Mnohorozměrná analýza (NMST539)

Z. Hlávka

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravděpodobnosti a matematické statistiky
www.karlin.mff.cuni.cz/~hlavka

Mnohorozměrná analýza (NMST539)

- Mnohorozměrná data.
- Opakování: lineární algebra (matice).
- Mnohorozměrné normální rozdělení, Wishartovo a Hotellingovo rozdělení.
- Metoda hlavních komponent, faktorová analýza.
- Mnohorozměrné škálování, shluková a diskriminační analýza.
- Kanonické korelace, korespondenční analýza.
- Další metody (hloubka dat, SIR, projection pursuit).

Použitá literatura:

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

Härdle, W. K., & Simar, L. (2014). *Applied multivariate statistical analysis, 4th edition*. Springer Science & Business Media.

Anděl, J. (1985). *Matematická statistika*. SNTL.

Týden 1

Předpokládané znalosti: základní maticové operace (sčítání, násobení apod.)

Mnohorozměrná data:

- grafické znázornění,
- matice dat a popisné statistiky.

Comparison of Batches

Figure: An old Swiss 1000-franc bill.

Example

Example: The authorities have measured

X_1 = length of the bill
X_2 = height of the bill (left)
X_3 = height of the bill (right)
X_4 = distance of the inner frame to the lower border
X_5 = distance of the inner frame to the upper border
X_6 = length of the diagonal of the central picture.

Example: Swiss bank data

The dataset consists of 200 measurements on Swiss bank notes. One half of these bank notes are genuine, the other half are forged bank notes.

It is important to be able to decide whether a given banknote is genuine.

We want to derive a good rule that separates the genuine and forged banknotes.

Which measurement is most informative? We have to visualize the difference.

Graphics

Computers allow easy construction of informative plots:

- 1D Boxplot, histogram, kernel density estimator (KDE), dotplot, ...
- 2D Histogram, KDE, scatterplot.
- 3D 3D scatterplot.
- 4+ Scatterplot matrix, parallel coordinates, Chernoff-Flury faces, Andrew's curves.

One typically needs static graphics (PDF) for reports and interactive graphics for data exploration.

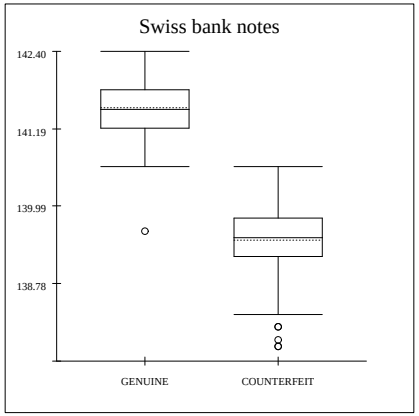


Figure: Variables X_6 (diagonal) of bank notes, the genuine at the left. → `MVAboxbank6`

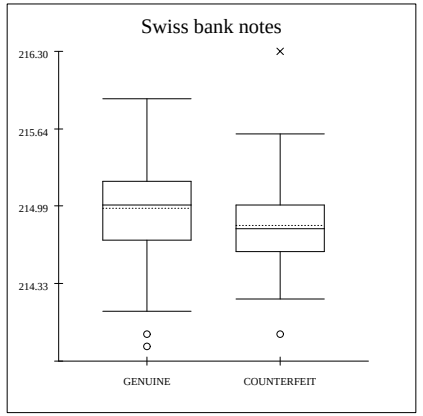


Figure: Variables X_1 (length) of bank notes, the genuine at the left. → `MVAboxbank1`

Boxplots

- Median and mean bar indicate the central locations.
- The relative location of median (and mean) in the box is a measure of skewness.
- The length of box and whiskers is a measure of spread.
- The length of whiskers indicate the tail length of the distribution.
- Outlying points are marked as "$\star$" or"$\bullet$" outside the outside bars.
- Boxplots do not indicate multi modality or clusters.
- If we compare the relative size and location of the boxes we are comparing distributions.

Histograms

The histogram counts relative frequencies of observations x_i falling into predefined bins:

$$\widehat{f}_h(x) = n^{-1}h^{-1}\sum_{j\in\mathbb{Z}}\sum_{i=1}^{n} \mathrm{I}\{x_i \in B_j(x_0,h)\}\,\mathrm{I}\{x \in B_j(x_0,h)\}$$

- the histogram is a simple estimator of a probability density,
- h is a smoothing parameter and controls the width of the histogram bins.

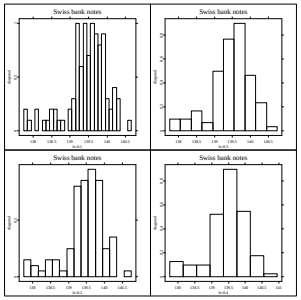


Figure: Diagonal of forged bank notes. Histograms with $x_0 = 137.8$ and $h = 0.1$ (upper left), $h = 0.2$ (lower left), $h = 0.3$ (upper right), $h = 0.4$ (lower right).

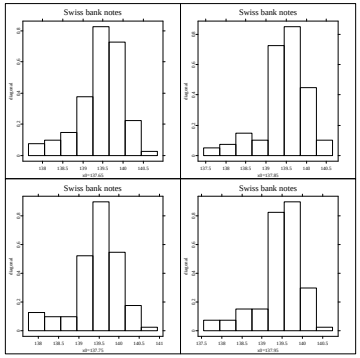


Figure: Diagonal of forged bank notes. Histogram with $h = 0.4$ and origins $x_0 = 137.65$ (upper left), $x_0 = 137.75$ (lower left),$x_0 = 137.85$ (upper right), $x_0 = 137.95$ (lower right).

For $x \in B_j$ (assuming that the density $f(x)$ is 'reasonable'), it is easy to calculcate the bias $E\widehat{f}_h(x) - f(x) \doteq f'(m_j)(m_j - x)$ and variance $\text{var}\,\widehat{f}_h(x) \doteq \frac{1}{nh} f(x)$.

It follows that the Mean Squared Error is

$$\text{MSE}\{\widehat{f}_h(x)\} = \frac{1}{mn} f(x) + \{f'(m_j)\}^2 (m_j - x)^2 + o(h) + o(1/nh).$$

By integrating MSE and taking limits, we easily obtain

$$\text{AMISE}\{\widehat{f}_h(x)\} = \frac{1}{nh} + \frac{h^2}{12} \|f'\|_2^2$$

leading the asymptotically optimal bandwidth $h_0 = \{6/(n\|f'\|_2^2)\}^{1/3}$.

Histograms

- Modes of the density correspond to strong peaks in the histogram.
- Histograms with the same h need not be identical because they also depend on the origin x_0 of the grid.
- The consequence of a too large h is a too flat, unstructured histogram (large bias).
- A too small binwidth h results in a wiggly histogram (large variance).
- It is recommended to use averaged histograms (so-called kernel density estimators).

Kernel density estimators (KDEs)

Kernel density estimator is a natural generalization of a histogram (by shifting the "bin", we obtain smooth estimator of the underlying probability density).

Histogram (at the center of a bin) can be written as

$$\widehat{f}_h(x) = n^{-1}(2h)^{-1}\sum_{i=1}^{n} I(|x - x_i| \leq h)$$

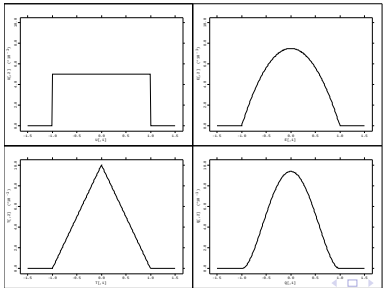
$K(u) = I(|u| \leq 1)/2$

$$\widehat{f}_h(x) = n^{-1}h^{-1}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

K is the so-called kernel.

Common kernel functions

$K(u) = \frac{1}{2}I(\lvert u\rvert \leq 1)$	Uniform
$K(u) = (1 - \lvert u\rvert)I(\lvert u\rvert \leq 1)$	Triangle
$K(u) = \frac{3}{4}(1 - u^2)I(\lvert u\rvert \leq 1)$	Epanechnikov
$K(u) = \frac{15}{16}(1 - u^2)^2 I(\lvert u\rvert \leq 1)$	Quartic (Biweight)
$K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2}) = \varphi(u)$	Gaussian



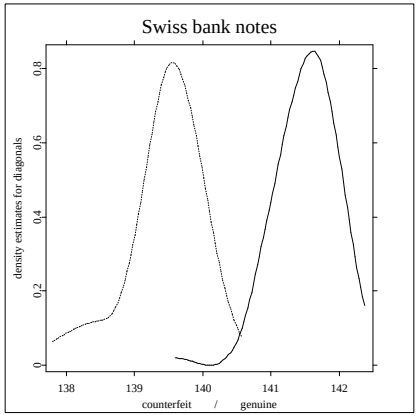


Figure: Densities of diagonals of genuine and forged bank notes. Automatic density estimates. → `MVAdenbank`

The bias of KDE

$$\text{Bias}\widehat{f}_h(x) = \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2)$$

is of a smaller order than the bias of histogram.

