Multivariate Statistical Analysis

- 1. Aspects of Multivariate Analysis
- 2. Principal Components
- 3. Factor Analysis
- 4. Discrimination and Classification
- 5. Clustering

Johnson, R.A., Wichern, D.W. (1982): Applied Multivariate Statistical Analysis, Prentice Hall.

1. Aspects of Multivariate Analysis

Multivariate data arise whenever $p \geq 1$ variables are recorded. Values of these variables are observed for n distinct item, individuals, or experimental trials.

We use the notation x_{ij} to indicate the particular value of the ith variable that is observed on the jth item, or trial.

Thus, n measurements on p variables are displayed as $p \times n$ **random** matrix $\boldsymbol{X}$:

	Item 1	Item 2	...	Item j	...	Item n
Variable 1:	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}
Variable 2:	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Variable i:	x_{i1}	x_{i2}	...	x_{ij}	...	x_{in}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Variable p:	x_{p1}	x_{p2}	...	x_{pj}	...	x_{pn}

Estimating Moments:

Suppose, $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$ are the population moments. Based on a sample of size n, these quantities can be estimated by their empirical versions:

Sample Mean:

$$\overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}\,, \quad i = 1, \ldots, p$$

Sample Variance:

$$s_i^2 = s_{ii} = \frac{1}{n-1}\sum_{j=1}^{n} \left(x_{ij} - \overline{x}_i\right)^2, \quad i = 1, \ldots, p$$

Sample Covariance:

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n} \left(x_{ij} - \overline{x}_i\right)\left(x_{kj} - \overline{x}_k\right), \quad i = 1, \ldots, p\,, \quad k = 1, \ldots, p\,.$$

Summarize all elements s_{ik} into the $p \times p$ sample variance-covariance matrix $\boldsymbol{S} = \left(s_{ik}\right)_{i,k}$.

Assume further, that the $p \times p$ population correlation matrix $\boldsymbol{\rho}$ is estimated by the sample correlation matrix $\boldsymbol{R}$ with entries

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii} s_{kk}}}, \quad i = 1, \ldots, p, \quad k = 1, \ldots, p,$$

where $r_{ii} = 1$ for all $i = 1, \ldots, p$.

```
> aimu <- read.table("aimu.dat", header=TRUE)
> attach(aimu)
> options(digits=2)

> mean(aimu[ ,3:8])
   age height weight    fvc   fev1   fevp
    30    177     77    553    460     83
```

```
> cov(aimu[ ,3:8])
         age height weight  fvc fev1  fevp
age      110  -16.9   16.5 -233 -302 -20.8
height   -17   45.5   34.9  351  275  -1.9
weight    16   34.9  109.6  325  212  -7.6
fvc     -233  351.5  324.7 5817 4192 -86.5
fev1    -302  275.2  212.0 4192 4347 162.5
fevp     -21   -1.9   -7.6  -87  162  41.3

> cor(aimu[ ,3:8])
          age height weight   fvc  fev1   fevp
age      1.00 -0.239   0.15 -0.29 -0.44 -0.309
height -0.24  1.000   0.49  0.68  0.62 -0.043
weight  0.15  0.494   1.00  0.41  0.31 -0.113
fvc    -0.29  0.683   0.41  1.00  0.83 -0.177
fev1   -0.44  0.619   0.31  0.83  1.00  0.384
fevp   -0.31 -0.043  -0.11 -0.18  0.38  1.000
```

Distances:

Consider the point $P = (x_1, x_2)$ in the plane. The straight line (Euclidian) distance, $d(O, P)$, from P to the origin $O = (0, 0)$ is (Pythagoras)

$$d(O, P) = \sqrt{x_1^2 + x_2^2}\,.$$

In general, if P has p coordinates so that $P = (x_1, x_2, \ldots, x_p)$, the Euclidian distance is

$$d(O, P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}\,.$$

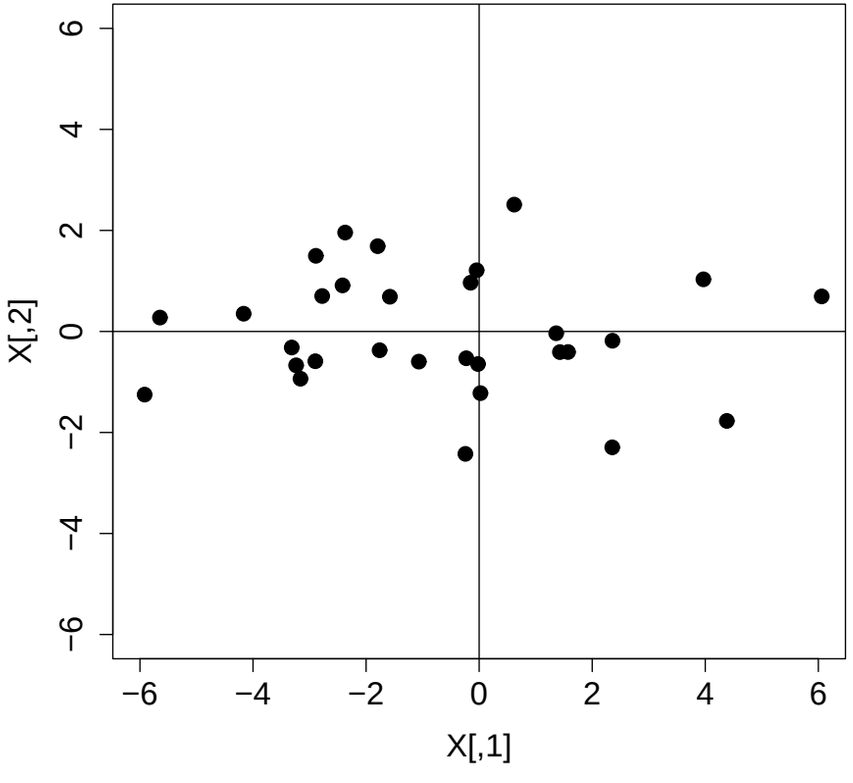
The distance between 2 arbitrary points P and $Q = (y_1, y_2, \ldots, y_p)$ is given by

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}\,.$$

Each coordinate contributes equally to the calculation of the Euclidian distance. It is often desirable to weight the coordinates.

Statistical distance should account for differences in variation and correlation. Suppose we have n pairs of measurements on 2 independent variables x_1 and x_2:

```
> X <- mvrnorm(30, mu=c(0, 0), Sigma=matrix(c(9,0,0,1), 2, 2));  plot(X)
```



Variability in x_1 direction is much larger than in x_2 direction! Values that are a given deviation from the origin in the x_1 direction are not as $surprising$ as are values in x_2 direction.

It seems reasonable to weight an x_2 coordinate more heavily than an x_1 coordinate of the same value when computing the distance to the origin.

Compute the statistical distance from the standardized coordinates

$$x_1^* = \frac{x_1}{\sqrt{s_{11}}} \qquad \text{and} \qquad x_2^* = \frac{x_2}{\sqrt{s_{22}}}$$

as

$$d(O,P) = \sqrt{(x_1^*)^2 + (x_2^*)^2} = \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}\,.$$

This can be generalized to accommodate the calculation of statistical distance from an arbitrary point $P = (x_1, x_2)$ to any *fixed* point $Q = (y_1, y_2)$. If the coordinate variables vary independent of one other, the distance from P to Q is

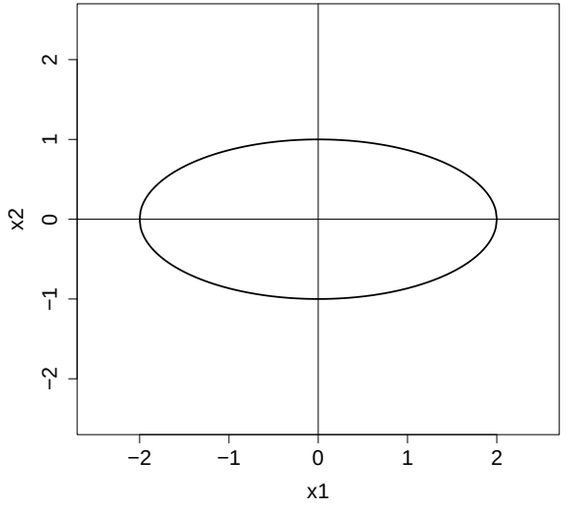
$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}}\,.$$

The extension to more than 2 dimensions is straightforward.

Let $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$. Assume again that Q is fixed. The statistical distance from P to Q is

$$d(P, Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}.$$

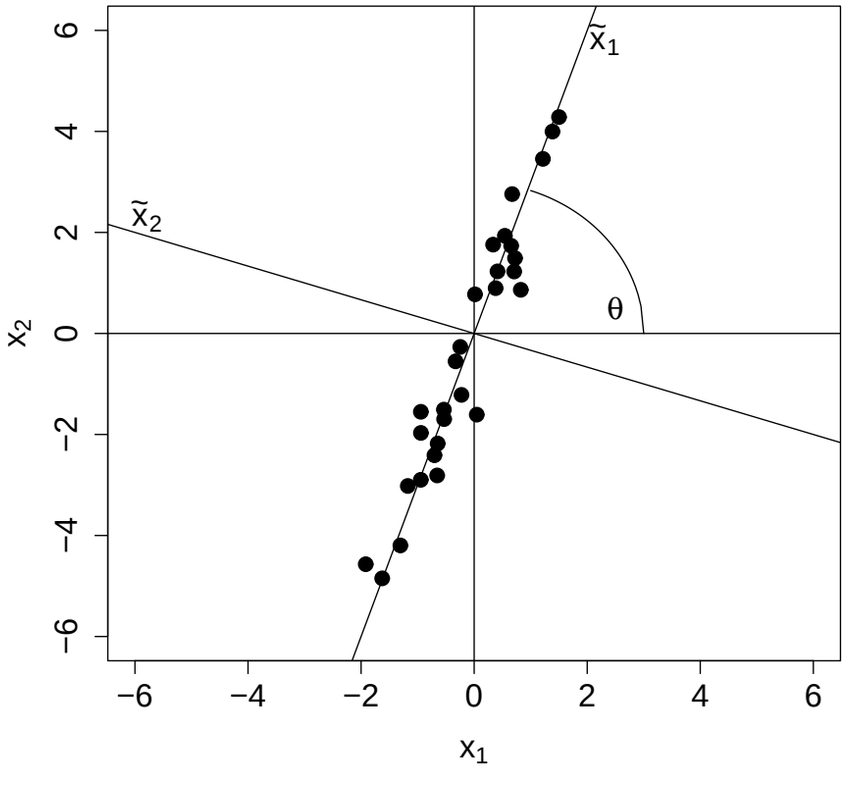
- The distance of P to the origin is obtained by setting $y_1 = y_2 = \cdots = y_p = 0$.
- If $s_{11} = s_{22} = \cdots = s_{pp}$, the Euclidian distance is appropriate.



Consider a set of paired measurements (x_1, x_2) with $\overline{x}_1 = \overline{x}_2 = 0$, and $s_{11} = 4$, $s_{22} = 1$. Suppose the x_1 measurements are unrelated to the x_2 ones. We measure the squared distance of an arbitrary $P = (x_1, x_2)$ to $(0, 0)$ by $d^2(O, P) = x_1^2/4 + x_2^2/1$. All points with constant distance 1 satisfy: $x_1^4/4 + x_2^2/1 = 1$, an Ellipse centered at $(0, 0)$.

This definition of statistical distance still does not include most of the important cases because of the assumption of independent coordinates.

```
> X <- mvrnorm(30, mu=c(0, 0), Sigma=matrix(c(1,2.9,2.9,9), 2, 2))
> plot(X);  abline(h=0, v=0);  abline(0, 3);  abline(0, -1/3)
```



Here, the x_1 measurements do not vary independently of x_2. The coordinates exhibit a tendency to be large or small together. Moreover, the variability in the x_2 directions is larger than in x_1 direction.

What is a meaningful measure of distance? Actually, we can use what we have already introduced!

But before, we only have to rotate the coordinate system through the angle θ and label the rotated axes $\tilde{x}_1$ and $\tilde{x}_2$.

Now, we define the distance of a point $P = (x_1, x_2)$ from the origin $(0, 0)$ as

$$d(O, P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}},$$

where $\tilde{s}_{ii}$ denotes the sample variance computed with the (rotated) $\tilde{x}_i$ measurements.

Alternative measures of distance can be useful, provided they satisfy the properties

1. $d(P, Q) = d(Q, P)$,
2. $d(P, Q) > 0$ if $P \neq Q$,
3. $d(P, Q) = 0$ if $P = Q$,
4. $d(P, Q) \leq d(P, R) + d(R, Q)$, R being any other point different to P and Q.

