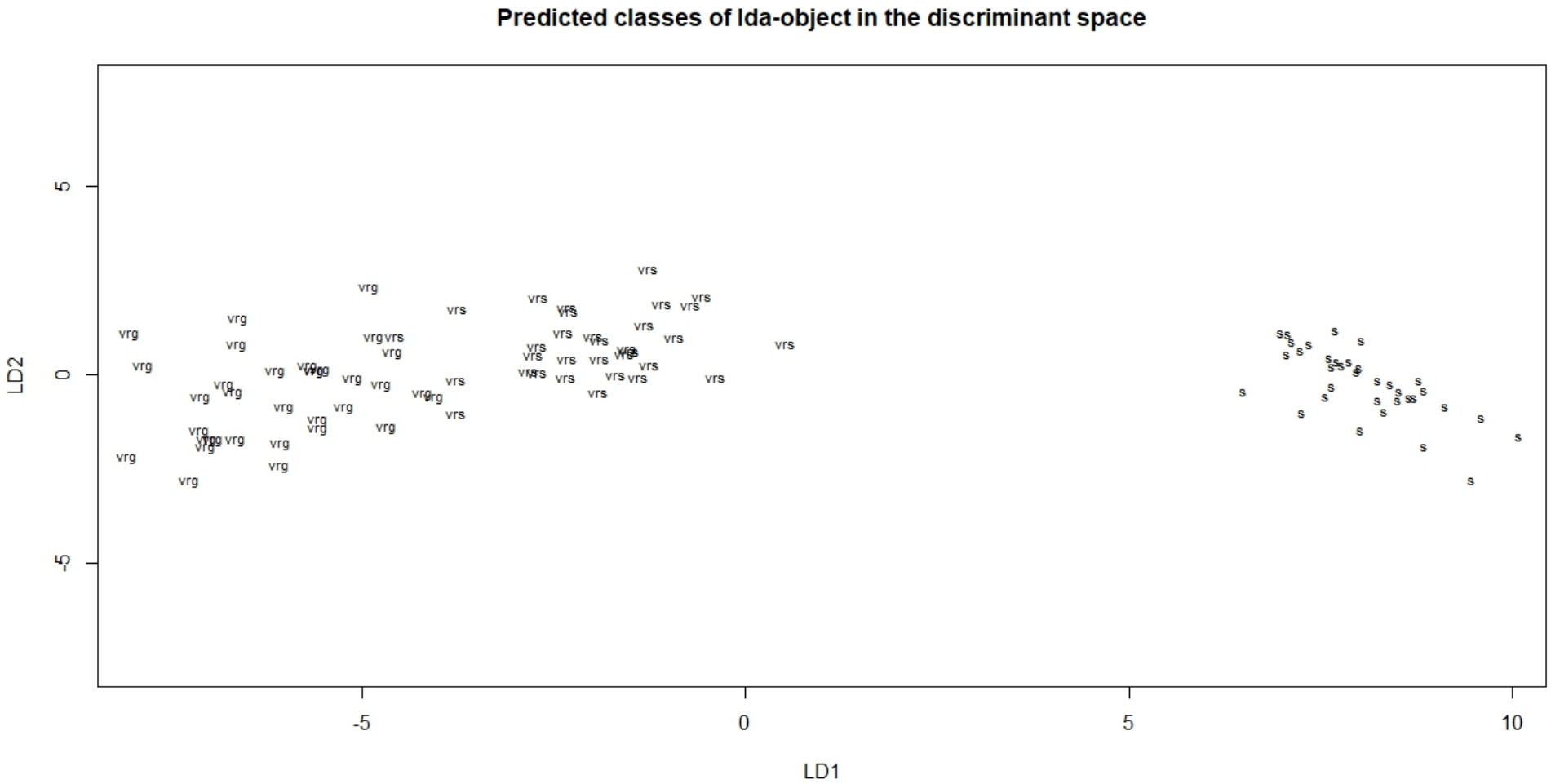
```
> ldac= predict(z, iris[-train, ])$x
> dim(ldac)
[1]     50      2
> round(ldac, 4)
```

	$LD1$	$LD2$
2	7.4258	0.8547
4	7.2207	0.7783

44	6.6734	− 1.2856
53	− 2.6510	0.1306
57	− 2.5407	− 0.7721
.....	.....	.....
96	− 1.0994	0.6612
104	− 5.7356	0.4450
105	− 7.1222	− 0.8370
.....	.....	.....
134	− 3.9243	1.0642
135	− 5.1113	2.3789
.....	.....	.....
145	− 7.2241	− 2.5645
147	− 5.4886	0.3465