Proceeding similarly, it is straghtforward that the Asymptotic Mean Integrated Squred Error is

$$\text{AMISE}(\widehat{f}_h) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2\|f''\|_2^2$$

leading the asymptotically optimal bandwidth

$$h_0 = \left(\frac{\|K\|_2^2}{n\|f''\|_2^2\mu_2(K)\}^2}\right)^{1/5}.$$

Choice of the bandwidth

Assuming normality and using Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2})$, the unknown constants can be calculated and we obtain the so-called **Silverman's rule of thumb:**

$$h_G = 1.06\hat{\sigma} n^{-\frac{1}{5}},$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{x=1}^{n}(x_i - \bar{x})^2$

Using Quartic kernel, the constants are somewhat different: $h_Q = 2.62 h_G$.

In practice, one must be very careful because statistical software may assume another standardization of the kernel function (i.e., the bandwidth parameter may be multiplied by some constant).

Cross-validation is a popular bandwidth selection method (producing somewhat unstable results).

KDEs in R

Libraries: KernSmooth, ks, sm.

See R task-views on CRAN:

- 1D density(), bkde(KernSmooth), locpoly(KernSmooth),
- 3D sm.density (sm)
- 6D kde(ks)

Unfortunately, multivariate KDEs have slow rates of convergence (so-called *curse of dimensionality*) — see Modern Statistical Methods (NMST434) for more details.

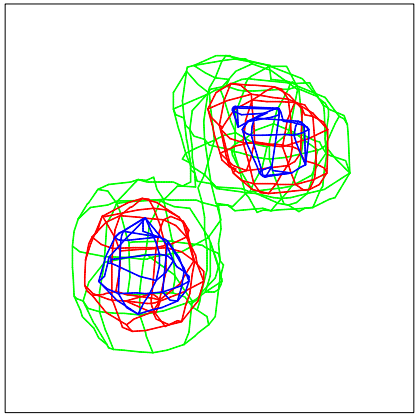


Figure: Contours of the density of X_4, X_5, X_6 of genuine and forged bank notes.
→ `MVAcontbank3`

Kernel densities

- Kernel densities estimate distribution densities by the kernel method.
- The bandwidth h determines the degree of smoothness of the estimate $\widehat{f}$.
- A simple (but not necessarily correct) way to find a good bandwidth is to compute the rule of thumb bandwidth $h_G = 1.06\widehat{\sigma} n^{-1/5}$. This bandwidth is to be used only in combination with a Gaussian kernel φ.
- Kernel density estimates are a good descriptive tool for seeing modes, location, skewness, tails, asymmetry etc.

Scatterplots

- Rotation of data
- Separation lines
- Draftman plot
- Brushing
- Parallel coordinate plots

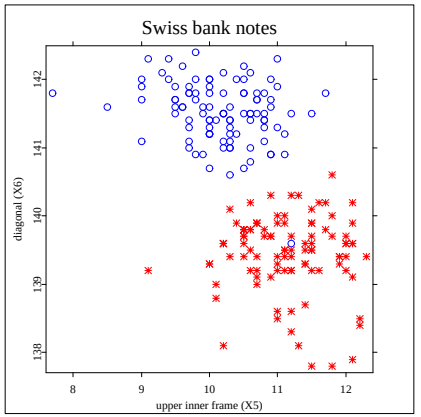


Figure: 2D scatterplot for X_5 vs. X_6 of the bank notes. Genuine notes are circles, forged are stars. → `MVAscabank56`

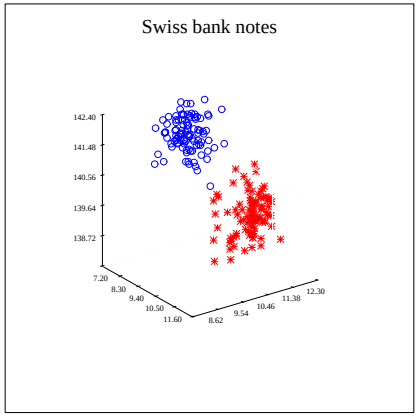


Figure: 3D Scatterplot for (X_4, X_5, X_6) of the bank notes. Genuine notes are circles, forged are stars. → `MVAscabank456`

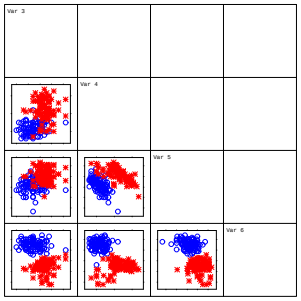


Figure: Draftman plot of the bank notes. The pictures in the left column show (X_3, X_4), (X_3, X_5) and (X_3, X_6), in the middle we have (X_4, X_5) and (X_4, X_6), and in the lower right is (X_5, X_6). → `MVAdrafbank1`

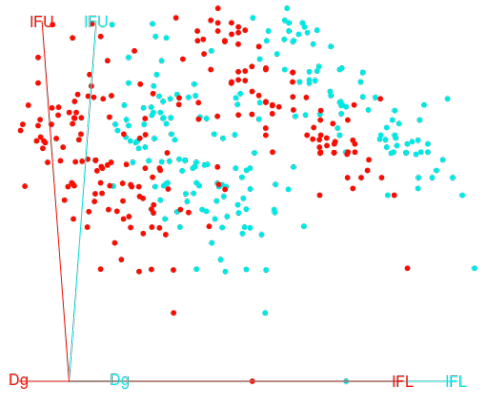


Figure: Stereo plot of the bank notes — (X_4, X_5, X_6).

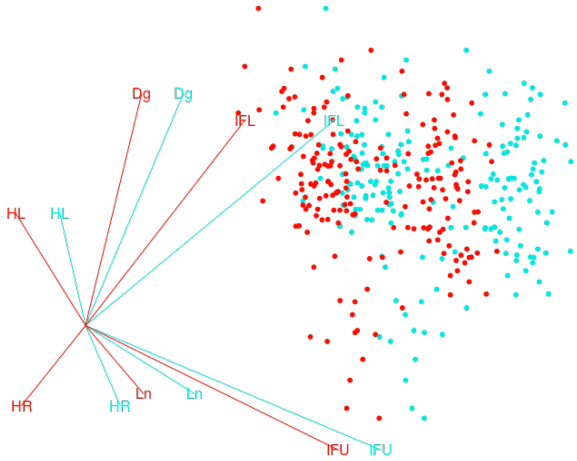


Figure: Stereo plot of the bank notes — all variables.

Scatterplots

- Scatterplots in two and three dimensions help us in seeing separated points or clouds.
- They help us in judging positive or negative dependence.
- Draftman scatterplot matrices are useful for detecting structures conditioned on values of certain variables.
- As the brush of a scatterplot matrix is moving in the point cloud we can study conditional dependence (e.g., in Ggobi).

Parallel coordinate plots

- based on a orthogonal coordinate system
- allows to see more than four dimensions

Idea:
Instead of plotting observations in an orthogonal coordinate system one draws their coordinates in a system of parallel axes. This way of representation is however sensitive to the order of the variables.

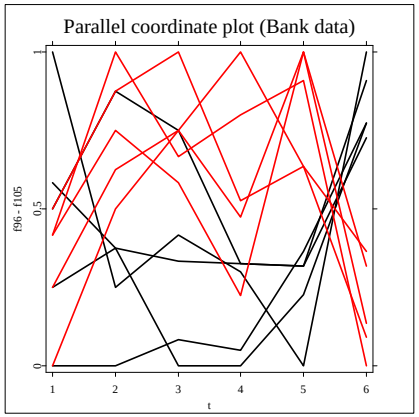


Figure: Parallel coordinate plot of observations 96–105 → `MVAparcoo1`

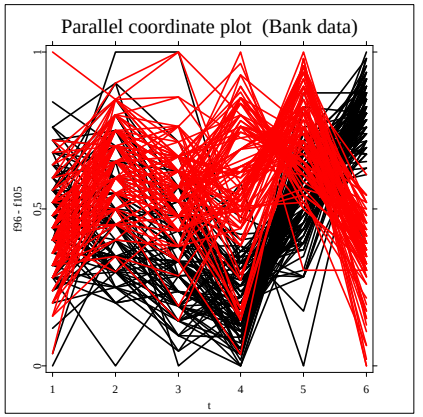


Figure: The full bank Data set. Genuine banknotes displayed as solid lines. The forged bank notes are shown as dashed lines. → `MVAparcoo2`

Parallel coordinate plots

- Parallel coordinate plots overcome the visulisation problem of the Cartesian coordinate system for dimensions greater than 4.
- Outliers are seen as outlying polygon curves.
- The order of variables is still important for detection of subgroups for example.
- Subgroups may be screened by selective coloring in an interactive manner.