Principle Components (PCA)

Now we try to explain the variance-covariance structure through a few **linear** combinations of the original p variables $X_1, X_2, \ldots, X_p$ (data reduction).

Let a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^t$ have $p \times p$ population variance-covariance matrix $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

Denote the eigenvalues of $\boldsymbol{\Sigma}$ by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the arbitrary linear combinations with fixed vectors $\boldsymbol{\ell}_i$

$$\begin{aligned}
Y_1 &= \boldsymbol{\ell}_1^t \boldsymbol{X} = \ell_{11} X_1 + \ell_{21} X_2 + \cdots + \ell_{p1} X_p \\
Y_2 &= \boldsymbol{\ell}_2^t \boldsymbol{X} = \ell_{12} X_1 + \ell_{22} X_2 + \cdots + \ell_{p2} X_p \\
&\vdots \qquad \vdots \\
Y_p &= \boldsymbol{\ell}_p^t \boldsymbol{X} = \ell_{1p} X_1 + \ell_{2p} X_2 + \cdots + \ell_{pp} X_p
\end{aligned}$$

For these

$$\begin{aligned} \mathsf{var}(Y_i) &= \mathsf{var}(\boldsymbol{\ell}_i^t \boldsymbol{X}) = \boldsymbol{\ell}_i^t \boldsymbol{\Sigma} \boldsymbol{\ell}_i \\ \mathsf{cov}(Y_i, Y_k) &= \mathsf{cov}(\boldsymbol{\ell}_i^t \boldsymbol{X}, \boldsymbol{\ell}_k^t \boldsymbol{X}) = \boldsymbol{\ell}_i^t \boldsymbol{\Sigma} \boldsymbol{\ell}_k \end{aligned}$$

We define as **principal components** those linear combinations $Y_1, Y_2, \ldots, Y_p$, which are **uncorrelated** and whose variances are as **large** as possible.

Since increasing the length of $\boldsymbol{\ell}_i$ would also increase the variances, we restrict our search onto vectors $\boldsymbol{\ell}_i$, which are of unit length, i.e. $\sum_j \ell_{ij}^2 = \boldsymbol{\ell}_i^t \boldsymbol{\ell}_i = 1$.

Procedure:

1. the first principal component is the linear combination $\boldsymbol{\ell}_1^T \boldsymbol{X}$ that maximizes var$(\boldsymbol{\ell}_1^t \boldsymbol{X})$ subject to $\boldsymbol{\ell}_1^t \boldsymbol{\ell}_1 = 1$.

2. the second principal component is the linear combination $\boldsymbol{\ell}_2^T \boldsymbol{X}$ that maximizes var$(\boldsymbol{\ell}_2^t \boldsymbol{X})$ subject to $\boldsymbol{\ell}_2^t \boldsymbol{\ell}_2 = 1$ and with cov$(\boldsymbol{\ell}_1^t \boldsymbol{X}, \boldsymbol{\ell}_2^t \boldsymbol{X}) = 0$ (uncorrelated with the first one).

3. the ith principal component is the linear combination $\boldsymbol{\ell}_i^T \boldsymbol{X}$ that maximizes var$(\boldsymbol{\ell}_i^t \boldsymbol{X})$ subject to $\boldsymbol{\ell}_i^t \boldsymbol{\ell}_i = 1$ and with cov$(\boldsymbol{\ell}_i^t \boldsymbol{X}, \boldsymbol{\ell}_k^t \boldsymbol{X}) = 0$, for $k < i$ (uncorrelated with all the previous ones).

How to find all these vectors $\boldsymbol{\ell}_i$?

We will use well known some results from matrix theory.

Result 1: Let $\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$ and let $\boldsymbol{\Sigma}$ have the eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \ldots, (\lambda_p, \boldsymbol{e}_p)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the ith principal component, $i = 1, \ldots, p$, is given by

$$Y_i = \boldsymbol{e}_i^t \boldsymbol{X} = e_{1i} X_1 + e_{2i} X_2 + \ldots + e_{pi} X_p \,.$$

With this choices

$$\begin{aligned} \mathrm{var}(Y_i) &= \boldsymbol{e}_i^t \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i \,, \\ \mathrm{cov}(Y_i, Y_k) &= \boldsymbol{e}_i^t \boldsymbol{\Sigma} \boldsymbol{e}_k = 0 \,. \end{aligned}$$

Thus, the principal components are uncorrelated and have variances equal to the eigenvalues of $\boldsymbol{\Sigma}$.

If some λ_i are equal, the choice of the corresponding coefficient vectors $\boldsymbol{e}_i$, and hence Y_i, are not unique.

Result 2: Let $Y_1 = \boldsymbol{e}_1^t \boldsymbol{X}$, $Y_2 = \boldsymbol{e}_2^t \boldsymbol{X}$, ..., $Y_p = \boldsymbol{e}_p^t \boldsymbol{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \text{var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{var}(Y_i)\,.$$

Thus, the **total population variance** equals the sum of the eigenvalues. Consequently, the proportion of total variance due to (explained by) the kth principal component is

$$0 < \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} < 1$$

If most (e.g. 80 to 90%) of the total population variance (for large p) can be attributed to the first one, two, or three principal components, then these components can $replace$ the original p variables without much loss of information.

The magnitude of e_{ik} measures the importance of the kth variable to the ith principal component. In particular, e_{ik} is proportional to the correlation coefficient between Y_i and X_k.

Result 3: If $Y_1 = \boldsymbol{e}_1^t \boldsymbol{X}$, $Y_2 = \boldsymbol{e}_2^t \boldsymbol{X}$, ..., $Y_p = \boldsymbol{e}_p^t \boldsymbol{X}$ are the principal components from the variance-covariance matrix $\boldsymbol{\Sigma}$, then

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

are the correlation coefficients between the components Y_i and the variables X_k.

It is informative to consider principal components derived from multivariate normal random variables. Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ having density function

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) .$$

Then the $\boldsymbol{\mu}$ centered ellipsoids of constant density are

$$(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2 .$$

In the two-dimensional case $\boldsymbol{x} = (x_1, x_2)^t$ this equals

$$\frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right] = c^2 .$$

These ellipsoids have axes $\pm c\sqrt{\lambda_i}\boldsymbol{e}_i$, $i = 1, \ldots, p$.

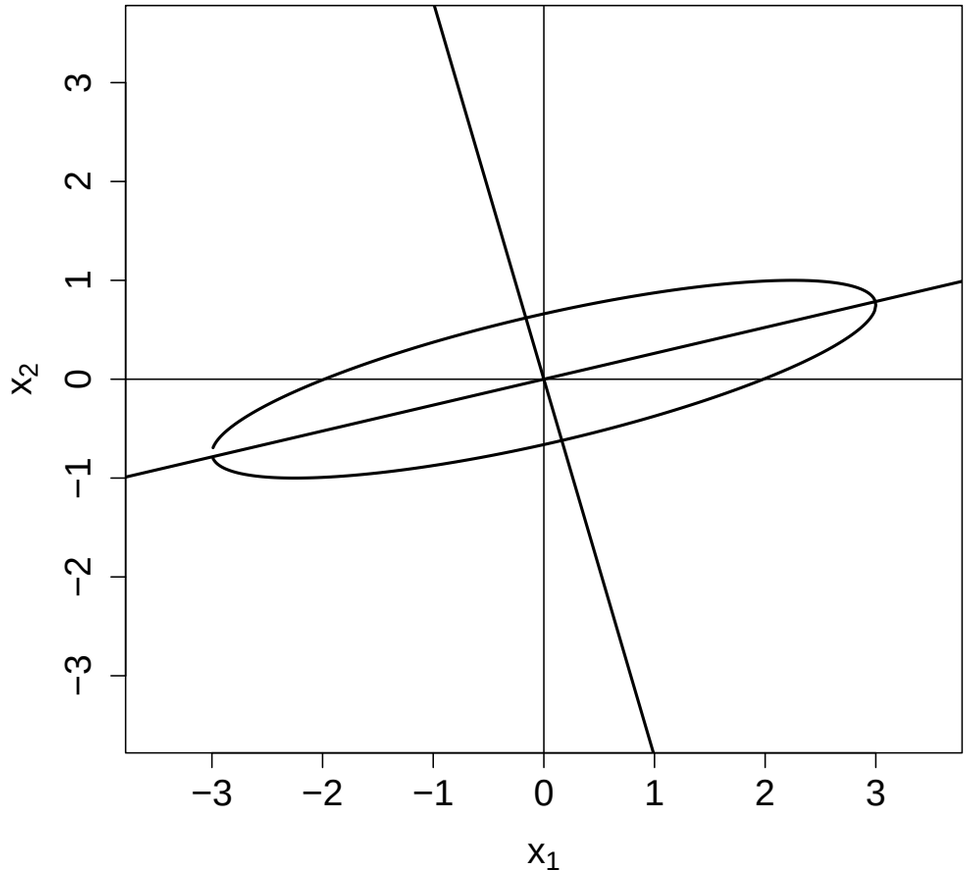
Example: Suppose $\boldsymbol{x} = (x_1, x_2)^t \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0, 0)^t$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} = 9 & \sigma_{12} = 9/4 \\ \sigma_{21} = 9/4 & \sigma_{22} = 1 \end{pmatrix}$$

giving $\rho_{12} = (9/4)/\sqrt{9 \cdot 1} = 3/4$.

The eigen-analysis of $\boldsymbol{\Sigma}$ results in

```
> sigma <- matrix(c(9, 9/4, 9/4, 1), 2, 2)
> e <- eigen(sigma, symmetric=TRUE);  e
 $values
 [1] 9.58939  0.41061

 $vectors
          [,1]     [,2]
 [1,] -0.96736  0.25340
 [2,] -0.25340 -0.96736
```




```
# check length of eigenvectors
> e$vectors[2,1]^2+e$vectors[1,1]^2
 [1] 1
> e$vectors[2,2]^2+e$vectors[1,2]^2
 [1] 1

# slopes of major & minor axes
> e$vectors[2,1]/e$vectors[1,1]
 [1] 0.2619511
> e$vectors[2,2]/e$vectors[1,2]
 [1] -3.817507

# endpoints of of major&minor axes
> sqrt(e$values[1])*e$vectors[,1]
 [1] -2.9956024 -0.7847013
> sqrt(e$values[2])*e$vectors[,2]
 [1]  0.1623767 -0.6198741
```

These results also hold for $p \geq 2$. Set $\boldsymbol{\mu} = \mathbf{0}$ in what follows.

$$\begin{aligned} c^2 = \boldsymbol{x}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x} &= \frac{1}{\lambda_1}(\boldsymbol{e}_1^t \boldsymbol{x})^2 + \frac{1}{\lambda_2}(\boldsymbol{e}_2^t \boldsymbol{x})^2 + \cdots + \frac{1}{\lambda_p}(\boldsymbol{e}_p^t \boldsymbol{x})^2 \,, \\ &= \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2 \end{aligned}$$

and this equation defines an ellipsoid (since the λ_i are positive) in a coordinate system with axes $y_1, y_2, \ldots, y_p$ lying in the directions of $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$. If λ_1 is the largest eigenvalue, then the major axes lies in the direction of $\boldsymbol{e}_1$. The remaining minor axes lie in the directions defined by $\boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$. Thus the principal components lie in the directions of the axes of the constant density ellipsoid.

Principal Components obtained from Standardized Variables

Instead of using $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^t$ we now calculate the principal components from $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)^t$, where

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \, .$$

In matrix notation this equals

$$\boldsymbol{Z} = \left(\boldsymbol{V}^{1/2}\right)^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \, ,$$

where the diagonal standard deviation matrix $\boldsymbol{V}^{1/2}$ is defined as

$$\boldsymbol{V}^{1/2} = \begin{pmatrix} \sqrt{\sigma_{11}} & & \\ & \ddots & \\ & & \sqrt{\sigma_{pp}} \end{pmatrix} \, .$$

Clearly E($\boldsymbol{Z}$) $= \mathbf{0}$ and var($\boldsymbol{Z}$) $= (\boldsymbol{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\boldsymbol{V}^{1/2})^{-1} = \boldsymbol{\rho}$.

Principal Components of $\boldsymbol{Z}$ will be obtained from the eigenvalues λ_i and eigenvectors $\boldsymbol{e}_i$ of $\boldsymbol{\rho}$ of $\boldsymbol{X}$. Theses are, in general, **not the same** as the ones derived from $\boldsymbol{\Sigma}$.