The plot of the predicted class labels of the lda-object z in the discriminant space can be done using the command

```
> plot(z, abbrev=T, main="Predicted classes of lda-object
+ in the discriminant space")
```


Predicted classes of lda-object in the discriminant space
LD2
LD1
5
0
-5
-5
0
5
10
vrg
vrs
s

Finally, we note that quadratic discrimination can be done accordingly, replacing "lda" by "qda":

```
> zq= qda(Species ~ ., iris, prior= c(1,1,1)/3, subset= train)
> predict(zq, iris[-train, ])$class
 [1] seto seto seto seto seto seto seto seto seto seto seto seto seto seto
[14] seto seto vers vers vers vers vers vers vers vers vers vers vers vers
[27] vers vers vers vers vers vers vers virg virg virg virg virg virg
[40] virg virg virg virg vers virg virg virg virg virg virg
```

This classification coincides with that obtained on the basis of the lda-object *z*.

Remark 1: A full Bayesian approach to the discrimination problem using Gaussian mixture modeling (GMM) with prior Wishart distributions for the unknown covariance matrices of the different classes has been developed in the paper by Kazianka, Mulyk and Pilz (JAS, 2011). This paper includes an application to skin cancer data classification.

Bayes Deep Learning

Remark 2: More recently, deep neural networks have gained popularity for classification problems in high and ultra-high dimensions. Bayes deep learning methods often make use of the variational Bayes principle, where the complex posterior distribution is approximated by a more convenient distribution minimizing the Kullback-Leibler distance to the posterior (instead of approximating it on the basis of time-consuming MCMC methods). For Bayesian classification and (3D point cloud) segmentation with applications in the automotive industry see e.g.

`https://www.mdpi.com/1099-4300/23/3/301`

7. Principal Component Analysis (PCA)

This method dates back to Hotelling (1933). It is still one of the most important methods of data compression and dimension reduction.

Problems:

- high-dimensinonal data matrix $X_{n\times p}$
- many (highly) correlated variables (features)

Goals:

- Reduction of the number of features
- Improved interpretation of the observed variables
- De-correlation of the features
- Improvement of the numerical stability of statistical inference

Applications are manifold, e.g.:

- Statistical Quality Control (industrial manufacturing, chemical processes,. . .)

- Pattern Recognition Problems (handwritten documents, language, image processing,. . .)
- ”Feature" Extraction (Machine Learning)

7.1 Principal component transformation

We aim at a projection of our p-dimensional data into a subspace of lower dimension $r < p$.
This is effected by an affine-linear transformation of the observations:

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \mapsto \underline{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} = \underset{(r\times p)}{A}\,\underline{X} + \underset{(r\times 1)}{\underline{b}}$$

such that the following two requirements are met:

R 1: minimum information loss in terms of total variability, i.e.

$$V_t(\underline{X}) = \sum_{i=1}^{p} \mathrm{Var}(X_i) \approx \sum_{i=1}^{r} \mathrm{Var}(Y_i) = V_t(\underline{Y})$$

R 2: Uncorrelatedness, i.e. $\mathrm{Cov}(\underline{Y}) = D$, where D is a diagonal matrix.

PC-transformation

Starting point: r= p,
assume that $\mathbb{E}(\underline{X}) = \underline{\mu}$, $\mathrm{Cov}(\underline{X}) = \Sigma$ exist,
no assumption of normality is needed!

Consider the spectral decomposition: $\Sigma = U\Lambda U^T$,
where $U = (\underline{u}_1, ..., \underline{u}_p)$ is the matrix of orthonormal eigenvectors and Λ the diagonal matrix of the ordered eigenvalues of $\Sigma = \mathrm{Cov}(\underline{X})$, i.e.

$$UU^T = U^T U = I_p \text{ and } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p);\ \lambda_1 \geqslant \ldots \geqslant \lambda_p > 0.$$

The directions of the maximum variability in the data correpond to the directions of the eigenvectors of $\Sigma = \mathrm{Cov}(\underline{X})$. Since the eigenvalues of Λ are ordered in decreasing sequence, the direction u_i contains more information than direction u_j whenever $i > j$. The above requirements R 1 and R 2 therefore suggest to consider the data in the coordinate system with basis vectors $\underline{u}_1, \ldots, \underline{u}_p$.

PC-transformation

The affine-linear principal component transformation $\underline{Y} = A\underline{X} + \underline{b}$ of the centered data then reads:

$$\underline{Y} = U^T(\underline{X} - \underline{\mu}) \text{ , } i.e.\ A := U^T,\ \underline{b} := -U^T\underline{\mu} \quad (PCT)$$

The components of the transformed vector $\underline{Y}$ are called **principal components**.