Faces

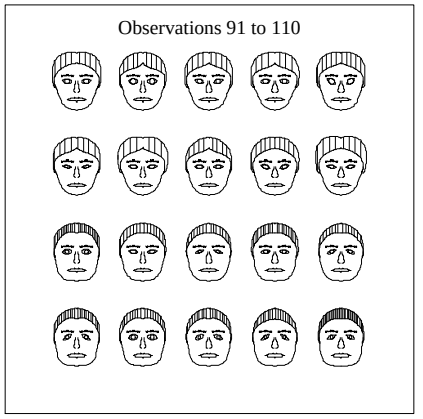


Figure: Flury faces for observations 91 to 110 of the bank notes. → `MVAfacebank`

Six variables to the following face elements

X_1 = 1, 19 (eye sizes)
X_2 = 2, 20 (pupil sizes)
X_3 = 4, 22 (eye slants)
X_4 = 11, 29 (upper hair lines)
X_5 = 12, 29 (lower hair lines)
X_6 = 13, 14, 31, 32 (face lines and darkness of hair)

```
library(aplpack)
faces(bank2)
faces(bank2[91:110])
```


Faces

- faces can be used to detect subgroups in multivariate data
- subgroups are characterized by similar looking faces
- outliers are identified by extreme faces (e.g. dark hair)
- if one element of X is unusual the corresponding face element changes a lot in shape

Summary Statistics

$\mathcal{X}(n \times p)$ data matrix

$$\mathcal{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

mean

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix} = n^{-1}\mathcal{X}^\top 1_n$$

Covariance matrix

$$\mathcal{S} = n^{-1}\mathcal{X}^\top\mathcal{X} - \overline{x}\,\overline{x}^\top$$
$$= n^{-1}(\mathcal{X}^\top\mathcal{X} - n^{-1}\mathcal{X}^\top 1_n 1_n^\top \mathcal{X}) = n^{-1}\mathcal{X}^\top\mathcal{H}\mathcal{X}$$

Centering matrix

$$\mathcal{H} = \mathcal{I}_n - n^{-1}1_n 1_n^\top$$

centered data: $S = n^{-1}\mathcal{X}^\top\mathcal{X}$

$\mathcal{D} = \mathrm{diag}(s_{X_jX_j})$, where $X_j, j = 1, \ldots, p$ are the columns of $\mathcal{X}$

Correlation matrix $\mathcal{R} = \mathcal{D}^{-1/2}\mathcal{S}\mathcal{D}^{-1/2}$

Linear transformations

$\mathcal{A}$ $(q \times p)$ matrix

$$\mathcal{Y} = \mathcal{X}\mathcal{A}^\top = (y_1, \ldots, y_n)^\top$$
$$\overline{y} = \mathcal{A}\overline{x}$$
$$\mathcal{S}_\mathcal{Y} = \mathcal{A}\mathcal{S}_\mathcal{X}\mathcal{A}^\top$$

Example: $\overline{x} = (1, 2)^\top$

$y = 4x, x \in I\!R^2$

$\overline{y} = 4\overline{x} = (4, 8)^\top$

Mahalanobis Transformation

$$z_i = \mathcal{S}^{-1/2}(x_i - \overline{x}), \quad i = 1, \ldots, n,$$
$$\mathcal{S}_\mathcal{Z} = n^{-1}\mathcal{Z}^\top\mathcal{H}\mathcal{Z} = \mathcal{I}_p,$$
$$\overline{\mathcal{Z}} = 0.$$

where $\mathcal{H}$ is the centering matrix.

Mahalanobis transformation leads to standardized uncorrelated zero mean data matrix $\mathcal{Z}$.

Summary Statistics

- The center of gravity of a data matrix is given by its mean vector $\overline{x} = n^{-1}\mathcal{X}^\top 1_n$.
- The dispersion of the observations in a data matrix is given by the empirical covariance matrix $\mathcal{S} = n^{-1}\mathcal{X}^\top\mathcal{H}\mathcal{X}$.
- The empirical correlation matrix is given by $\mathcal{R} = \mathcal{D}^{-1/2}\mathcal{S}\mathcal{D}^{-1/2}$.
- A linear transformation $\mathcal{Y} = \mathcal{X}\mathcal{A}^\top$ of a data matrix $\mathcal{X}$ has mean $\mathcal{A}\overline{x}$ and empirical covariance $\mathcal{A}\mathcal{S}_\mathcal{X}\mathcal{A}^\top$.
- The Mahalanobis transformation is a linear transformation $z_i = \mathcal{S}^{-1/2}(x_i - \overline{x})$ which gives a standardized, uncorrelated data matrix $\mathcal{Z}$.

Týden 2

Opakování základní maticové algebry:

- spektrální rozklad matice a kvadratické formy.

Náhodné vektory:

- mnohorozměrná distribuční funkce a hustota,
- podmíněná a marginální rozdělení,
- momenty,
- mnohorozměrné normální rozdělení.

A short Excursion into Matrix Algebra

$$\mathcal{A}_{(n\times p)} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}$$

Definition	Notation
Transpose	$\mathcal{A}^\top$
Sum	$\mathcal{A}+\mathcal{B}$
Difference	$\mathcal{A}-\mathcal{B}$
Scalar product	$c\cdot\mathcal{A}$
Product	$\mathcal{A}\cdot\mathcal{B}$
Rank	$rank(\mathcal{A})$
Trace	$tr(\mathcal{A})$
Determinant	$det(\mathcal{A})=\lvert\mathcal{A}\rvert$
Inverse	$\mathcal{A}^{-1}$
Generalised Inverse	$\mathcal{A}^- : \mathcal{A}\mathcal{A}^-\mathcal{A}=\mathcal{A}$

Name	Definition	Notation	Example
scalar	$p=n=1$	a	3
column vector	$p=1$	a	$\begin{pmatrix}1\\3\end{pmatrix}$
row vector	$n=1$	$a^\top$	$\begin{pmatrix}1 & 3\end{pmatrix}$
vector of ones	$\underbrace{(1,\ldots,1)}_{n}^\top$	1_n	$\begin{pmatrix}1\\1\end{pmatrix}$
vector of zeros	$\underbrace{(0,\ldots,0)}_{n}^\top$	0_n	$\begin{pmatrix}0\\0\end{pmatrix}$
square matrix	$n=p$	$\mathcal{A}(p\times p)$	$\begin{pmatrix}2 & 0\\0 & 2\end{pmatrix}$

Name	Definition	Notation	Example
diagonal matrix	$a_{ij}=0,\ i\neq j,\ n=p$	$\mathrm{diag}(a_{ii})$	$\begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}$
identity matrix	$\mathrm{diag}\underbrace{(1,\ldots,1)}_{p}$	$\mathcal{I}_p$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$
unit matrix	$a_{ij}\equiv 1,\ n=p$	$1_n 1_n^\top$	$\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}$
symmetric matrix	$a_{ij}=a_{ji}$		$\begin{pmatrix}1 & 2\\2 & 3\end{pmatrix}$

Name	Definition	Example
null matrix	$a_{ij}=0$	$\begin{pmatrix}0 & 0\\0 & 0\end{pmatrix}$
upper triangular matrix	$a_{ij}=0, i<j$	$\begin{pmatrix}1 & 2 & 4\\0 & 1 & 3\\0 & 0 & 1\end{pmatrix}$
idempotent matrix	$\mathcal{A}^2=\mathcal{A}$	$\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\\\frac{1}{2} & \frac{1}{2}\end{pmatrix}$
orthogonal matrix	$\mathcal{A}^T\mathcal{A}=I=\mathcal{A}\mathcal{A}^\top$	$\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix}$