Result 4: The ith principal component of the standardized variables $\boldsymbol{Z}$ with var($\boldsymbol{Z}$) $= \boldsymbol{\rho}$ is given by

$$Y_i = \boldsymbol{e}_i^t \boldsymbol{Z} = \boldsymbol{e}_i^t \left(\boldsymbol{V}^{1/2}\right)^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \,, \qquad i = 1, \ldots, p \,.$$

Moreover,

$$\sum_{i=1}^{p} \text{var}(Y_i) = \sum_{i=1}^{p} \text{var}(Z_i) = p \,.$$

Thus, the proportion explained by the kth principal component is λ_k/p and

$$\rho_{Y_i, Z_k} = \boldsymbol{e}_{ki}\sqrt{\lambda_i} \,.$$

Example cont'ed: Let again $\boldsymbol{x} = (x_1, x_2)^t \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0, 0)^t$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 9 & 9/4 \\ 9/4 & 1 \end{pmatrix} \qquad \Longrightarrow \qquad \boldsymbol{\rho} = \begin{pmatrix} 1 & 3/4 \\ 3/4 & 1 \end{pmatrix}.$$

The eigen-analysis of $\boldsymbol{\rho}$ now results in:

```
> rho <- matrix(c(1, 3/4, 3/4, 1), 2, 2)
> e <- eigen(rho, symmetric=TRUE);  e
 $values
 [1] 1.75 0.25

 $vectors
         [,1]     [,2]
 [1,] 0.70711  0.70711
 [2,] 0.70711 -0.70711
```

The total population variance is $p = 2$, and $1.75/2 = 87.5\%$ of this variance is already explained by the first principal component.

The principal components from $\boldsymbol{\rho}$ are

$$\begin{aligned} Y_1 &= 0.707Z_1 + 0.707Z_2 = 0.707\frac{X_1}{3} + 0.707\frac{X_2}{1} = 0.236X_1 + 0.707X_2 \\ Y_2 &= 0.707Z_1 - 0.707Z_2 = 0.707\frac{X_1}{3} - 0.707\frac{X_2}{1} = 0.236X_1 - 0.707X_2\,, \end{aligned}$$

whereas those from $\boldsymbol{\Sigma}$ have been

$$\begin{aligned} Y_1 &= -0.967X_1 - 0.253X_2 \\ Y_2 &= +0.253X_1 - 0.967X_2\,. \end{aligned}$$

The important first component has explained $9.589/10 = 95.6\%$ of the total variability and is dominated by X_1 (because of its large variance). When the variables are standardized however, the resulting variables contribute equally to the principal components. Variables should be standardized, if they are measured on very different scales.

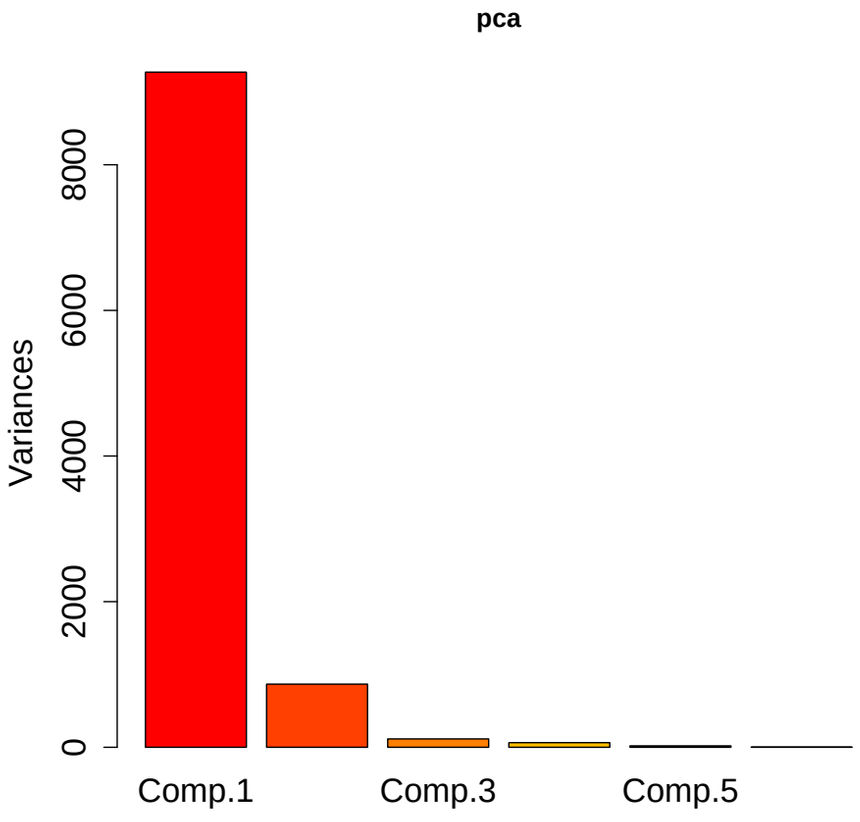
Summarizing Sample Variation by Principal Components

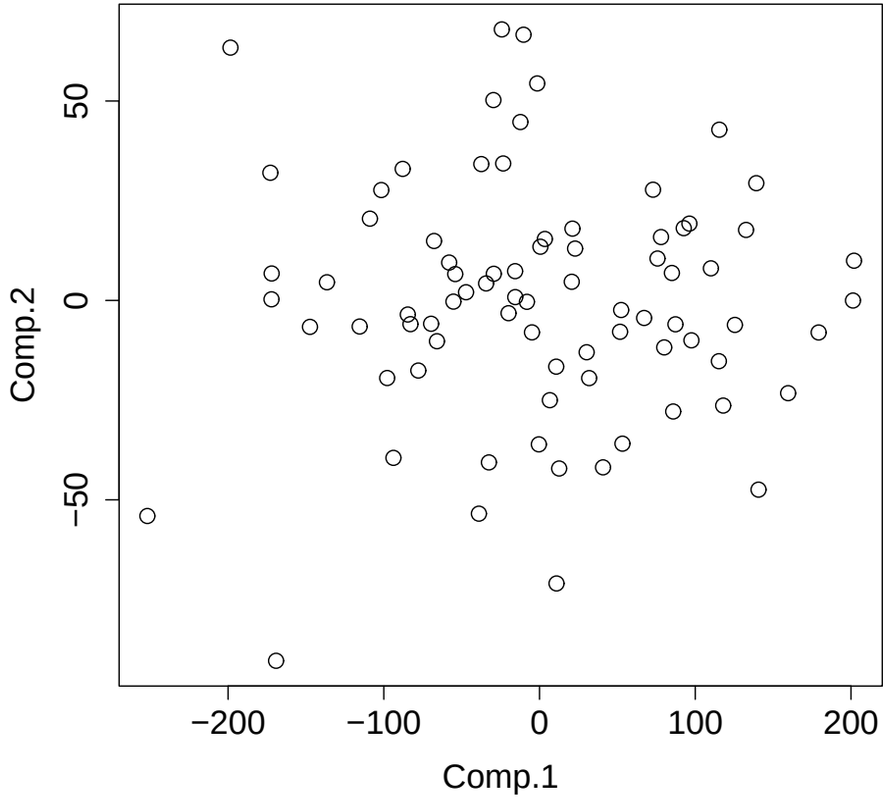
So far we have dealt with population means $\boldsymbol{\mu}$ and variances $\boldsymbol{\Sigma}$. If we analyze a sample then we have to replace $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ by their empirical versions $\boldsymbol{S}$ and $\overline{\boldsymbol{x}}$. The eigenvalues and eigenvectors are then based on $\boldsymbol{S}$ or $\boldsymbol{R}$ instead of $\boldsymbol{\Sigma}$ or $\boldsymbol{\rho}$.

```
> library(mva)
> attach(aimu)
> options(digits=2)
> pca <- princomp(aimu[ , 3:8])
> summary(pca)
 Importance of components:
                         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  Comp.6
 Standard deviation         96.3 29.443 10.707 7.9581 4.4149 1.30332
 Proportion of Variance      0.9  0.084  0.011 0.0061 0.0019 0.00016
 Cumulative Proportion       0.9  0.981  0.992 0.9980 0.9998 1.00000

> pca$center  # the means that were subtracted
   age height weight    fvc   fev1   fevp
    30    177     77    553    460     83
```

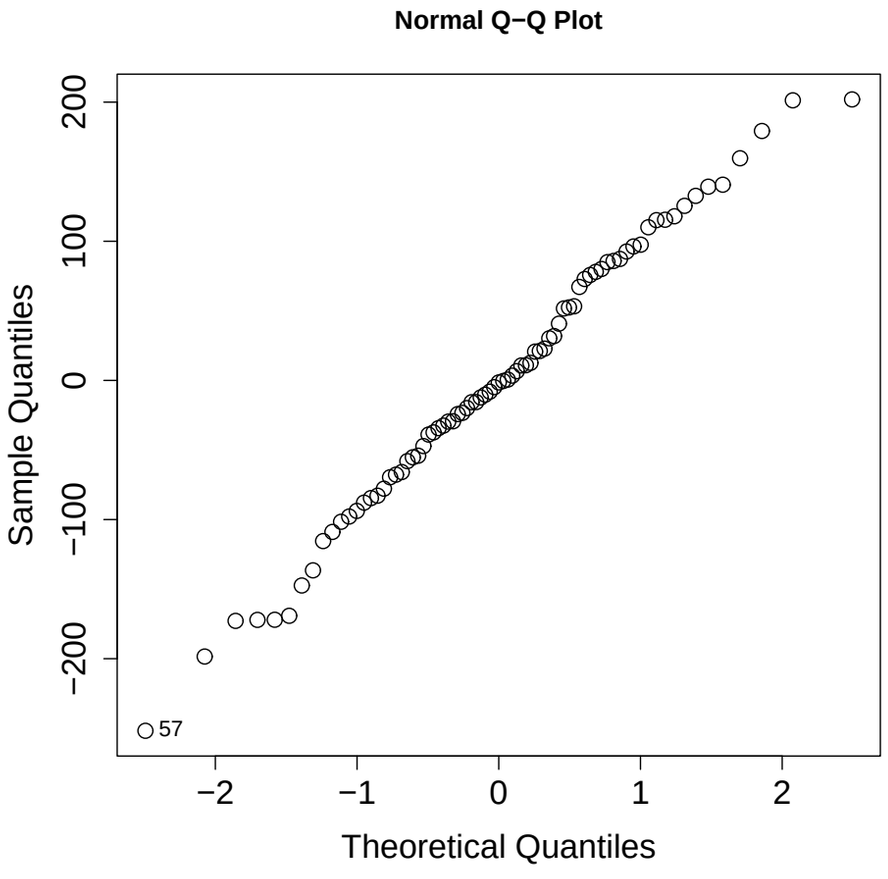
```
> pca$scale  # the scalings applied to each variable
   age height weight    fvc   fev1   fevp
     1      1      1      1      1      1

> pca$loadings #  a matrix whose columns contain the eigenvectors

 Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
 age            -0.109  0.645  0.747  0.110
 height                 0.119 -0.246  0.960
 weight                 0.745 -0.613 -0.251
 fvc    -0.763 -0.624                  0.133
 fev1   -0.641  0.741                 -0.164
 fevp           0.212                  0.976

                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
 SS loadings      1.00   1.00   1.00   1.00   1.00   1.00
 Proportion Var   0.17   0.17   0.17   0.17   0.17   0.17
 Cumulative Var   0.17   0.33   0.50   0.67   0.83   1.00
```