Corollary 7.1: The principal components are uncorrelated with variances equal to the eigenvalues of Σ, i.e. $\text{Cov}(\underline{Y}) = \Lambda$.

Proof: $\text{Cov}(\underline{Y}) = \text{Cov}(U^T\underline{X}) = U^T\text{Cov}(\underline{X})U$

$$= U^T\Sigma U = U^T U\Lambda U^T U$$

$$= \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p).$$

Hence, the requirement $R\,2$ above is met.

PC-transformation

Corollary 7.2: The principal component transformation (PCT) preserves the total variability of the data, i.e. $\sum_{i=1}^{r} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(X_i)$.

Proof: $\sum_{i=1}^{r} \mathrm{Var}(Y_i) = \mathrm{tr}(\mathrm{Cov}(\underline{Y})) = \mathrm{tr}(\Lambda)$

$$= \mathrm{tr}(U\Lambda U^T) = \mathrm{tr}(\Sigma)$$

$$= \sum_{i=1}^{p} \mathrm{Var}(X_i).$$

Thus, requirement $R\,2$ is satisfied, too, with exact equality since $r = dim(\underline{Y}) = dim(\underline{X}) = p$.

It is common practice to standardize the vector of PCs:

$$\underline{Y}_{st} = \Lambda^{-1/2}\underline{Y}$$

is called the vector of **standardized PCs**. In fact, we have

$$\mathrm{Cov}(\underline{Y}_{st}) = \Lambda^{-1/2}\,\mathrm{Cov}(\underline{Y})\,\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda\Lambda^{-1/2} = I_p.$$

PC-transformation

Moreover, $\underline{Y}_{st}$ is centered, i.e.

$$\mathbb{E}(\underline{Y}_{st}) = \mathbb{E}\left(\Lambda^{-1/2}U^T(\underline{X} - \underline{\mu})\right) = \Lambda^{-1/2}U^T\left(\mathbb{E}(\underline{X}) - \underline{\mu}\right) = \underline{0}.$$

Corollary 7.3: The covariances between the PCs and the original variables of $\underline{X}$ are given by:

$$\mathrm{Cov}(X_i, Y_j) = u_{ij}\lambda_j; \quad i, j = 1, \ldots, p$$

Proof:
$$\begin{aligned}
\mathrm{Cov}(\underline{X}, \underline{Y}) &= \mathbb{E}\left[(\underline{X} - \mathbb{E}(\underline{X}))(\underline{Y} - \mathbb{E}(\underline{Y})^T\right] \\
&= \mathbb{E}\left[(\underline{X} - \mathbb{E}(\underline{X}))\underline{Y}^T\right] \\
&= \mathbb{E}\left[(\underline{X} - \mathbb{E}(\underline{X}))(\underline{X} - \underline{\mu})^T U\right] \\
&= \mathrm{Cov}(\underline{X})U = \Sigma U = U\Lambda U^T U = U\Lambda.
\end{aligned}$$

The impact of the original variables X_i on the PCs Y_j will be measured

Loadings

through the correlations

$$\mathrm{Cor}(X_i, Y_j) = \frac{\mathrm{Cov}(X_i, Y_j)}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(Y_j)}} = \frac{u_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}} = u_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}}$$

These values are collected in the matrix

$$L = (l_{ij}) = (\mathrm{Corr}(X_i, Y_j))_{i,j=1,\ldots,p}.$$

The matrix L is called the **matrix of loadings**.

Corollary 7.4: The matrix of loadings can be represented as

$$\underset{p\times p}{L} = D^{-1/2}U\Lambda^{1/2} = (u_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}}) \qquad i,j = 1,\ldots,p$$

where $D = \mathrm{diag}(\Sigma) := \begin{pmatrix} \sigma_{11} & 0 & \ldots & 0 \\ 0 & \sigma_{22} & \ldots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & \sigma_{pp} \end{pmatrix}$.

7.2 Sample PCs, loadings and scores

Up to here we considered PCs and loadings with known mean $\underline{\mu}$ and covariance matrix Σ.
We now make the transition to sample-based PCs. This means to replace $\underline{\mu}$ and Σ by the corresponding sample estimates:

$$\hat{\underline{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \qquad \widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\underline{X}_i - \hat{\underline{\mu}})(\underline{X}_i - \hat{\underline{\mu}})^T.$$

Accordingly, we set: $U =$ matrix of eigenvectors of $\widehat{\Sigma}$,
$\Lambda =$ diagonal matrix of eigenvalues of $\widehat{\Sigma}$.