Properties of a Square Matrix

For any $\mathcal{A}(n \times n)$ and $\mathcal{B}(n \times n)$ and any scalar c

$$\begin{aligned} tr(\mathcal{A}+\mathcal{B}) &= tr(\mathcal{A}) + tr(\mathcal{B}) \\ tr(c\mathcal{A}) &= c\ tr(\mathcal{A}) \\ |c\mathcal{A}| &= c^n|\mathcal{A}| \\ tr(\mathcal{A}\mathcal{B}) &= tr(\mathcal{B}\mathcal{A}) \\ |\mathcal{A}\mathcal{B}| &= |\mathcal{B}\mathcal{A}| \\ |\mathcal{A}\mathcal{B}| &= |\mathcal{A}||\mathcal{B}| \\ |\mathcal{A}^{-1}| &= |\mathcal{A}|^{-1} \end{aligned}$$

Eigenvalues and Eigenvectors

Square matrix $\mathcal{A}(n \times n)$

eigenvalue $\lambda = Eval(\mathcal{A})$

eigenvector $\gamma = Evec(\mathcal{A})$

$$\mathcal{A}\gamma = \lambda\gamma$$

Eigenvalues describe the 'size' of the matrix $\mathcal{A}$:

$$\begin{aligned} |\mathcal{A}| &= \prod_{j=1}^{n} \lambda_j \\ tr(\mathcal{A}) &= \sum_{j=1}^{n} \lambda_j \end{aligned}$$

Matrix Algebra

- The determinant $|\mathcal{A}|$ is a product of the eigenvalues of $\mathcal{A}$.
- The inverse of a matrix $\mathcal{A}$ exists if $|\mathcal{A}| \neq 0$.
- The trace $\text{tr}(\mathcal{A})$ is the sum of the eigenvalues of $\mathcal{A}$.
- The sum of the traces of two matrices equals the trace of the sum of the two matrices.
- The trace $\text{tr}(\mathcal{A}\mathcal{B})$ equals $\text{tr}(\mathcal{B}\mathcal{A})$.

Spectral Decomposition

Every real symmetric matrix $\mathcal{A}(p \times p)$ can be written as:

$$\begin{aligned} \mathcal{A} &= \Gamma\Lambda\Gamma^\top \\ &= \sum_{j=1}^{p} \lambda_j \gamma_j \gamma_j^\top \\ \Lambda &= diag(\lambda_1, \cdots, \lambda_p) \\ \Gamma &= (\gamma_1, \cdots, \gamma_p), \end{aligned}$$

where the matrix Γ is orthogonal (i.e. $\Gamma^\top\Gamma = \mathcal{I}_p$).

Spectral decomposition allows easier calculation of powers of the matrix $\mathcal{A}$ (very useful is the inverse $\mathcal{A}^{-1}$ and 'inverse square root' $\mathcal{A}^{-1/2}$).

Quadratic forms

$\mathcal{A}(p \times p)$ symmetric matrix

$$Q(x) = x^T \mathcal{A} x = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} x_i x_j$$

Definiteness

$Q(x) > 0$ for all $x \neq 0$	positive definite (pd) ,
$Q(x) \geq 0$ for all $x \neq 0$	positive semidefinite (psd) .

$\mathcal{A}$ is pd (psd) iff $Q(x) = x^T \mathcal{A} x$ is pd (psd).

Example:

$Q(x) = x^T \mathcal{A} x = x_1^2 + x_2^2$, $\mathcal{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
eigenvalues: $\lambda_1 = \lambda_2 = 1$ positive definite

$Q(x) = (x_1 - x_2)^2$, $\mathcal{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$
eigenvalues $\lambda_1 = 2, \lambda_2 = 0$ positive semidefinite

$Q(x) = x_1^2 - x_2^2$
eigenvalues $\lambda_1 = 1, \lambda_2 = -1$ indefinite.

$\mathcal{A} > 0$ if and only if all $\lambda_i > 0$, $i = 1, \ldots, p$

Maximization of quadratic forms

Theorem: $\mathcal{A}$, $\mathcal{B}$ symmetric, $\mathcal{B} > 0$. The maximum of $x^T \mathcal{A} x$ under the constraint $x^T \mathcal{B} x = 1$ is given by:

$\max_{\{x: x^\top \mathcal{B} x = 1\}} \quad x^T \mathcal{A} x = \lambda_1$
$\lambda_1 = 1.Eval\{\mathcal{B}^{-1}\mathcal{A}\}$
$\operatorname{argmax}_{\{x: x^\top \mathcal{B} x = 1\}} \quad x^T \mathcal{A} x = 1.Evec\{\mathcal{B}^{-1}\mathcal{A}\}$

Proof: the proof will be given during derivation of principal components.

Quadratic forms

- A quadratic form can be described by a symmetric quadratic matrix $\mathcal{A}$.
- Quadratic forms can always be diagonalized.
- Positive definiteness of a quadratic form is equivalent to positiveness of the eigenvalues of the matrix $\mathcal{A}$.
- Maximum and minimum of a quadratic form under constraints can be expressed in terms of eigenvalues.

Geometrical aspects

Distance function $d : \mathbb{R}^{2p} \to \mathbb{R}_+$ $d^2(x, y) = (x - y)^T \mathcal{A}(x - y), \quad \mathcal{A} > 0$
$\mathcal{A} = \mathcal{I}_p$, Euclidean distance
$E_d = \{x \in \mathbb{R}^p \mid (x - x_0)^\top (x - x_0) = d^2\}$

Example: $x \in \mathbb{R}^2, x_0 = 0, x_1^2 + x_2^2 = 1$

Norm of a vector

$$\|x\| = d(0, x) = \sqrt{x^T x}$$

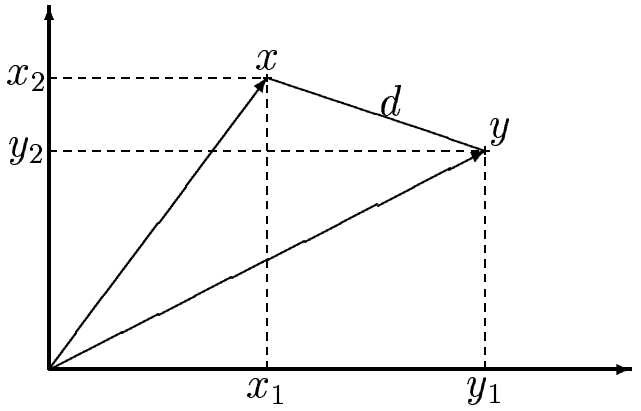


Figure: Distance d.

$d^2(x, y) = (x - y)^T (x - y)$

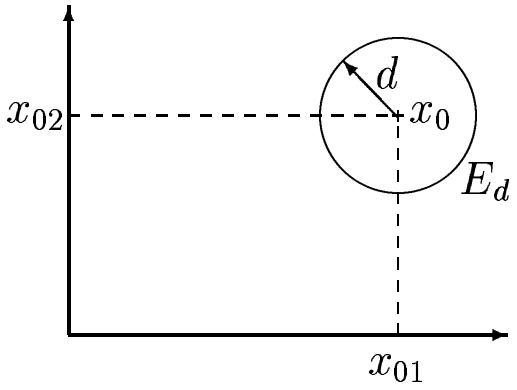


Figure: Iso–distance sphere.

$\mathcal{A} = \mathcal{I}_2$, $(x_1 - x_{01})^2 + (x_2 - x_{02})^2 = d^2$

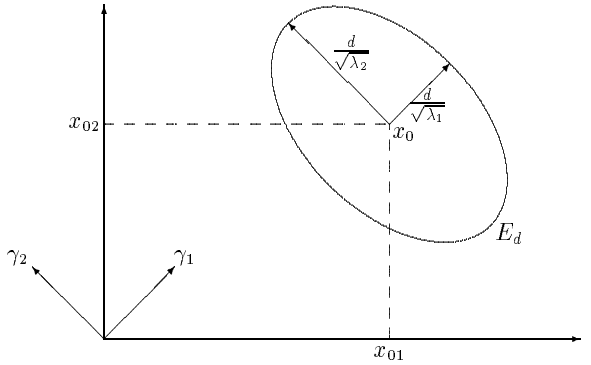


Figure: Iso–distance ellipsoid.