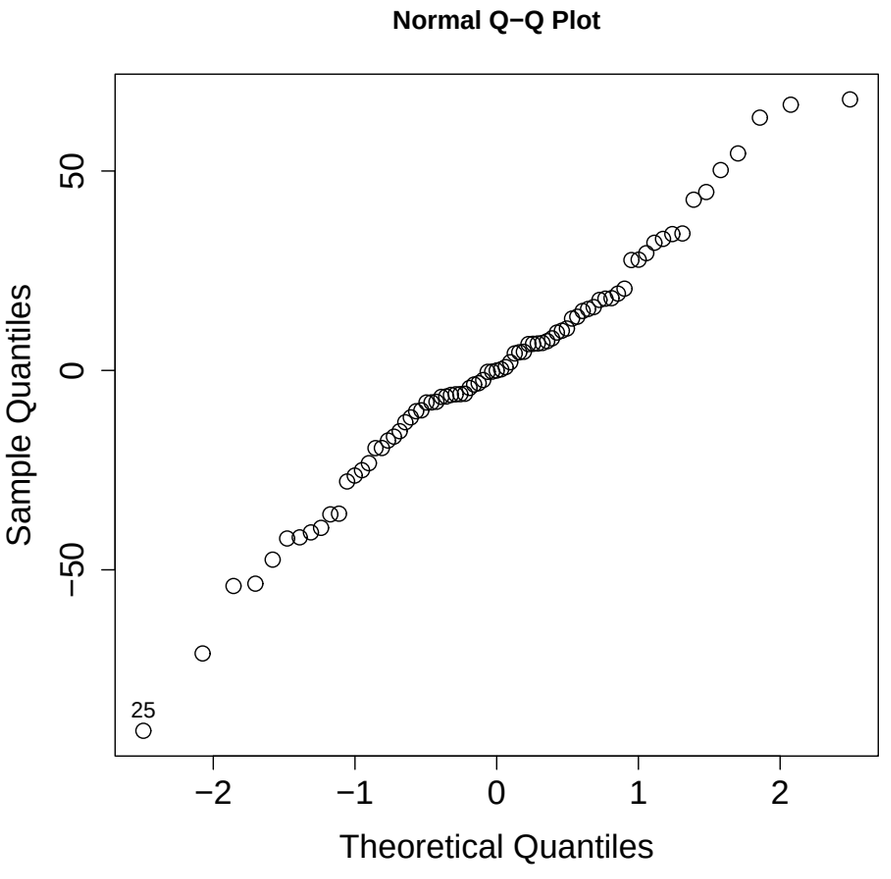
```
> pca$scores  # values of the p principal components for each observation

     Comp.1  Comp.2 Comp.3  Comp.4  Comp.5  Comp.6
1    22.84  12.998   4.06  13.131  -1.908  0.0408
2  -147.40  -6.633  -5.14  14.009  -2.130 -0.2862
3   159.64 -23.255   9.60   0.059   5.372 -0.8199
:
78   52.42  -2.409   1.68   9.169   3.716  0.6386
79  -82.87  -5.951   7.82  11.068   0.834 -0.4171

> plot(pca) # or screeplot(pca)
> plot(pca$scores[ , 1:2])
> identify(qqnorm(pca$scores[, 1]));  identify(qqnorm(pca$scores[, 2]))
```





Observations 57 and 25 are a bit outside the ellipsoid.

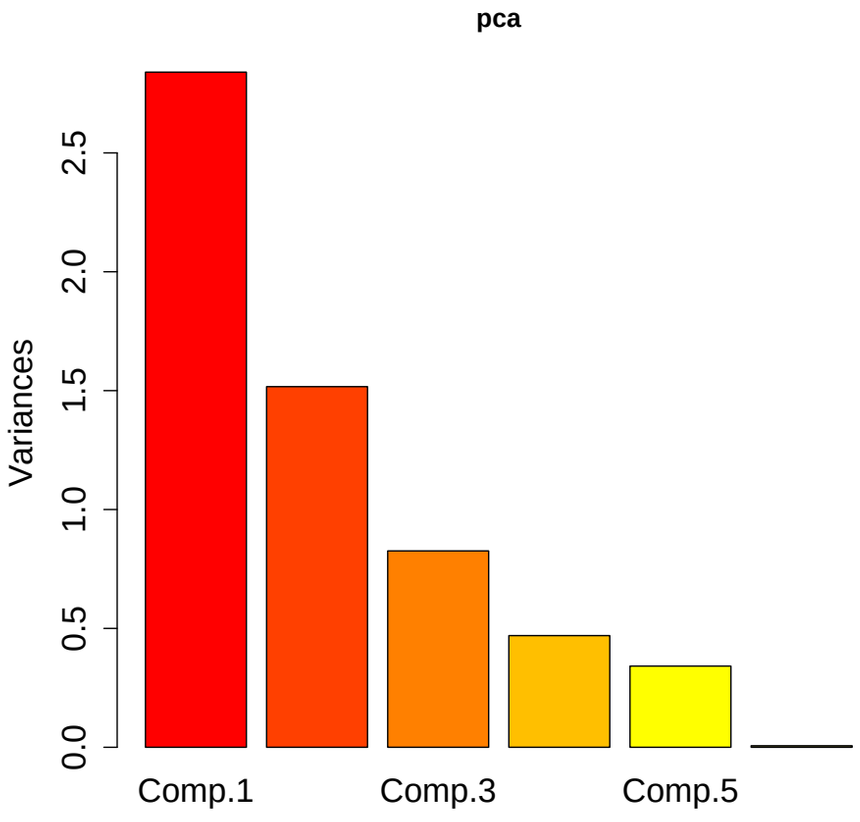

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
200
100
0
−100
−200
−2
−1
0
1
2
57

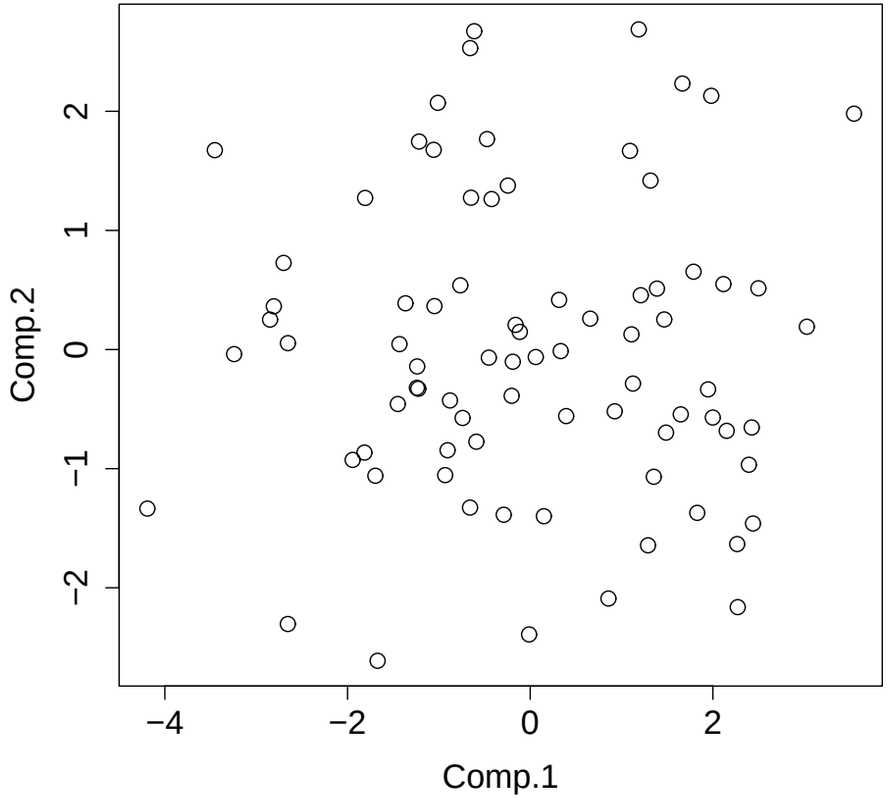

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
50
0
−50
−2
−1
0
1
2
25

If we base the analysis on the sample correlation matrix, we get

```
> pca <- princomp(aimu[ , 3:8], cor=TRUE)
> summary(pca)
 Importance of components:
                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
 Standard deviation         1.69   1.23   0.91  0.685  0.584 0.0800
 Proportion of Variance     0.47   0.25   0.14  0.078  0.057 0.0011
 Cumulative Proportion      0.47   0.73   0.86  0.942  0.999 1.0000

> pca$center
   age height weight    fvc   fev1   fevp
    30    177     77    553    460     83

> pca$scale
   age height weight    fvc   fev1   fevp
  10.4    6.7   10.4   75.8   65.5    6.4
```

```
> pca$loadings
 Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
 age     0.264 -0.535  0.446  0.633  0.211
 height -0.497 -0.172        -0.207  0.824
 weight -0.316 -0.449  0.541 -0.494 -0.402
 fvc    -0.534 -0.149 -0.278  0.373 -0.270  0.635
 fev1   -0.540  0.217         0.411 -0.168 -0.674
 fevp           0.643  0.650         0.110  0.375

                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
 SS loadings      1.00   1.00   1.00   1.00   1.00   1.00
 Proportion Var   0.17   0.17   0.17   0.17   0.17   0.17
 Cumulative Var   0.17   0.33   0.50   0.67   0.83   1.00
```





Apart from observations 57 and 25 the plot appears to be reasonable elliptical.

Factor Analysis

Purpose of this (controversial) technique is to describe (if possible) the covariance relationships among many variables in terms of a few underlying but unobservable, random quantities called **factors**.

Suppose variables can be grouped by their correlations. All variables within a group are highly correlated among themselves but have small correlations with variables in a different group. It is conceivable that each such group represents a single underlying construct (factor), that is responsible for the correlations.

E.g., correlations from the group of test scores in French, English, Mathematics suggest an underlying $intelligence\ factor$. A second group of variables representing $physical\ fitness\ scores$ might correspond to another factor.

Factor analysis can be considered as an extension of principal component analysis. Both attempt to approximate the covariance matrix $\mathbf{\Sigma}$.

The Orthogonal Factor Model

The $p \times 1$ random vector $\boldsymbol{X}$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model postulates that $\boldsymbol{X}$ linearly depend on some unobservable random variables $F_1, F_2, \ldots, F_m$, called *common factors* and p additional sources of variation $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$, called *errors* or sometimes *specific factors*.

The factor analysis model is

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \epsilon_2 \\ \vdots &\quad \vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \epsilon_p \end{aligned}$$

or in matrix notation

$$\underbrace{\boldsymbol{X} - \boldsymbol{\mu}}_{(p\times 1)} = \underbrace{\boldsymbol{L}}_{(p\times m)} \underbrace{\boldsymbol{F}}_{(m\times 1)} + \underbrace{\boldsymbol{\epsilon}}_{(p\times 1)} .$$

The coefficient ℓ_{ij} is called *loading* of the ith variable on the jth factor, so $\boldsymbol{L}$ is the matrix of factor loadings. Notice, that the p deviations $X_i - \mu_i$ are expressed in terms of $p + m$ random variables $F_1, \ldots, F_m$ and $\epsilon_1, \ldots, \epsilon_p$, which are all **unobservable**. (This distinguishes the factor model from a regression model, where the explanatory variables F_j can be observed.)

There are too many unobservable quantities in the model. Hence we need further assumptions about $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$. We assume that

$$\mathsf{E}(\boldsymbol{F}) = \mathbf{0}\,, \qquad \mathsf{cov}(\boldsymbol{F}) = \mathsf{E}(\boldsymbol{F}\boldsymbol{F}^t) = \boldsymbol{I}$$

$$\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}\,, \qquad \mathsf{cov}(\boldsymbol{\epsilon}) = \mathsf{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t) = \boldsymbol{\psi} = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ 0 & 0 & \ldots & \psi_p \end{pmatrix}.$$

and $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent, so

$$\mathsf{cov}(\boldsymbol{\epsilon}, \boldsymbol{F}) = \mathsf{E}(\boldsymbol{\epsilon}\boldsymbol{F}^t) = \mathbf{0}\,.$$

This defines the **orthogonal factor model** and implies a covariance structure for $\boldsymbol{X}$. Because of

$$\begin{aligned}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t &= (\boldsymbol{LF}+\boldsymbol{\epsilon})(\boldsymbol{LF}+\boldsymbol{\epsilon})^t \\ &= (\boldsymbol{LF}+\boldsymbol{\epsilon})((\boldsymbol{LF})^t+\boldsymbol{\epsilon}^t) \\ &= \boldsymbol{LF}(\boldsymbol{LF})^t+\boldsymbol{\epsilon}(\boldsymbol{LF})^t+(\boldsymbol{LF})\boldsymbol{\epsilon}^t+\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\end{aligned}$$

we have

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{cov}(\boldsymbol{X}) = \mathsf{E}\left((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t\right) \\ &= \boldsymbol{L}\,\mathsf{E}\left(\boldsymbol{F}\boldsymbol{F}^t\right)\boldsymbol{L}^t+\mathsf{E}\left(\boldsymbol{\epsilon}\boldsymbol{F}^t\right)\boldsymbol{L}^t+\boldsymbol{L}\,\mathsf{E}\left(\boldsymbol{F}\boldsymbol{\epsilon}^t\right)+\mathsf{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\right) \\ &= \boldsymbol{L}\boldsymbol{L}^t+\boldsymbol{\psi}\,.\end{aligned}$$

Since $(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{F}^t=(\boldsymbol{LF}+\boldsymbol{\epsilon})\boldsymbol{F}^t=\boldsymbol{LF}\boldsymbol{F}^t+\boldsymbol{\epsilon}\boldsymbol{F}^t$ we further get

$$\mathrm{cov}(\boldsymbol{X},\boldsymbol{F})=\mathsf{E}\left((\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{F}^t\right)=\mathsf{E}\left(\boldsymbol{LF}\boldsymbol{F}^t+\boldsymbol{\epsilon}\boldsymbol{F}^t\right)=\boldsymbol{L}\,\mathsf{E}(\boldsymbol{F}\boldsymbol{F}^t)+\mathsf{E}(\boldsymbol{\epsilon}\boldsymbol{F}^t)=\boldsymbol{L}\,.$$

That proportion of var$(X_i) = \sigma_{ii}$ contributed by the m common factors is called the ith **communality** h_i^2. The proportion of var(X_i) due to the specific factor is called the **uniqueness**, or **specific variance**. I.e.,

$$\begin{aligned} \text{var}(X_i) &= \text{communality} + \text{specific variance} \\ \sigma_{ii} &= \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 + \psi_i \,. \end{aligned}$$

With $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2$ we get

$$\sigma_{ii}^2 = h_i^2 + \psi_i \,.$$

The factor model assumes that the $p(p+1)/2$ variances and covariances of $\boldsymbol{X}$ can be reproduced by the pm factor loadings ℓij and the p specific variances ψ_i. For $p = m$, the matrix $\boldsymbol{\Sigma}$ can be reproduced exactly as $\boldsymbol{L}\boldsymbol{L}^t$, so $\boldsymbol{\psi}$ is the zero matrix. If m is small relative to p, then the factor model provides a simple explanation of $\boldsymbol{\Sigma}$ with fewer parameters.