$\Longrightarrow L = D^{-1/2} U \Lambda^{1/2}$ with $D = \text{diag}(\hat{\Sigma})$

Recall the centered and standardized data matrices introduced in Section 1.5:

$$X_c = X - \underline{1}_p \, \hat{\mu}^T \;, \; X_{st} = X_c \; D^{-1/2}$$

Sample PCs

Changing from the sample vectors $\underline{x}_i = (x_{i1}, \ldots, x_{ip})^T; i = 1, ..., n;$ which form the rows of the data matrix X, to the standardized sample principal component vectors

$$\Lambda^{-1/2}\underline{y}_i = \Lambda^{-1/2}U^T(\underline{x}_i - \underline{\hat{\mu}})$$

leads us to the so-called **scores matrix**

$$Y^{(s)} = (X - \underline{1}_p\, \underline{\hat{\mu}}^T)U\Lambda^{-1/2}.$$

The rows of this matrix contain the coordinates of the samples in the new PC-coordinate system, which is spanned by the orthonormal eigenvectors of $\widehat{\Sigma}$.

Theorem

The standardized data matrix X_{st} can be factorized as

$$X_{st} = Y^{(s)} * L^T = (\textit{scores matrix}) * (\textit{matrix of loadings})^T.$$

Factorization of X_{st}

Proof:

$$\begin{aligned} Y^{(s)} \cdot L^T &= (X - \underline{1}_n \, \hat{\underline{\mu}}^T) U \Lambda^{-1/2} \Lambda^{1/2} U^T D^{-1/2} \\ &= (X - \underline{1}_n \, \hat{\underline{\mu}}^T) U U^T D^{-1/2} \\ &= (X - \underline{1}_n \, \hat{\underline{\mu}}^T) D^{-1/2} \\ &= X_c D^{-1/2} = X_{st} \ . \end{aligned}$$

This decomposition will be used in the so-called **Biplot**-representations (double coordinate system comprised of scores and loadings) below.

7.3 Dimension reduction, number of relevant PCs

The reduction of the dimensionality of the data will be accomplished by deleting PCs which correspond to the smallest eigenvalues of $\hat{\Sigma}$, these carry the least information.
The decision about how many PCs could be neglected is often of subjective nature.

Choosing the number of PCs

In the sequel we will provide some (visual) aids to support the choice of a suitable number of PCs. Recall that the variability in our data is measured through

$$V_t = \sum_{i=1}^{p} \mathrm{Var}(Y_i) = \mathrm{tr}(\mathrm{Cov}(\underline{Y})) = \mathrm{tr}(\mathrm{Cov}(\underline{X})) = \sum_{i=1}^{p} \lambda_i.$$

We will focus on the main carriers of data variability, i.e. on the k largest eigenvalues, $1 \leq k < p$.

1) Choose k such that

$$V_k := \frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_k + ... + \lambda_p} \geq V_0$$

where V_0 is a prespecified percentage, customary choices are $V_0 = 0.9$ or $V_0 = 0.95$.

2) Sequential Likelihood Ratio Test: we test the hypothesis

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = ... = \lambda_p = 0$$

Choice of k

To test H_0, start with $k = 0$, then increase k, step by step and stop at the first value of k which leads to an acceptance of H_0.
As test statistic we use

$$T_R = (n-1)(p-k) \ln \left\{ m_a(R)/m_g(R) \right\} \overset{as.}{\approx} \chi^2_q$$

where R is the sample correlation matrix and

$m_a(R) = \frac{\lambda_{k+1}+\ldots+\lambda_p}{p-k}$ (arithmetic mean of the eigenvalues of R)

$m_g(R) = \sqrt[p-k]{\lambda_{k+1} \cdot \ldots \cdot \lambda_p}$ (geometric mean of the eigenvalues of R)

$q = \frac{1}{2}(p-k+2)(p-k-1)$ (degrees of freedom).