$E_d = \{x : (x - x_0)^\top \mathcal{A}(x - x_0) = d^2\}$, $\gamma_j = Evec(\mathcal{A})$, $\mathcal{A} > 0$

Angle between two Vectors

Angle of vectors x and y can be calculated as

$$\cos\theta = \frac{x^\top y}{\|x\|\,\|y\|}$$

Norm of a vector

$$\|x\| = d(0,x) = \sqrt{x^\top x}$$

Unit vectors

$$\{x : \|x\| = 1\}$$

Angle between two Vectors

Example: Angle = Correlation

Observations $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$
$\overline{x} = \overline{y} = 0$

$$\rho_{XY} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \cos\theta$$

Correlation corresponds to angle between $x, y \in I\!R^p$.

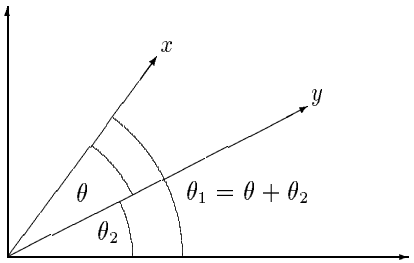


Figure: Angle between vectors.

$$\cos\theta = \frac{x^\top y}{\|x\|\|y\|} = \frac{x_1y_1+x_2y_2}{\|x\|\|y\|} = \cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2$$

Projection

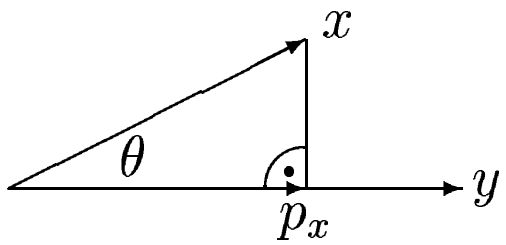


Figure: Projection.

$$p_x = y(y^T y)^{-1} y^T x = \frac{x^\top y}{\|y\|^2} y$$

Column space

$\mathcal{X}(n \times p)$ data matrix

$$C(\mathcal{X}) = \{x \in I\!R^n \mid \exists a \in I\!R^p \text{ so that } \mathcal{X}a = x\}$$

projection matrix
$\mathcal{P}(n \times n)$, $\mathcal{P} = \mathcal{P}^\top = \mathcal{P}^2$ ($\mathcal{P}$ is idempotent)
let $b \in I\!R^n$, $a = \mathcal{P}b$ is the projection of b on $C(\mathcal{P})$

Projection on $C(\mathcal{X})$

$$\mathcal{X}(n \times p), \quad \mathcal{P} = \mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top$$

$\mathcal{P}\mathcal{X} = \mathcal{X}$, $\mathcal{P}$ is a projector, $\mathcal{P}\mathcal{P} = \mathcal{P}$.

$$\mathcal{Q} = \mathcal{I}_n - \mathcal{P}, \mathcal{Q}^2 = \mathcal{Q}$$

$$p_x = \frac{y^\top x}{\|y\|^2} y$$

$$\mathcal{P}\mathcal{X} = \mathcal{X}$$

$$\mathcal{Q}\mathcal{X} = 0$$

Geometrical aspects

- A distance between two p-dimensional points x, y is a quadratic form $(x-y)^\top\mathcal{A}(x-y)$ in the vectors of differences $(x-y)$. A distance defines the norm of a vector.
- Iso-distance curves of a point x_0 are all those points which have the same distance from x_0. Iso-distance curves are ellipsoids whose principal axes are determined by the direction of the eigenvectors. The half-length of principal axes is proportional to the inverse of the roots of the eigenvalues of $\mathcal{A}$.
- The angle between to vectors x and y is given by $\cos\theta = \frac{x^\top\mathcal{A}y}{\|x\|_\mathcal{A}\ \|y\|_\mathcal{A}}$ w.r.t. the metric $\mathcal{A}$.

Geometrical aspects

- For the Euclidean distance with $\mathcal{A} = \mathcal{I}$ the correlation between two centered data vectors x and y is given by the cosine of the angle between them, i.e. $\cos\theta = \rho_{XY}$.
- The projection $\mathcal{P} = \mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top$ is the projection in the column space $C(\mathcal{X})$ of $\mathcal{X}$.
- The projection of $x \in I\!R^n$ on $y \in I\!R^n$ is given by $p_x = \frac{y^\top x}{\|y\|^2}y$.

Random vector

Let us assume that random variables $X_1, \ldots, X_p$ are defined on the probability space $(\Omega, \mathcal{A}, P)$. In this setup, the vector $(X_1, \ldots, X_p)^\top$ is called *random vector*.

Theorem: The p-dimensional random vector $X = (X_1, \ldots, X_p)^\top$ is a measurable function from $(\Omega, \mathcal{A}, P)$ to $(\mathbb{R}^p, \mathcal{B}_p)$

Proof: See Theorem II.1.1 in Anděl (1985).

The function

$$F(x_1, \ldots, x_p) = P(X_1 < x, \ldots X_p < x_p)$$

is the multivariate (joint) cummulative *distribution function* of the random vector $X = (X_1, \ldots, X_p)^\top$.

Multivariate density

A random vector X is absolutely continuous if there exists a probability *density* function (pdf), $f(.)$, such that

$$F(x) = \int_{-\infty}^{x} f(u)du.$$

For random vector $X = (X_1, \ldots, X_p)^\top$, we define (one-dimensional):

marginal distributions of X_i, $i = 1, \ldots, p$,
conditional distributions of $X_i | X_j = x_j$, $i, j \in \{1, \ldots, p\}$.

The expressions for marginal and conditional densities are easy to derive.

In the same way, marginal and conditional distributions are defined for all subvectors:

$X = (X_1, X_2)^\top, \quad X_1 \in \mathbb{R}^k \quad X_2 \in \mathbb{R}^{p-k}$

marginal density of X_1 is $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2$

conditional density of X_2 (conditioned on $X_1 = x_1$)
$f_{X_2|X_1=x_1}(x_2) = f(x_1, x_2)/f_{X_1}(x_1)$

Example:

$$f(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & 0 \le x_1, x_2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

$f(x_1, x_2)$ is a density since

$$\int f(x_1, x_2)dx_1x_2 = \frac{1}{2}\left[\frac{x_1^2}{2}\right]_0^1 + \frac{3}{2}\left[\frac{x_2^2}{2}\right]_0^1 = \frac{1}{4} + \frac{3}{4} = 1.$$

Example: The marginal densities

$$f_{X_1}(x_1) = \int f(x_1,x_2)dx_2 = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2\right) dx_2 = \frac{1}{2}x_1 + \frac{3}{4};$$
$$f_{X_2}(x_2) = \int f(x_1,x_2)dx_1 = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2\right) dx_1 = \frac{3}{2}x_2 + \frac{1}{4}.$$

The conditional densities

$$f(x_2 \mid x_1) = \frac{\frac{1}{2}x_1 + \frac{3}{2}x_2}{\frac{1}{2}x_1 + \frac{3}{4}} \quad \text{and} \quad f(x_1 \mid x_2) = \frac{\frac{1}{2}x_1 + \frac{3}{2}x_2}{\frac{3}{2}x_2 + \frac{1}{4}}.$$

Definition of (statistical) independence

Absolutely continuous random vectors X_1, X_2 are independent iff $f(x) = f(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$.

Two random variables may have identical marginals but different joint distribution.

Example:

$$f(x_1,x_2) = 1, \quad 0 < x_1, x_2 < 1,$$
$$f(x_1,x_2) = 1 + \alpha(2x_1 - 1)(2x_2 - 1), \quad 0 < x_1,\ x_2 < 1, \quad -1 \le \alpha \le 1.$$

$$f_{X_1}(x_1) = 1, \quad f_{X_2}(x_2) = 1.$$

$$\int_0^1 1 + \alpha(2x_1 - 1)(2x_2 - 1)dx_2 = 1 + \alpha(2x_1 - 1)[x_2^2 - x_2]_0^1 = 1.$$

Moments

$EX \in \mathbb{R}^p$ denotes the p-dimensional vector of expected values of the random vector X

$$EX = \begin{pmatrix} EX_1 \\ \vdots \\ EX_p \end{pmatrix} = \int x f(x)dx = \begin{pmatrix} \int x_1 f(x)dx \\ \vdots \\ \int x_p f(x)dx \end{pmatrix} = \mu.$$

The properties of expected value follow from the properties of the integral:

$$E(\alpha X + \beta Y) = \alpha EX + \beta EY$$

If X and Y are independent then

$$E(XY^\top) = \int xy^\top f(x,y)dxdy$$

$$= \int xf(x)dx \int y^\top f(y)dy = EXEY^\top$$

Definition of variance matrix (Σ)

$$\Sigma = Var(X) = E(X-\mu)(X-\mu)^\top$$

We say that random vector X has a distribution with the vector of expected values μ and the covariance matrix Σ,

$$X \sim (\mu, \Sigma)$$

Properties of the Covariance Matrix

Elements of Σ are variances and covariances of the components of the random vector X:

$\Sigma = (\sigma_{X_iX_j})$
$\sigma_{X_iX_j} = \text{cov}(X_i, X_j)$
$\sigma_{X_iX_i} = \text{var}(X_i)$

Computational formula:

$\Sigma = E(XX^\top) - \mu\mu^\top$

Variance matrix is positive semidefinite:

$\Sigma \geq 0$

(variance $a^\top \Sigma a$ of any linear combination $a^\top X$ cannot be negative).