Drawbacks:

• Most covariance matrices can not be factored as $\boldsymbol{L}\boldsymbol{L}^t + \boldsymbol{\psi}$, where $m << p$.

• There is some inherent ambiguity associated with the factor model: let $\boldsymbol{T}$ be any $m \times m$ orthogonal matrix so that $\boldsymbol{T}\boldsymbol{T}^t = \boldsymbol{T}\boldsymbol{T} = \boldsymbol{I}$. then we can rewrite the factor model as

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}^t\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\epsilon}\,.$$

Since with $\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{T}$ and $\boldsymbol{F}^* = \boldsymbol{T}\boldsymbol{F}$ we also have

$$\mathsf{E}(\boldsymbol{F}^*) = \boldsymbol{T}\,\mathsf{E}(\boldsymbol{F}) = \boldsymbol{0}\,, \quad \text{and} \quad \mathsf{cov}(\boldsymbol{F}^*) = \boldsymbol{T}^t\,\mathsf{cov}(\boldsymbol{F})\boldsymbol{T} = \boldsymbol{T}^t\boldsymbol{T} = \boldsymbol{I}\,,$$

it is impossible to distinguish the loadings in $\boldsymbol{L}$ from those in $\boldsymbol{L}^*$. The factors $\boldsymbol{F}$ and $\boldsymbol{F}^*$ have the same statistical properties.

Methods of Estimation

With observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ on $\boldsymbol{X}$, factor analysis seeks to answer the question: Does the factor model with a smaller number of factors adequately represent the data?

The sample covariance matrix $\boldsymbol{S}$ is an estimator of the unknown population covariance matrix $\boldsymbol{\Sigma}$. If the off-diagonal elements of $\boldsymbol{S}$ are small, the variables are not related and a factor analysis model will not prove useful. In these cases, the *specific variances* play the dominant role, whereas the major aim of factor analysis is to determine a few important *common factors*.

If $\boldsymbol{S}$ deviate from a diagonal matrix then the initial problem is to estimate the factor loadings $\boldsymbol{L}$ and specific variances $\boldsymbol{\psi}$. Two methods are very popular: the *principal component method* and the *maximum likelihood method*. Both of these solutions can be *rotated* afterwards in order to simplify the interpretation of the factors.

The Principal Component Approach:

Let $\boldsymbol{\Sigma}$ have eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Then

$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^t + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^t + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^t \,.$$

Thus we define

$$\boldsymbol{L}^t = (\sqrt{\lambda_1}\boldsymbol{e}_1, \sqrt{\lambda_2}\boldsymbol{e}_2, \ldots, \sqrt{\lambda_p}\boldsymbol{e}_p)$$

to get a factor analysis model with as many factors as variables ($m = p$) and specific variances $\psi_i = 0$ for all i i.e.

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^t + \boldsymbol{0} = \boldsymbol{L}\boldsymbol{L}^t \,.$$

This is not very useful, however, if the last eigenvalues are relatively small we neglect the contributions of $\lambda_{m+1}\boldsymbol{e}_{m+1}\boldsymbol{e}_{m+1}^t + \lambda_{m+2}\boldsymbol{e}_{m+2}\boldsymbol{e}_{m+2}^t + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^t$ to $\boldsymbol{\Sigma}$ above.

This gives us the approximation

$$\boldsymbol{\Sigma} \approx \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^t + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^t + \cdots + \lambda_m \boldsymbol{e}_m \boldsymbol{e}_m^t = \boldsymbol{L}\boldsymbol{L}^t \,,$$

where $\boldsymbol{L}$ is now a $(m \times p)$ matrix of coefficients as required. This representation assumes that the specific factors $\boldsymbol{\epsilon}$ are of minor importance. If specific factors are included in the model, their variances may be taken to be the diagonal elements of $\boldsymbol{\Sigma} - \boldsymbol{L}\boldsymbol{L}^t$ and the approximation becomes

$$\boldsymbol{\Sigma} \approx \boldsymbol{L}\boldsymbol{L}^t + \boldsymbol{\psi} \,,$$

where $\psi_i = \sigma_{ii}^2 - \sum_j \ell_{ij}^2$.

To apply this approach to data, it is customary first to center the observations (this does not change the sample covariance structure) and to consider $\boldsymbol{x}_j - \overline{\boldsymbol{x}}$.

If the units of the variables are not of the same size then it is desirable to work with the standardizes variables $z_{ij} = (x_{ij} - \overline{x}_i)/\sqrt{s_{ii}}$ having sample variance $\boldsymbol{R}$.

Applying the above technique onto $\boldsymbol{S}$ or $\boldsymbol{R}$ is known as the *principal component solution*.

By the definition of $\hat{\psi}_i = s_{ii} - \sum_j \hat{\ell}_{ij}^2$, where $\hat{\boldsymbol{\ell}}_i$ are the eigenvectors of $\boldsymbol{S}$ (or $\boldsymbol{R}$), the diagonal elements of $\boldsymbol{S}$ are equal to the diagonal elements of $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^t + \hat{\boldsymbol{\psi}}$. However, the off-diagonal elements of $\boldsymbol{S}$ are not usually reproduced by $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^t + \hat{\boldsymbol{\psi}}$.

• How to determine the number of factors, m?

Consider the residual matrix of a m factor model

$$\boldsymbol{S} - (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^t + \hat{\boldsymbol{\psi}})$$

with zero diagonal elements. If the other elements are also small we will take the m factor model to be appropriate.

Ideally, the contributions of the first few factors to the sample variance should be large. The contribution to the sample variance s_{ii} from the first common factor is $\hat{\ell}_{i1}^2$. The contribution to the *total sample variance*, $s_{11} + s_{22} + \cdots + s_{pp}$, from the first common factor is

$$\sum_{i=1}^{p} \hat{\ell}_{i1}^2 = \left(\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1\right)^t \left(\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1\right) = \hat{\lambda}_1 \,.$$

In general, the proportion of total sample variance due to the jth factor is

$$\frac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}}$$

for a factor analysis of $\boldsymbol{S}$, or

$$\frac{\hat{\lambda}_j}{p}$$

for a factor analysis of $\boldsymbol{R}$.

Software packages sometimes set m equal to the number of eigenvalues of $\boldsymbol{R}$ largen than 1 (if the correlation matrix is factored), or equal m to the number of positive eigenvalues of $\boldsymbol{S}$. (Be careful when using these rules blindly!)

Example: In a consumer-preference study, a number of customers were asked to rate several attributes of a new product. The correlation matrix of the responses was calculated.

Taste	1.00	0.02	0.96	0.42	0.01
Good buy for money	0.02	1.00	0.13	0.71	0.85
Flavor	0.96	0.13	1.00	0.50	0.11
Suitable for snack	0.42	0.71	0.50	1.00	0.79
Provides lots of energy	0.01	0.85	0.11	0.79	1.00

```
> library(mva)
> R <- matrix(c(1.00,0.02,0.96,0.42,0.01,
                0.02,1.00,0.13,0.71,0.85,
                0.96,0.13,1.00,0.50,0.11,
                0.42,0.71,0.50,1.00,0.79,
                0.01,0.85,0.11,0.79,1.00), 5, 5)
> eigen(R)
 $values
 [1] 2.85309042 1.80633245 0.20449022 0.10240947 0.03367744
```

```
$vectors
           [,1]        [,2]        [,3]       [,4]         [,5]
[1,] 0.3314539  0.60721643  0.09848524  0.1386643  0.701783012
[2,] 0.4601593 -0.39003172  0.74256408 -0.2821170  0.071674637
[3,] 0.3820572  0.55650828  0.16840896  0.1170037 -0.708716714
[4,] 0.5559769 -0.07806457 -0.60158211 -0.5682357  0.001656352
[5,] 0.4725608 -0.40418799 -0.22053713  0.7513990  0.009012569
```

The first 2 eigenvalues of $\boldsymbol{R}$ are the only ones being larger than 1. These two will account for

$$\frac{2.853 + 1.806}{5} = 0.93$$

of the total (standardized) sample variance. Thus we decide to set $m = 2$.

There is no special function available in R allowing to get the estimated factor loadings, communalities, and specific variances (uniquenesses). Hence we directly calculate those quantities.

```
> L <- matrix(rep(0, 10), 5, 2)  # factor loadings
> for (j in 1:2) L[ ,j] <- sqrt(eigen(R)$values[j]) * eigen(R)$vectors[ ,j]
       [,1]   [,2]
 [1,] 0.560  0.816
 [2,] 0.777 -0.524
 [3,] 0.645  0.748
 [4,] 0.939 -0.105
 [5,] 0.798 -0.543
> h2 <- diag(L %*% t(L));  h2    # communalities
 [1] 0.979 0.879 0.976 0.893 0.932
> psi <- diag(R) - h2;  psi      # specific variances
 [1] 0.0205 0.1211 0.0241 0.1071 0.0678
> R - (L %*% t(L) + diag(psi))   # residuals
         [,1]   [,2]    [,3]   [,4]    [,5]
 [1,]  0.0000  0.013 -0.0117 -0.020  0.0064
 [2,]  0.0126  0.000  0.0205 -0.075 -0.0552
 [3,] -0.0117  0.020  0.0000 -0.028  0.0012
 [4,] -0.0201 -0.075 -0.0276  0.000 -0.0166
 [5,]  0.0064 -0.055  0.0012 -0.017  0.0000
```

Thus we would judge a 2-factor model providing a good fit to the data. The large communalities indicate that this model accounts for a large percentage of the sample variance of each variable.

A Modified Approach – The Principle Factor Analysis

We describe this procedure in terms of a factor analysis of $\boldsymbol{R}$. If

$$\boldsymbol{\rho} = \boldsymbol{L}\boldsymbol{L}^t + \boldsymbol{\psi}$$

is correctly specified, then the m common factors should account for the off-diagonal elements of $\boldsymbol{\rho}$, as well as the communality portions of the diagonal elements

$$\rho_{ii} = 1 = h_i^2 + \psi_i \, .$$

If the specific factor contribution ψ_i is removed from the diagonal or, equivalently, the 1 replaced by h_i^2 the resulting matrix is $\boldsymbol{\rho} - \boldsymbol{\psi} = \boldsymbol{L}\boldsymbol{L}^t$.