3) Delete all PCs Y_i, for which

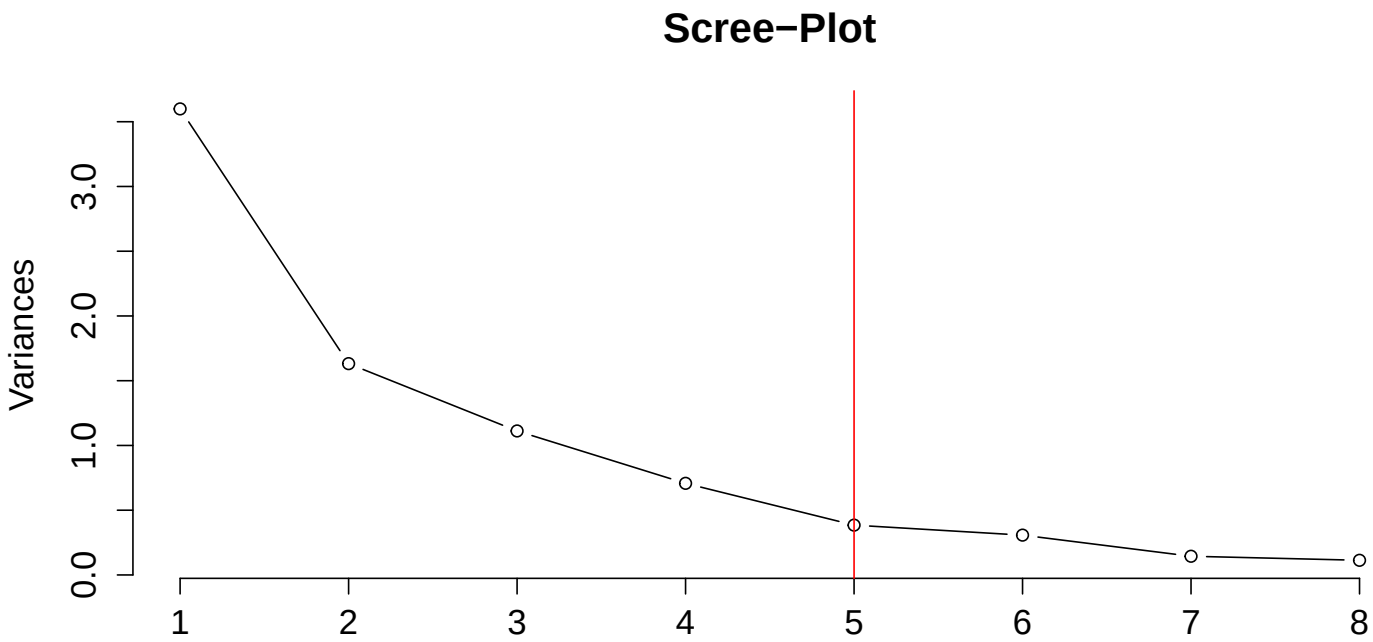
$$\lambda_i < \overline{\lambda} = \frac{1}{p} \sum_{j=1}^{p} \lambda_j.$$

If the sample correlation matrix R is used, instead of $\widehat{\Sigma}$, then this means: Delete all PCs Y_i for which $\lambda_i < 1$.

Choice of k

4) Scree-Plot (visual way of finding k):

Plots of the descending eigenvalues exhibit a typical pattern: first they strongly decrease and, from some threshold onwards, they only gradually decrease. We then see a clear breakpoint in the plot and neglect all the PCs corresponding to eigenvalues on the right-hand side of this point.



Example data set

7.4 R-Implementation

We analyze the data frame "state.x77".

```
> str(state.x77)
 num[1:50, 1:8] 3615 365 2212 2110 21198...
 - attr(, "dimnames")= List of 2
  ..: chr[1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
  ..: chr[1:8] "Population" "Income" "Illiteracy" "LifeExp" ...
> pca.state= prcomp(state.x77,scale=T)    # PC - object
> pca.state
Standard deviations (1, .., p=8):
[1]  1.8971 1.2775 1.0549 0.8411 0.6202 0.5545 0.3801 0.3364
Rotation (n x k) = (8 x 8):
                PC1       PC2       PC3  .....  PC8
Population  0.12643   0.41087  -0.65633  ..... 0.21925
Income     -0.29883   0.51898  -0.10036  .....-0.06029
```

Illiteracy	0.46767	0.05297	0.07090	..... 0.33869
Life Exp	-0.41161	-0.08166	-0.35993	.....-0.52743
Murder	0.44426	0.30695	0.10847	.....-0.67825
HS Grad	-0.42468	0.29877	0.04971	..... 0.30724
Frost	-0.35741	-0.15358	0.38711	.....-0.02834
Area	-0.03338	0.58762	0.51038	.....-0.01320

The command "$prcomp(data, scale = T)$" provides us with the standard deviations of the PCs and the matrix of loadings (Rotation). We recommend to work with the standardized data matrix X_{st} (scale=T).

The scores-matrix $Y^{(s)}$ can be accessed via

$>$ pca.state$x

	PC1	PC2	PC3	.....	PC8
Alabama	3.78989	-0.23478	0.22932	.....-0.53492	
Alaska	-1.05314	5.45618	4.24059	..... 0.11848	
Arizona	0.86743	0.74506	0.07727	..... 0.52455	

.....