Properties of Variances and Covariances

$$\text{var}(a^\top X) = a^\top \text{ Var}(X) \ a = \sum_{i,j} a_i a_j \sigma_{X_iX_j}$$

$$\text{Var}(\mathcal{A}X + b) = \mathcal{A} \text{ Var}(X) \ \mathcal{A}^\top$$

$$\text{Cov}(X+Y, Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$$

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Cov}(X,Y) + \text{Cov}(Y,X) + \text{Var}(Y)$$

$$\text{Cov}(\mathcal{A}X, \mathcal{B}Y) = \mathcal{A} \text{ Cov}(X,Y) \ \mathcal{B}^\top.$$

Conditional Expectations

(Absolutely continuous) random vector $X = (X_1, X_2)$

Conditional expectation of X_2, given $X_1 = x_1$:

$$E(X_2 \mid x_1) = \int x_2 f(x_2 \mid x_1) \ dx_2$$

and conditional expectation of X_1, given $X_2 = x_2$:

$$E(X_1 \mid x_2) = \int x_1 f(x_1 \mid x_2) \ dx_1$$

The conditional expectation $E(X_2 \mid x_1)$ is a function of x_1 (it is the expected value of X_2 if we know that corresponding $X_1 = x_1$—typical example of this setup is simple linear regression, where $E(Y \mid X = x) = x\beta$).

Moments

- The expectation of a random vector X is $\mu = \int x f(x)\, dx$, the covariance matrix $\Sigma = \mathrm{Var}(X) = E(X-\mu)(X-\mu)^\top$. We denote $X \sim (\mu, \Sigma)$.
- Expectations are linear, i.e., $E(\alpha X + \beta Y) = \alpha EX + \beta EY$. If X, Y are independent then $E(XY^\top) = EXEY^\top$.
- The covariance between two random vectors X, Y is $\Sigma_{XY} = \mathrm{Cov}(X,Y) = E(X-EX)(Y-EY)^\top = E(XY^\top) - EXEY^\top$. If X, Y are independent then $\mathrm{Cov}(X,Y) = 0$.
- The Conditional Expectation $E(X_2|X_1)$ is the MSE best approximation of X_2 by a function of X_1.

Multivariate Normal (Multinormal) Distribution

The pdf of a multinormal is (assuming that Σ has full rank):

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

$X \sim N_p(\mu, \Sigma)$

Expected value is $EX = \mu$,

Variance matrix of X is $\mathrm{Var}\{X\} = \Sigma > 0$.

(what is the meaning of the quadratic form $(x-\mu)^T\Sigma^{-1}(x-\mu)$ in the formula for density?)

Geometry of the $N_p(\mu, \Sigma)$ Distribution

Density of $N_p(\mu, \Sigma)$ is constant on ellipsoids of the form

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = d^2$$

If $X \sim N_p(\mu, \Sigma)$, then the variable $Y = (X-\mu)^\top \Sigma^{-1}(X-\mu)$ is χ^2_p distributed, since the Mahalonobis transformation $Z = \Sigma^{-1/2}(X-\mu) \sim N_p(0, \mathcal{I}_p)$ and $Y = Z^T Z = \sum_{j=1}^p Z_j^2$.

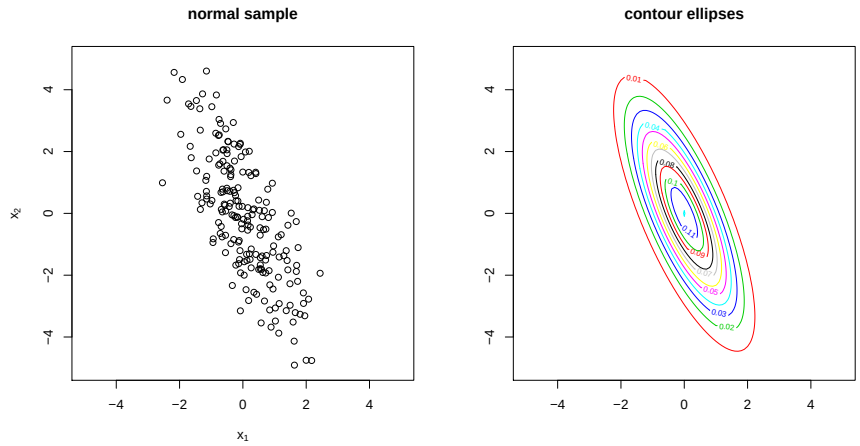


Figure: Scatterplot of normal sample and contour ellipses for $\mu = \binom{3}{2}$ and $\Sigma = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix}$ → `SMScontnorm`

Singular Normal Distribution

Definition of "Normal" distribution in case that the matrix Σ is singular—we use its eigenvalues λ_i and the generalized inverse Σ^-:

$\text{rank}(\Sigma) = k < p, \quad \lambda_1 \cdots \lambda_k$

$$\frac{(2\pi)^{-k/2}}{(\lambda_1 \cdots \lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^- (x-\mu)\right\}$$

Σ^- = G-inverse

Multinormal Distribution

- The pdf of a p-dimensional multinormal $X \sim N_p(\mu, \Sigma)$ is

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

  The contour curves of a multinormal are ellipsoids with half-lengths proportional to $\sqrt{\lambda_i}$, where λ_i denote the eigenvalues of Σ.
- The Mahalanobis transformation transforms $X \sim N_p(\mu, \Sigma)$ to $Y = \Sigma^{-1/2}(X-\mu) \sim N_p(0, \mathcal{I}_p)$. Vice versa, one can create a $X \sim N_p(\mu, \Sigma)$ from $Y \sim N_p(0, \mathcal{I}_p)$ via $X = \Sigma^{1/2} Y + \mu$.

Multinormal Distribution

- If the covariance matrix Σ is singular (i.e., $\text{rank}(\Sigma) < p$) then it defines a singular normal distribution.
- The density of a singular normal distribution is given by

$$\frac{(2\pi)^{-k/2}}{(\lambda_1 \cdots \lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^- (x-\mu)\right\},$$

  where Σ^- denotes the G-inverse of Σ.

Týden 3

Mnohorozměrné normální rozdělení:

- hustota transformovaného náhodného vektoru.
- centrální limitní věta a transformace,
- vlastnosti mnohorozměrného normálního rozdělení.

Transformations

Theorem: Assume that random vector $(X_1, \ldots, X_p)^\top$ has density $p(x)$ and that t is an injective and regular function on a set G such that $\int_G p(x)dx = 1$. Let τ denote the inverse function to $t : G \to t(G)$. Then the random vector $Y = t(X)$ has the density

$$q(y) = \begin{cases} p\{\tau(y)\}\,\mathrm{abs}(|\mathcal{J}(y)|) & \text{for } y \in t(G), \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{J}(y)$ denotes the Jacobian of the inverse function τ.

Proof: See Theorem III.2.5 in Anděl (1985).

Density of a linear transformation

$$Y = \mathcal{A}X + b, \quad \mathcal{A} \quad \text{nonsingular}$$

$$X = \mathcal{A}^{-1}(Y - b)$$

$$\mathcal{J} = \mathcal{A}^{-1}$$

$$f_Y(y) = \mathrm{abs}(|\mathcal{A}|^{-1}) f_X\{\mathcal{A}^{-1}(y - b)\}$$

Starting from $X \sim N_p(0_p, \mathcal{I}_p)$, it is now easy to calculate the p-dimensional density of $Y = \Sigma^{1/2}X + \mu \sim N_p(\mu, \Sigma)$ (assuming that Σ has full rank).

Note that the multivariate standard normal density (or characteristic function) can be defined as a product of univariate standard normal densities $\exp(-x^2/2)/\sqrt{2\pi}$ (or characteristic functions $\exp(-t^2/2)$).

Multivariate Normal distribution

Elementary properties

pdf:

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\tfrac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right\}$$

$$E(X) = \mu, \ \mathrm{Var}(X) = \Sigma > 0$$

Linear transformations

Linear transformations turn normal random variables into normal random variables.