Suppose initial estimates ψ_i^* are available. Then we replace the ith diagonal element of $\boldsymbol{R}$ by $h_i^{*2} = 1 - \psi_i^*$, and obtain the reduced correlation matrix $\boldsymbol{R}_r$, which is now factored as

$$\boldsymbol{R}_r \approx \boldsymbol{L}_r^* \boldsymbol{L}_r^{*t} .$$

The *principle factor method* of factor analysis employs the estimates

$$\boldsymbol{L}_r^* = \left[\sqrt{\hat{\lambda}_1^*}\hat{\boldsymbol{e}}_1^*, \sqrt{\hat{\lambda}_2^*}\hat{\boldsymbol{e}}_2^*, \ldots, \sqrt{\hat{\lambda}_m^*}\hat{\boldsymbol{e}}_m^* \right]$$

and

$$\hat{\psi}_i^* = 1 - \sum_{j=1}^{m} \ell_{ij}^{*2} ,$$

where $(\hat{\lambda}_i^*, \hat{\boldsymbol{e}}_i^*)$ are the (largest) eigenvalue-eigenvector pairs from $\boldsymbol{R}_r$. Re-estimate the communalities again and continue till convergence. As initial choice of h_i^{*2} you can use $1 - \psi_i^* = 1 - 1/r^{ii}$, where r^{ii} is the ith diagonal element of $\boldsymbol{R}^{-1}$.

Example cont'ed:

```
> h2 <- 1 - 1/diag(solve(R));  h2  # initial guess
 [1] 0.93 0.74 0.94 0.80 0.83

> R.r <- R;  diag(R.r) <- h2
> L.star <- matrix(rep(0, 10), 5, 2)  # factor loadings
> for (j in 1:2) L.star[ ,j] <- sqrt(eigen(R.r)$values[j]) * eigen(R.r)$vectors[ ,j]
> h2.star <- diag(L.star %*% t(L.star));  h2.star  # communalities
 [1] 0.95 0.76 0.95 0.83 0.88

> # apply 3 times to get convergence

> R.r <- R;  diag(R.r) <- h2.star
> L.star <- matrix(rep(0, 10), 5, 2)  # factor loadings
> for (j in 1:2) L.star[ ,j] <- sqrt(eigen(R.r)$values[j]) * eigen(R.r)$vectors[ ,j]
> h2.star <- diag(L.star %*% t(L.star));  h2.star  # communalities
 [1] 0.97 0.77 0.96 0.83 0.93
```

```
> L.star  # loadings
       [,1]  [,2]
 [1,] -0.60 -0.78
 [2,] -0.71  0.51
 [3,] -0.68 -0.71
 [4,] -0.90  0.15
 [5,] -0.77  0.58

>  1 - h2.star  # specific variances
 [1] 0.032 0.231 0.039 0.167 0.069
```

The principle components method for $\boldsymbol{R}$ can be regarded as a principal factor method with initial communality estimates of unity (or specific variance estimates equal to zero) and without iterating.

The only estimating procedure available in R is the maximum likelihood method. Beside the PCA method this is the only one, which is strongly recommended and shortly discussed now.

Maximum Likelihood Method

We now assume that the common factors $\boldsymbol{F}$ and the specific factors $\boldsymbol{\epsilon}$ are from a normal distribution. Then maximum likelihood estimates of the unknown factor loadings $\boldsymbol{L}$ and the specific variances $\boldsymbol{\psi}$ may be obtained.

This strategy is the only one which is implemented in R and is now applied onto our example.

Example cont'ed:

```
> factanal(covmat = R, factors=2)

 Call:
 factanal(factors = 2, covmat = R, rotation = "none")

 Uniquenesses: [1] 0.028 0.237 0.040 0.168 0.052

 Loadings:
```

```
     Factor1 Factor2
[1,]   0.976  -0.139
[2,]   0.150   0.860
[3,]   0.979
[4,]   0.535   0.738
[5,]   0.146   0.963

                Factor1 Factor2
SS loadings        2.24    2.23
Proportion Var     0.45    0.45
Cumulative Var     0.45    0.90
```

Factor Rotation

Since the original factor loadings are (a) not unique, and (b) usually not interpretable, we rotate them until a *simple structure* is achieved.

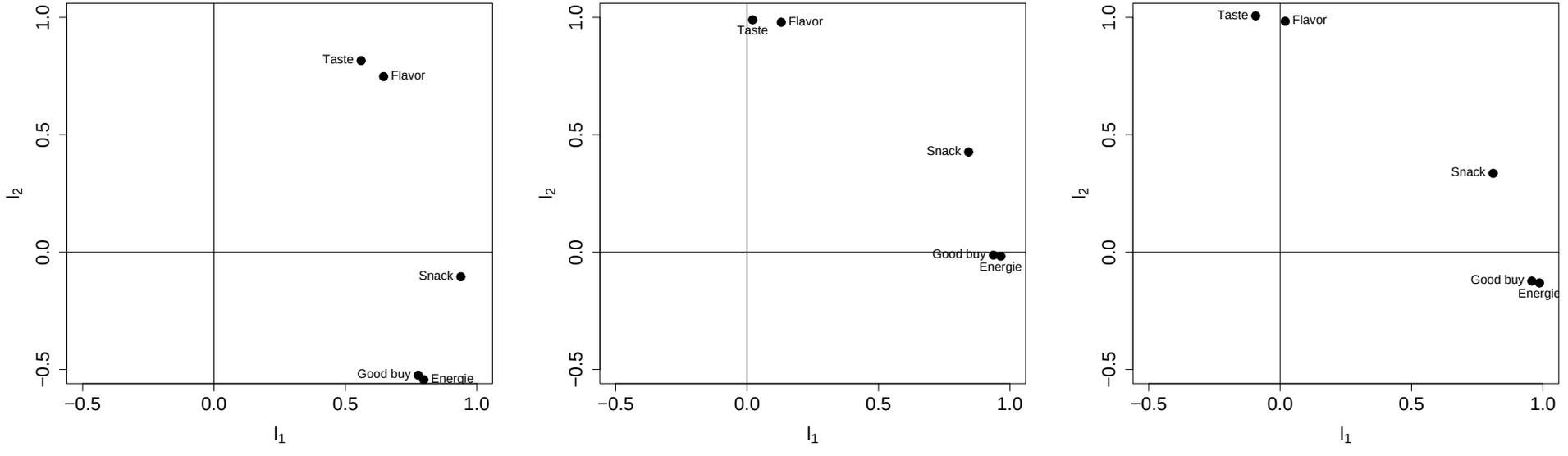
We concentrate on graphical methods for $m = 2$. A plot of the pairs of factor loadings $(\hat{\ell}_{i1}, \hat{\ell}_{i2})$, yields p points, each point corresponding to a variable. These points can be rotated by using either the `varimax` or the `promax` criterion.

Example cont'ed: Estimates of the factor loadings from the principal component approach were:

```
> L
      [,1]   [,2]
[1,] 0.560  0.816
[2,] 0.777 -0.524
[3,] 0.645  0.748
[4,] 0.939 -0.105
[5,] 0.798 -0.543
```

```
> varimax(L)
      [,1]   [,2]
[1,] 0.021  0.989
[2,] 0.937 -0.013
[3,] 0.130  0.979
[4,] 0.843  0.427
[5,] 0.965 -0.017
```

```
> promax(L)
      [,1]   [,2]
[1,] -0.093  1.007
[2,]  0.958 -0.124
[3,]  0.019  0.983
[4,]  0.811  0.336
[5,]  0.987 -0.131
```



After rotation it's much clearer to see that variables 2 (Good buy), 4 (Snack), and 5 (Energy) define factor 1 (high loadings on factor 1, small loadings on factor 2), while variables 1 (Taste) and 3 (Flavor) define factor 2 (high loadings on factor 2, small loadings on factor 1).

Johnson & Wichern call factor 1 a *nutrition factor* and factor 2 a *taste factor*.

Factor Scores

In factor analysis, interest is usually centered on the parameters in the factor model. However, the estimated values of the common factors, called $factor\ scores$, may also be required (e.g., for diagnostic purposes).

These scores are not estimates of unknown parameters in the usual sense. They are rather estimates of values for the unobserved random factor vectors. Two methods are provided in `factanal( ..., scores = )`: the `regression` method of Thomson, and the weighted least squares method of `Bartlett`.

Both these methods allows us to plot n such p-dimensional observations as n m-dimensional scores.

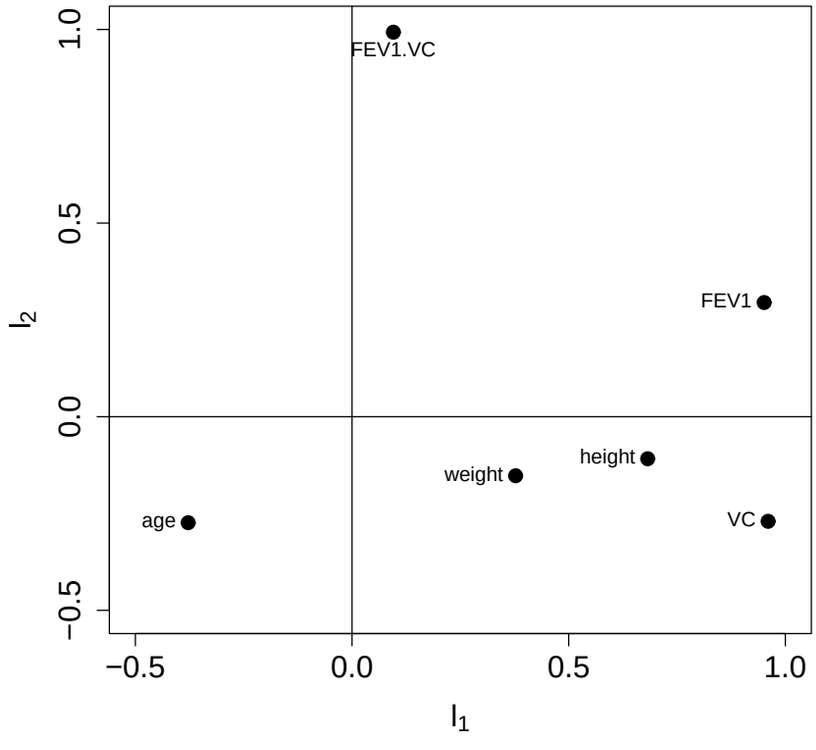
Example: A factor analytic analysis of the `fvc` data might be as follows:

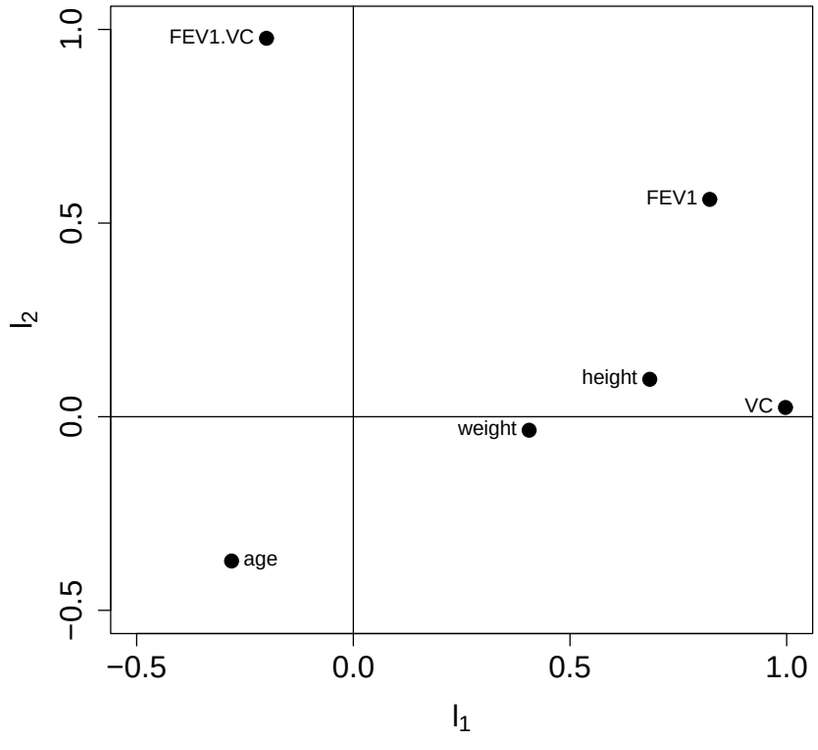
- calculate the maximum likelihood estimates of the loadings w/o rotation,
- apply a varimax rotation on these estimates and check plot of the loadings,
- estimate factor scores and plot them for the n observations.

```
> fa <- factanal(aimu[, 3:8], factors=2, scores="none", rotation="none"); fa
 Uniquenesses:
     age  height  weight      VC    FEV1 FEV1.VC
   0.782   0.523   0.834   0.005   0.008   0.005
 Loadings:
         Factor1 Factor2
 age     -0.378  -0.274
 height   0.682  -0.109
 weight   0.378  -0.153
 VC       0.960  -0.270
 FEV1     0.951   0.295
 FEV1.VC          0.993
```

```
                Factor1 Factor2
 SS loadings      2.587   1.256
 Proportion Var   0.431   0.209
 Cumulative Var   0.431   0.640
> L <- fa$loadings
> Lv <- varimax(fa$loadings);  Lv
 $loadings

         Factor1  Factor2
 age     -0.2810 -0.37262
 height   0.6841  0.09667
 weight   0.4057 -0.03488
 VC       0.9972  0.02385
 FEV1     0.8225  0.56122
 FEV1.VC -0.2004  0.97716

 $rotmat
         [,1]   [,2]
 [1,]  0.9559 0.2937
 [2,] -0.2937 0.9559
```