West Virginia	1.50662	-1.60198	0.49219	.....	0.33394
Wisconsin	- 1.75754	-0.63573	-0.42537	.....	-0.04113
Wyoming	- 1.48379	0.04226	1.35421	.....	0.13053

The variances of the PCs are

```
> pca.state.var=pca.state$sdev^2
> pca.state.var
```

```
[1] 3.5989 1.6319 1.1119 0.7075 0.3846 0.3075 0.1445
[8] 0.1132
```

The cumulative proportions $V_k; k = 1, ..., 8;$ of the PC-variances read as follows:

```
> cumsum(pca.state.var)/sum(pca.state.var)
```

```
[1] 0.4499 0.6539 0.7928 0.8813 0.9294 0.9678 0.9859
[8] 1.0000
```

The cumulative proportions can also be accessed via the "summary" command:

```
> summary(pca.state)
```

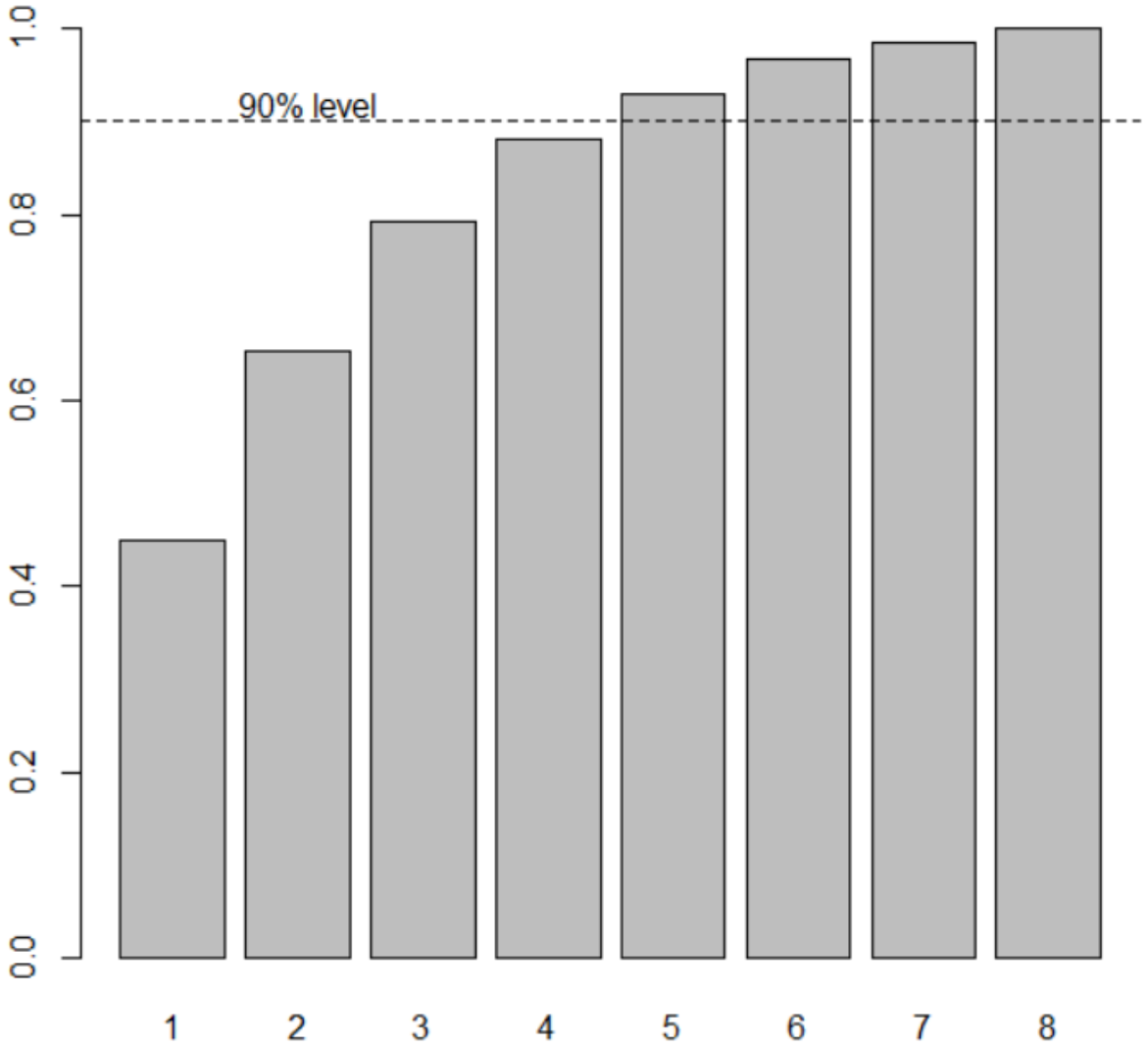
Importance of components:

	PC1	PC2	PC3	.....	PC7	PC8
Standard deviation	1.8971	1.2775	1.0545	.....	0.3801	0.3364
Proportion of Var.	0.4499	0.2040	0.1390	.....	0.0181	0.0142
Cumulative Prop.	0.4499	0.6539	0.7928	.....	0.9859	1.0000

Plot of the cumulative variance proportions:

```
> barplot(cumsum(pca.state.var)/sum(pca.state.var),
+ names=as.character(1:8))
> abline(h=0.9,lty=2)
> text(2, 0.92,"90% level")
```

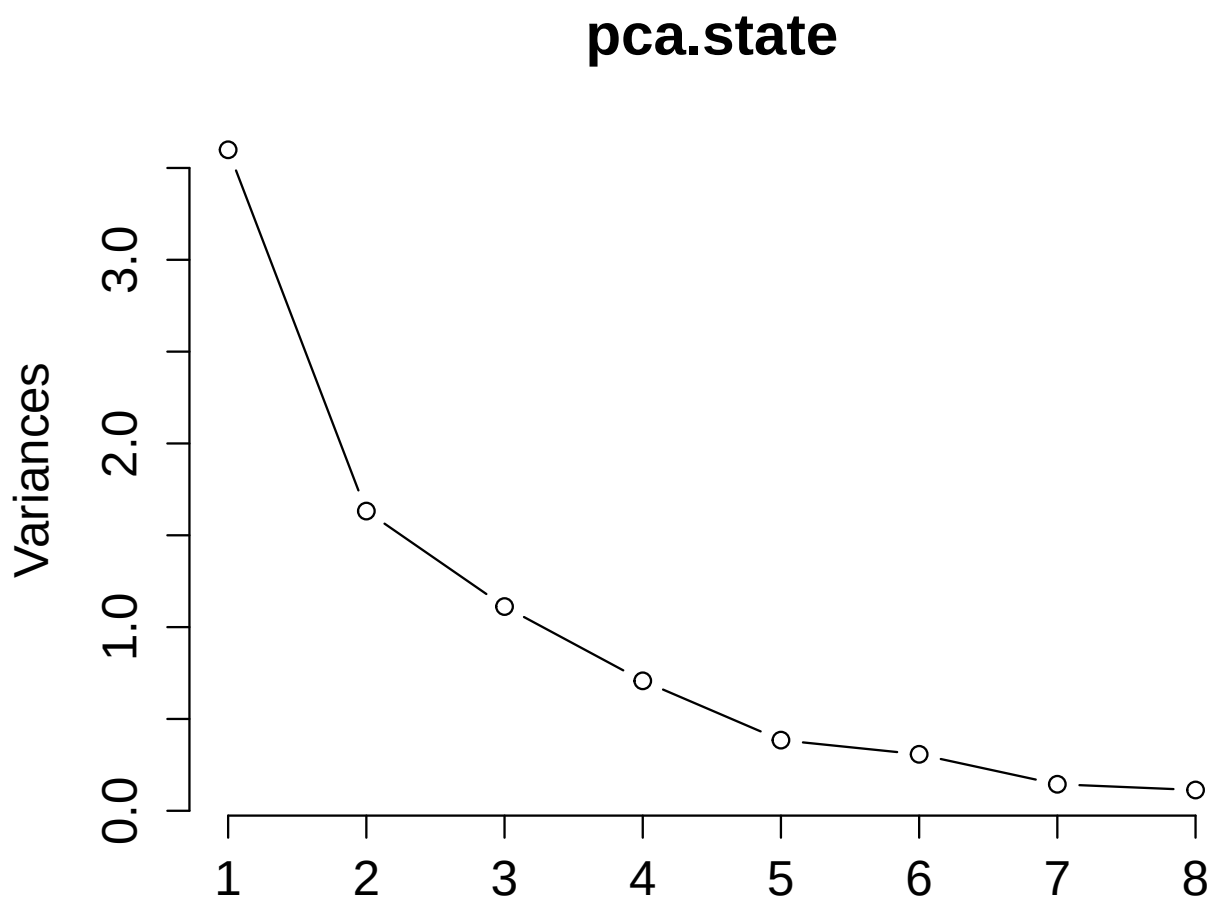
Cumulative variance proportions of the PCs



Scree-Plot of the PCs for state.x77

The scree-plot of the variances (eigenvalues) of the principal components can be generated easily via the "screeplot" command:

```
> screeplot(pca.state, type="l")
```