$X \sim N_p(\mu, \Sigma), A(p \times p)$ full rank, $c \in \mathbb{R}^p$

$$Y = \mathcal{A}X + c \ \sim N_p(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top).$$

Alternative definition (Cramér-Wold characterization)

We have defined $N_p(\mu, \Sigma)$ by writing down its density. Unfortunately, this approach has some disadvantages because we assumed that Σ has full rank.

Definition: X has multivariate Normal distribution if and only if $a^\top X$ is univariate normal for all $a \in \mathbb{R}^p$.

It easily follows that $Y = Ax + c$ has multivariate normal distribution even when A is not square and it does not have full rank.

The random variable $Z = t^\top X \ \sim N(t^\top \mu, t^\top \Sigma t)$ has the characteristic function

$$\phi_Z(s) = E(\exp isZ) = \exp(ist^\top \mu - s^2 t^\top \Sigma t/2).$$

Therefore, the characteristic function of X is

$$\phi_X(t) = E(\exp it^\top X) = E(\exp iZ) = \phi_Z(1) = \exp(it^\top \mu - t^\top \Sigma t/2).$$

Partitioned Matrices

$\mathcal{A}(n \times p)$

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{pmatrix}$$

$\mathcal{A}_{ij}(n_i \times p_j)$

$$\begin{aligned} \mathcal{A}+\mathcal{B} &= \begin{pmatrix} \mathcal{A}_{11}+\mathcal{B}_{11} & \mathcal{A}_{12}+\mathcal{B}_{12} \\ \mathcal{A}_{21}+\mathcal{B}_{21} & \mathcal{A}_{22}+\mathcal{B}_{22} \end{pmatrix} \\ \mathcal{B}^\top &= \begin{pmatrix} \mathcal{B}_{11}^\top & \mathcal{B}_{21}^\top \\ \mathcal{B}_{12}^\top & \mathcal{B}_{22}^\top \end{pmatrix} \\ \mathcal{A}\mathcal{B}^\top &= \begin{pmatrix} \mathcal{A}_{11}\mathcal{B}_{11}^\top + \mathcal{A}_{12}\mathcal{B}_{12}^\top & \mathcal{A}_{11}\mathcal{B}_{21}^\top + \mathcal{A}_{12}\mathcal{B}_{22}^\top \\ \mathcal{A}_{21}\mathcal{B}_{11}^\top + \mathcal{A}_{22}\mathcal{B}_{12}^\top & \mathcal{A}_{21}\mathcal{B}_{21}^\top + \mathcal{A}_{22}\mathcal{B}_{22}^\top \end{pmatrix} \end{aligned}$$

Inverse of Partitioned Matrix

$\mathcal{A}$ nonsingular, $\mathcal{A}_{11}, \mathcal{A}_{22}$ square matrices

$$\mathcal{A}^{-1} = \begin{pmatrix} \mathcal{A}^{11} & \mathcal{A}^{12} \\ \mathcal{A}^{21} & \mathcal{A}^{22} \end{pmatrix}$$

where

$$\begin{cases} \mathcal{A}^{11} &= (\mathcal{A}_{11} - \mathcal{A}_{12}\mathcal{A}_{22}^{-1}\mathcal{A}_{21})^{-1} &= (\mathcal{A}_{11\cdot 2})^{-1} \\ \mathcal{A}^{12} &= -(\mathcal{A}_{11\cdot 2})^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-1} & \\ \mathcal{A}^{21} &= -\mathcal{A}_{22}^{-1}\mathcal{A}_{21}(\mathcal{A}_{11\cdot 2})^{-1} & \\ \mathcal{A}^{22} &= \mathcal{A}_{22}^{-1} + \mathcal{A}_{22}^{-1}\mathcal{A}_{21}(\mathcal{A}_{11\cdot 2})^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-1} & \end{cases}$$

Determinant:

$$|\mathcal{A}| = |\mathcal{A}_{11}||\mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}| = |\mathcal{A}_{22}||\mathcal{A}_{11} - \mathcal{A}_{12}\mathcal{A}_{22}^{-1}\mathcal{A}_{21}|$$

Partioned Matrices

* For $\mathcal{A}$ nonsingular, $\mathcal{A}_{11}$, $\mathcal{A}_{22}$ square matrices,

$$\mathcal{A}^{-1} = \begin{pmatrix} \mathcal{A}^{11} & \mathcal{A}^{12} \\ \mathcal{A}^{21} & \mathcal{A}^{22} \end{pmatrix}$$

$$\begin{cases} \mathcal{A}^{11} &= (\mathcal{A}_{11} - \mathcal{A}_{12}\mathcal{A}_{22}^{-1}\mathcal{A}_{21})^{-1} &= (\mathcal{A}_{11\cdot 2})^{-1} \\ \mathcal{A}^{12} &= -(\mathcal{A}_{11\cdot 2})^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-1} & \\ \mathcal{A}^{21} &= -\mathcal{A}_{22}^{-1}\mathcal{A}_{21}(\mathcal{A}_{11\cdot 2})^{-1} & \\ \mathcal{A}^{22} &= \mathcal{A}_{22}^{-1} + \mathcal{A}_{22}^{-1}\mathcal{A}_{21}(\mathcal{A}_{11\cdot 2})^{-1}\mathcal{A}_{12}\mathcal{A}_{22}^{-1} & \end{cases}$$

* For $\mathcal{B} = \begin{pmatrix} 1 & b^\top \\ a & \mathcal{A} \end{pmatrix}$ we have $|\mathcal{B}| = |\mathcal{A} - ab^\top| = |\mathcal{A}||1 - b^\top\mathcal{A}^{-1}a|$.
* $(\mathcal{A} - ab^\top)^{-1} = \mathcal{A}^{-1} + \frac{\mathcal{A}^{-1}ab^\top\mathcal{A}^{-1}}{1-b^\top\mathcal{A}^{-1}a}$

Correlations and independence

Corollary: Let $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$, then X_1 is independent of X_2 if and only if X_1 and X_2 are uncorrelated.

Proof: factorization of density ($\Sigma \geq 0$) or characteristic function.

Interestingly, for two jointly multivariate Normal vectors (i.e., $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$), *pair-wise* independence of their components implies complete independence.

The independence of two linear transforms of a multinormal X can be shown via the following corollary.

Corollary: If $X \sim N_p(\mu, \Sigma)$, $\mathcal{A}$ and $\mathcal{B}$ matrices, then $\mathcal{A}X$ and $\mathcal{B}X$ are independent if and only if $\mathcal{A}\Sigma\mathcal{B}^\top = 0$.

Marginal and conditional distributions

Marginal distribution is just a special case of linear transform:

$$X_1 = (\mathcal{I}_q \;\; 0_q \times 0_p^\top)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

For *conditional distribution* $X_2|X_1 = x_1$ we have the following:

Theorem: The conditional distribution of X_2 given $X_1 = x_1$ is normal with mean $\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ and covariance $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, i.e.,

$$(X_2 \mid X_1 = x_1) \sim N_{p-r}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22.1}).$$

Proof: e.g. via the following lemma or by factorizing the density (using the formula for inverse of partitioned matrix).

Decomposition of Normal Random Vector

Lemma:

$$\begin{aligned} X &= \tbinom{X_1}{X_2}, \qquad X_1 \in \mathbb{R}^r \\ X_{2.1} &= X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1 \end{aligned}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

$$\begin{aligned} &\Rightarrow \quad X_1 \sim N_r(\mu_1, \Sigma_{11}), \\ &\text{independent} \\ &\Rightarrow \quad X_{2.1} \sim N_{p-r}(\mu_{2.1}, \Sigma_{22.1}) \end{aligned}$$

$$\begin{aligned} \mu_{2.1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 \\ \Sigma_{22.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned}$$

Example:

$$p = 2,\ r = 1,\ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 2 \end{pmatrix}$$

$$\Sigma_{11} = 1,\ \Sigma_{21} = -0.8,\ \Sigma_{22.1} = 2 - (0.8)^2 = 1.36.$$

$$\begin{aligned} \Rightarrow f_{X_1}(x_1) &= \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{x_1^2}{2}\right) \\ \Rightarrow f(x_2 \mid x_1) &= \tfrac{1}{\sqrt{2\pi(1.36)}} \exp\left\{-\tfrac{(x_2 - 0.8x_1)^2}{2\cdot(1.36)}\right\}. \end{aligned}$$

Theorem: If $X_1 \sim N_r(\mu_1, \Sigma_{11})$ and $(X_2|X_1 = x_1) \sim N_{p-r}(\mathcal{A}x_1 + b, \Omega)$ where Ω does not depend on x_1, then

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \Sigma),$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mathcal{A}\mu + b \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{11}\mathcal{A}^\top \\ \mathcal{A}\Sigma_{11} & \Omega + \mathcal{A}\Sigma_{11}\mathcal{A}^\top \end{pmatrix}.$$

Example: $X_2 \in \mathbb{R},\ X_1 \in \mathbb{R}^r$

$$E(X_2|X_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)$$

linear approximation!

$$\begin{aligned} X_2 &= E(X_2|X_1) + U,\ U \sim N_{p-r}(0, \Sigma_{22.1}) \\ &= \beta_0 + \beta^\top X_1 + U \end{aligned}$$

$$\begin{aligned} \sigma_{22} &= \text{var}(X_2)\ \Sigma = \begin{pmatrix} \Sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \\ &= \beta^\top \Sigma_{11}\beta + \sigma_{22.1} = \sigma_{21}\Sigma_{11}^{-1}\sigma_{12} + \sigma_{22.1} \end{aligned}$$

Consider the case where $r = p - 1$.