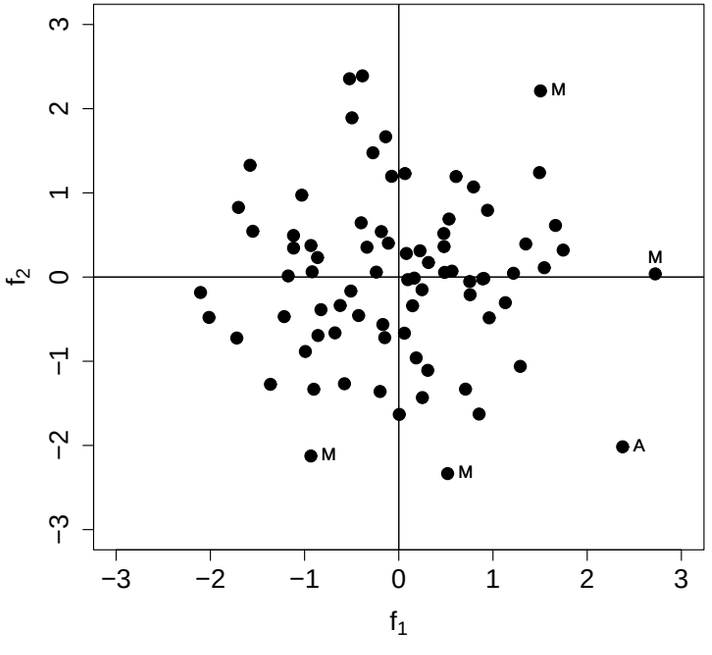
```
> plot(L); plot(Lv)
```






```
> s <- factanal(aimu[, 3:8], factors=2, scores="reg", rot="varimax")$scores
> plot(s);  i <- identify(s, region);  aimu[i, ]
```

	nr	year	age	height	weight	VC	FEV1	FEV1.VC	region
25	25	85	28	189	85	740	500	68	A
38	38	83	44	174	78	475	335	71	M
46	46	83	23	190	75	665	635	95	M
57	57	83	25	187	102	780	580	81	M
71	71	83	37	173	78	590	400	68	M



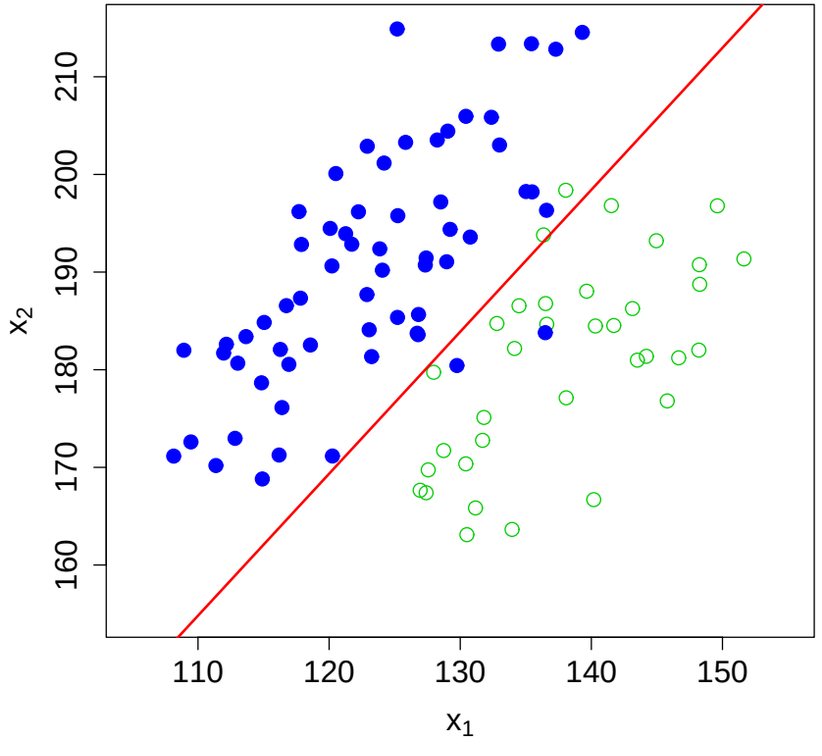
Discrimination and Classification

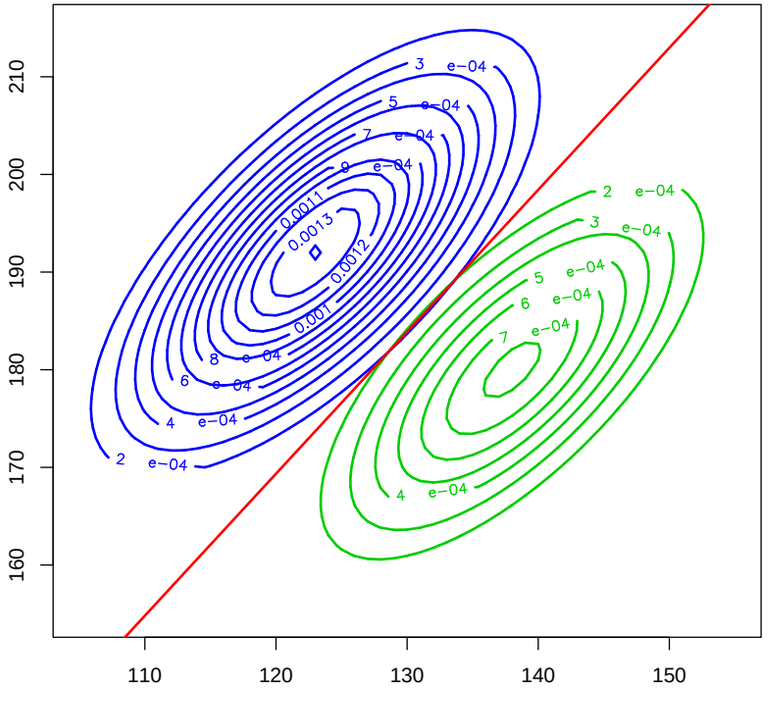
Discriminant analysis (DA) and classification are multivariate techniques concerned with $separating$ distinct sets of objects (observations) and with $allocating$ new objects to previously defined groups (defined by a categorial variable). There are several purposes for DA:

- (**Discrimination, separation**) To describe either graphically (low dimension) or algebraically, the differential features of objects from several known collections (populations, or groups).

- (**Classification, allocation**) To sort objects into 2 or more labelled classes. Thus, we derive a rule, that is used to optimally assign a new object to the labelled classes.

Consider 2 classes. Label these groups g_1, g_2. The objects are to be classified on the basis of measurements on a p variate random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. The observed values differ to some extend from one class to the other. Thus we assume that all objects $\boldsymbol{x}$ in class i have density $f_i(\boldsymbol{x})$, $i = 1, 2$.





Technique was introduced by R.A. Fisher. His idea was to transform the multivariate $\boldsymbol{x}$ to univariate y such that the y's derived from population g_1 and g_2 were separated as much as possible. He considered linear combinations of $\boldsymbol{x}$.

If we let μ_{1Y} be the mean of Y obtained from $\boldsymbol{X}$ belonging to g_1, and μ_{2Y} be the mean of Y obtained from $\boldsymbol{X}$ belonging to g_2, then he selected the linear combination that maximized the squared distance between μ_{1Y} and μ_{2Y} relative to the variability of the Y's.

We define

$$\boldsymbol{\mu}_1 = \mathsf{E}(\boldsymbol{X}|g_1)\,, \quad \text{and} \quad \boldsymbol{\mu}_2 = \mathsf{E}(\boldsymbol{X}|g_2)$$

and suppose the covariance matrix

$$\boldsymbol{\Sigma} = \mathsf{E}\left((\boldsymbol{X} - \boldsymbol{\mu}_i)(\boldsymbol{X} - \boldsymbol{\mu}_i)^t\right), \quad i = 1, 2$$

is the same for both populations (somewhat critical in practice).

We consider the linear combination

$$Y = \boldsymbol{\ell}^t \boldsymbol{X}$$

and get population-specific means

$$\begin{aligned} \mu_{1Y} &= \mathsf{E}(Y|g_1) = \mathsf{E}(\boldsymbol{\ell}^t \boldsymbol{X}|g_1) = \boldsymbol{\ell}^t \boldsymbol{\mu}_1 \\ \mu_{2Y} &= \mathsf{E}(Y|g_2) = \mathsf{E}(\boldsymbol{\ell}^t \boldsymbol{X}|g_2) = \boldsymbol{\ell}^t \boldsymbol{\mu}_2 \end{aligned}$$

but equal variance

$$\sigma_Y^2 = \mathsf{var}(Y) = \mathsf{var}(\boldsymbol{\ell}^t \boldsymbol{X}) = \boldsymbol{\ell}^t \,\mathsf{cov}(\boldsymbol{X})\boldsymbol{\ell} = \boldsymbol{\ell}^t \boldsymbol{\Sigma} \boldsymbol{\ell}\,.$$

The best linear combination is derived from the ratio ($\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$)

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{(\boldsymbol{\ell}^t \boldsymbol{\mu}_1 - \boldsymbol{\ell}^t \boldsymbol{\mu}_2)^2}{\boldsymbol{\ell}^t \boldsymbol{\Sigma} \boldsymbol{\ell}} = \frac{\boldsymbol{\ell}^t (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\ell}}{\boldsymbol{\ell}^t \boldsymbol{\Sigma} \boldsymbol{\ell}} = \frac{(\boldsymbol{\ell}^t \boldsymbol{\delta})^2}{\boldsymbol{\ell}^t \boldsymbol{\Sigma} \boldsymbol{\ell}}$$

Result: Let $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $Y = \boldsymbol{\ell}^t \boldsymbol{X}$, then

$$\frac{(\boldsymbol{\ell}^t \boldsymbol{\delta})^2}{\boldsymbol{\ell}^t \boldsymbol{\Sigma} \boldsymbol{\ell}}$$

is maximized by the choice

$$\boldsymbol{\ell} = c\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\big)$$

for any $c \neq 0$. Choosing $c = 1$ produces the linear combination

$$Y = \boldsymbol{\ell}^t \boldsymbol{X} = \big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\big)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$$

which is known as *Fisher's linear discriminant function*.

We can also employ this result as *classification device*. Let $y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0$ be the value of the discriminant function for a new observation $\boldsymbol{x}_0$ and let

$$m = \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

be the midpoint of the 2 univariate population means. It can be shown that

$$\mathsf{E}(Y_0|g_1) - m \geq 0 \qquad \text{and} \qquad \mathsf{E}(Y_0|g_2) - m < 0$$

That is, if $\boldsymbol{X}_0$ is from g_1, Y_0 is expected to be larger than the midpoint. If $\boldsymbol{X}_0$ is from g_2, Y_0 is expected to be smaller. Thus the classification rule is:

Allocate $\boldsymbol{x}_0$ to g_1 if: $\quad y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0 \geq m$

Allocate $\boldsymbol{x}_0$ to g_2 if: $\quad y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0 < m\,.$

Because the population moments are not known, we replace $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ by their empirical versions.

Suppose we have 2 data matrices $\boldsymbol{X}_1$ from g_1 and $\boldsymbol{X}_2$ from g_2 with n_1 and n_2 observations, from which we calculate both sample means $\overline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_2$, and sample covariance matrices $\boldsymbol{S}_1$, $\boldsymbol{S}_2$. Since it is assumed that the covariance matrices in the groups are the same, we combine (pool) $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ to derive a single estimate of $\boldsymbol{\Sigma}$. Hence we use the *pooled sample covariance* matrix

$$\boldsymbol{S}_p = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}$$

an unbiased estimate of $\boldsymbol{\Sigma}$. Now, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are replaced by $\overline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_2$, and $\boldsymbol{S}_p$ in the previous formulas to give *Fisher's sample linear discriminant function*

$$y = \hat{\boldsymbol{\ell}}^t \boldsymbol{x} = \left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\right)^t \boldsymbol{S}_p^{-1} \boldsymbol{x}\,.$$

The midpoint between both sample means is

$$\hat{m} = \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^t \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)$$

and the classification rule becomes

$$\begin{aligned} &\text{Allocate } \boldsymbol{x}_0 \text{ to } g_1 \text{ if:} && (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^t \boldsymbol{S}_p^{-1} \boldsymbol{x}_0 \geq \hat{m} \\ &\text{Allocate } \boldsymbol{x}_0 \text{ to } g_2 \text{ if:} && (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^t \boldsymbol{S}_p^{-1} \boldsymbol{x}_0 < \hat{m}\,. \end{aligned}$$

This idea can be easily generalized onto more than 2 classes. Moreover, instead of using a linear discriminant function we can also use a quadratic one.