Biplot of scores and loadings

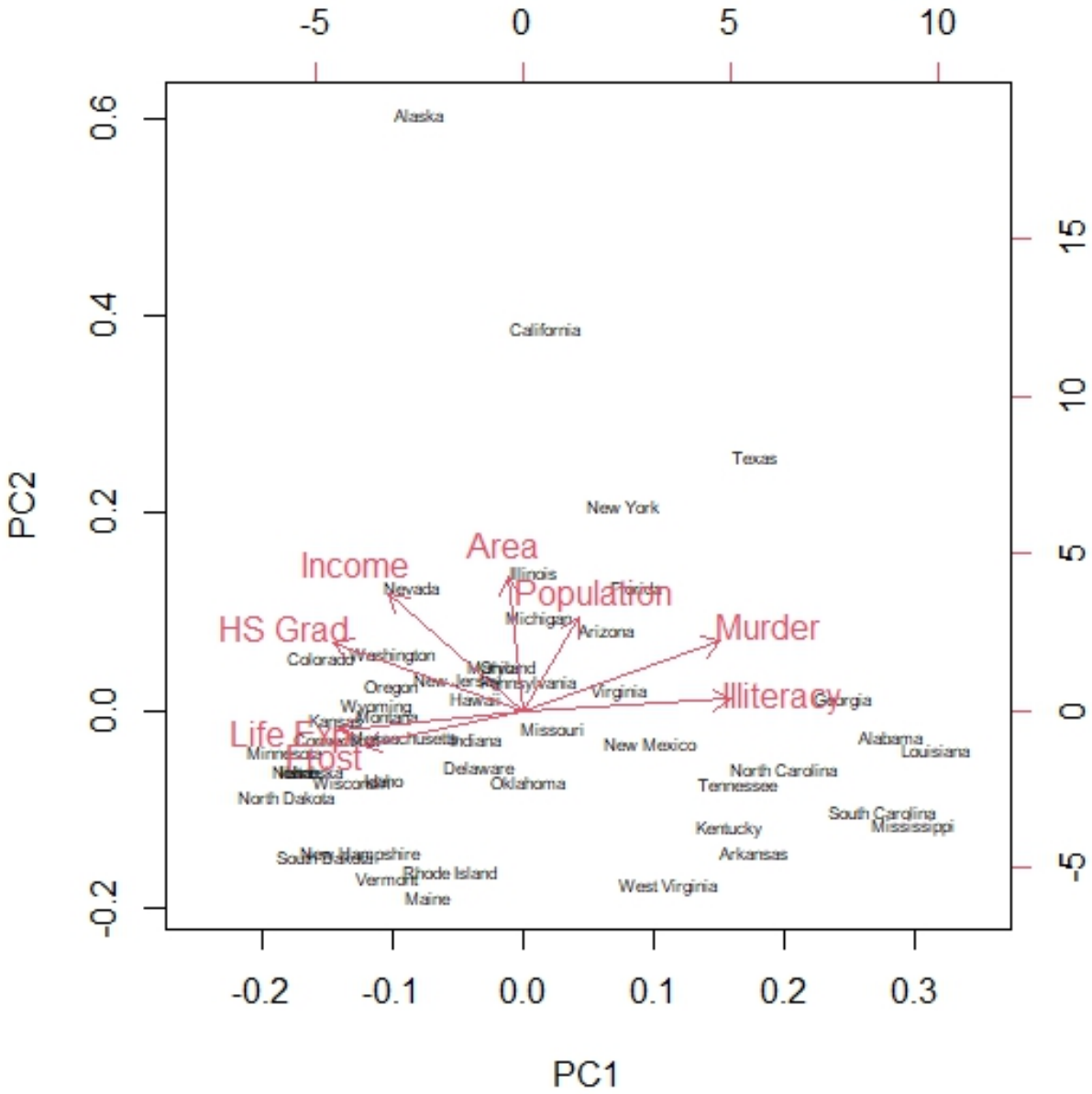
The biplot offers a joint graphical representation of the scores (representing the data objects) and loadings (representing the influence of variables in the PCs) in a double coordinate system.

For each sample, the first two principal components (PC1 and PC2) are plotted and, on the other hand, the loadings of PC1 and PC2 on the original variables, represented by vectors. Die $x-$coordinates represent the loadings of PC1 on the variables, the $y-$ coordinates represent the loadings of PC2 on the variables.

The length of the vectors characterize the importance of the variables. Vectors of approximately equal length and similar directions characterize variables with similar behaviour. On the other hand, sample points which are rather distant from each other, are highly dissimilar. Thus, very often, the biplot already indicates some grouping in the data. Also, potential outliers may be easily recognized.

```
> biplot(pca.state, cex=c(0.5,1), xlim=c(-0.25,0.35))
```

Biplot for the state.x77 data



```
> biplot(pca.state, cex=c(0.6,1.2), xlim=c(-0.25,0.35),
+ ylim=c(-0.2,0.25))
```