Example: Fisher's Iris data

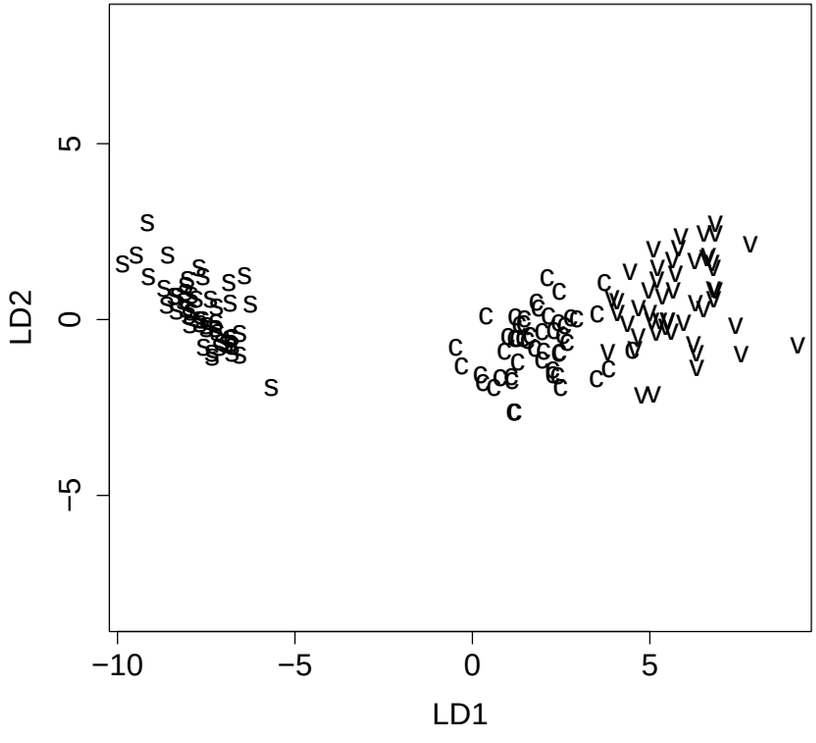
Data describing the sepal (Kelchblatt) width and length, and the petal (Blten-blatt) width and length of 3 different Iris species (Setosa, Versicolor, Virginica) were observed. There are 50 observation for each species.

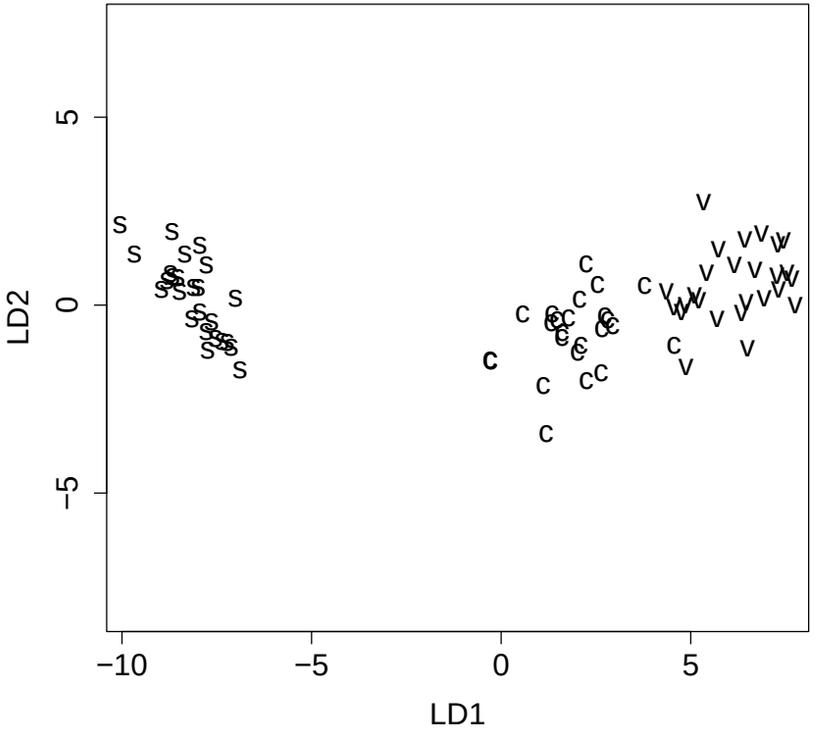
```
> library(MASS)
> data(iris3)
> Iris <- data.frame(rbind(iris3[,,1], iris3[,,2], iris3[,,3]),
                     Sp = rep(c("s","c","v"), rep(50,3)))
> z <- lda(Sp ~ Sepal.L.+Sepal.W.+Petal.L.+Petal.W., Iris, prior = c(1,1,1)/3)
 Prior probabilities of groups:
         c         s         v
 0.3333333 0.3333333 0.3333333

 Group means:
   Sepal.L. Sepal.W. Petal.L. Petal.W.
 c    5.936    2.770    4.260    1.326
 s    5.006    3.428    1.462    0.246
 v    6.588    2.974    5.552    2.026
```

```
Coefficients of linear discriminants:
                LD1         LD2
Sepal.L. -0.8293776  0.02410215
Sepal.W. -1.5344731  2.16452123
Petal.L.  2.2012117 -0.93192121
Petal.W.  2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088

> predict(z, Iris)$class
  [1] s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s
 [32] s s s s s s s s s s s s s s s s s s c c c c c c c c c c c c c
 [63] c c c c c c c c v c c c c c c c c c c c c v c c c c c c c c c
 [94] c c c c c c c v v v v v v v v v v v v v v v v v v v v v v v v
[125] v v v v v v v v v c v v v v v v v v v v v v v v v v v v v
```

```
> table(predict(z, Iris)$class, Iris$Sp)
     c  s  v
  c 48  0  1
  s  0 50  0
  v  2  0 49
> train <- sample(1:150, 75);  table(Iris$Sp[train])
  c  s  v
 24 25 26
> z1 <- lda(Sp ~ Sepal.L.+Sepal.W.+Petal.L.+Petal.W., Iris,
           prior = c(1,1,1)/3, subset = train)
> predict(z1, Iris[-train, ])$class
  [1] s s s s s s s s s s s s s s s s s s s s s s s c c c c c c
 [32] c c c c c c v c c c c c c c c c c c c v v v v v v v v v v
 [63] v v v c v v v v v v v v v
> table(predict(z1, Iris[-train, ])$class, Iris[-train, ]$Sp)
     c  s  v
  c 25  0  1
  s  0 25  0
  v  1  0 23
```

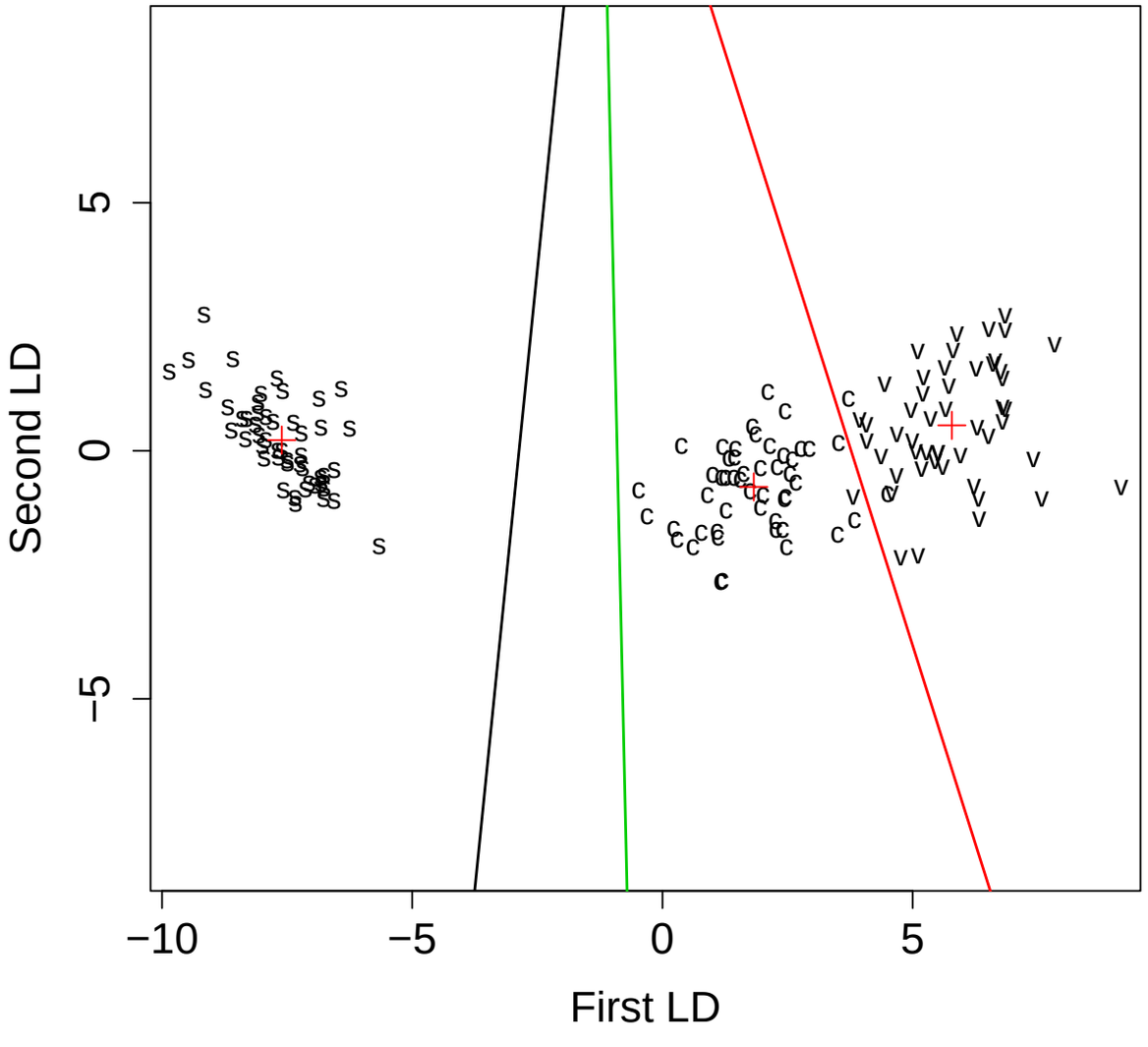
```
> plot(z);   plot(z1)
```





```
> ir.ld <- predict(z, Iris)$x  # => LD1 and LD2 coordinates
> eqscplot(ir.ld, type="n", xlab="First LD", ylab="Second LD") # eq. scaled axes
> text(ir.ld, as.character(Iris$Sp))  # plot LD1 vs. LD2

> # calc group-spec. means of LD1 & LD2
> tapply(ir.ld[ , 1], Iris$Sp, mean)
        c         s         v
 1.825049 -7.607600  5.782550
> tapply(ir.ld[ , 2], Iris$Sp, mean)
          c          s          v
 -0.7278996  0.2151330  0.5127666

> # faster alternative:
> ir.m <- lda(ir.ld, Iris$Sp)$means;  ir.m
         LD1        LD2
 c  1.825049 -0.7278996
 s -7.607600  0.2151330
 v  5.782550  0.5127666
> points(ir.m, pch=3, mkh=0.3, col=2)  # plot group means as "+"
```

```
> perp <- function(x, y, ...) {
+  m <- (x+y)/2                        # midpoint of the 2 group means
+  s <- -(x[1]-y[1])/(x[2]-y[2])       # perpendicular line through midpoint
+  abline(c(m[2]-s*m[1], s), ...)      # draw classification regions
+  invisible()
> }
> perp(ir.m[1,], ir.m[2,], col=1)  # classification decision b/w groups 1&2
> perp(ir.m[1,], ir.m[3,], col=2)  # classification decision b/w groups 1&3
> perp(ir.m[2,], ir.m[3,], col=3)  # classification decision b/w groups 2&3
```


Second LD
First LD
5
0
−5
−10
−5
0
5