Now $X_2 \in \mathbb{R}$ and B is a row vector $\beta^\top$ of dimension $(1 \times r)$

$$X_2 = \beta_0 + \beta^\top X_1 + U.$$

This means that the best MSE approximation of X_2 by a function of X_1 is a straight line.

The **marginal variance** of X_2 can be decomposed as

$$\sigma_{22} = \beta^\top \Sigma_{11}\beta + \sigma_{22.1} = \sigma_{21}\Sigma_{11}^{-1}\sigma_{12} + \sigma_{22.1}.$$

$$\rho^2_{2.1\ldots r} = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}}$$

is the square of the **multiple correlation** between X_2 and the r variables X_1.

Mahalanobis Transformation (Sphering)

If $X \sim N_p(\mu, \Sigma)$ then the Mahalanobis transformation is

$$Y = \Sigma^{-1/2}(X - \mu) \sim N_p(0, \mathcal{I}_p)$$

and it holds

$$Y^\top Y = (X - \mu)^\top\ \Sigma^{-1}(X - \mu) \sim \chi^2_p.$$

Notice that Y is random vector. $Y^\top Y$ is scalar which "measures the distance" between X and its expected value μ. $Y^\top Y$ can be easily used for testing (assuming that Σ is known).

In practice, we do not know Σ. The tests in this situation can be carried out using Wishart and Hotelling distributions (multivariate generalizations of χ^2 and Student's t distribution).

Elementary Properties

- If $X \sim N_p(\mu, \Sigma)$ then a linear transformation $\mathcal{A}X + c, \mathcal{A}(q \times p), c \in \mathbb{R}^q$ has distribution $N_q(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top)$.
- Two linear transformations $\mathcal{A}X$ and $\mathcal{B}X$ of $X \sim N_p(\mu, \Sigma)$ are independent if and only if $\mathcal{A}\Sigma\mathcal{B}^\top = 0$.
- If X_1 and X_2 are partitions of $X \sim N_p(\mu, \Sigma)$ then the conditional distribution of X_2 given $X_1 = x_1$ is normal again and X_1 is independent of X_2 if and only if $\Sigma_{12} = 0$.
- The conditional expectation of $(X_2|X_1)$ is a linear function for $\binom{X_1}{X_2} \sim N_p(\mu, \Sigma)$.
- The multiple correlation coefficient is the percentage of the variance of X_2 explained by the linear approximation $\beta_0 + \beta^\top X_1$.

Central Limit Theorems

Central Limit Theorem describes the (asymptotic) behaviour of sample mean

$X_1, X_2, \ldots, X_n$, i.i.d with $X_i \sim (\mu, \Sigma)$

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma) \qquad \text{for} \quad n \longrightarrow \infty.$$

The **CLT** can be easily applied for testing.

Normal distribution plays a central role in statistics.

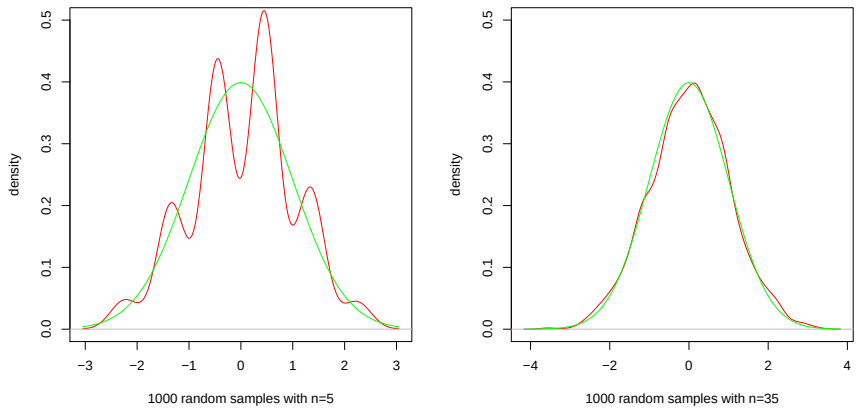


Figure: The CLT for Bernoulli distributed random variables. Sample size $n = 5$ (left) and $n = 35$ (right). → `SMScltbern`

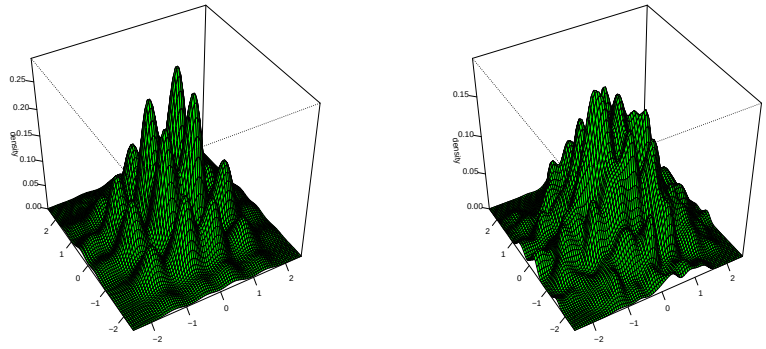


Figure: The CLT in the two-dimensional case. Sample size $n = 5$ (left) and $n = 500$ (right). → `SMScltbern3`

Transformation of statistics

If $\sqrt{n}(t - \mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma)$ and if $f = (f_1, \ldots, f_q)^\top : \mathbb{R}^p \to \mathbb{R}^q$ are real valued functions which are differentiable at $\mu \in \mathbb{R}^p$, then $f(t)$ is asymptotically normal with mean $f(\mu)$ and covariance $\mathcal{D}^\top \Sigma \mathcal{D}$, i.e.,

$$\sqrt{n}\{f(t) - f(\mu)\} \xrightarrow{\mathcal{L}} N_q(0, \mathcal{D}^\top \Sigma \mathcal{D}) \qquad \text{for} \quad n \longrightarrow \infty,$$

where

$$\mathcal{D} = \left(\frac{\partial f_j}{\partial t_i} \right)(t) \Big|_{t=\mu}$$

$(p \times q)$ matrix of all partial derivatives.

This theorem can be applied, e.g., to find the "variance stabilizing" transformation.

Example:

Suppose

$$\{X_i\}_{i=1}^n \sim (\mu, \Sigma); \quad \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad p = 2.$$

We have by CLT for $n \to \infty$

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N(0, \Sigma).$$

The distribution of $\begin{pmatrix} \overline{x}_1^2 - \overline{x}_2 \\ \overline{x}_1 + 3\overline{x}_2 \end{pmatrix}$?

This means to consider $f = (f_1, f_2)^\top$ with

$$f_1(x_1, x_2) = x_1^2 - x_2, \quad f_2(x_1, x_2) = x_1 + 3x_2, \quad q = 2.$$

Then $f(\mu) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and

$$\mathcal{D} = (d_{ij}), \quad d_{ij} = \left(\frac{\partial f_j}{\partial x_i} \right) \Big|_{x=\mu} = \begin{pmatrix} 2x_1 & 1 \\ -1 & 3 \end{pmatrix} \Big|_{x=0} = \begin{pmatrix} 0 & 1 \\ -1 & 3 \end{pmatrix}.$$

We have the covariance

$$\underset{\mathcal{D}^\top}{\begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}} \underset{\Sigma}{\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}} \underset{\mathcal{D}}{\begin{pmatrix} 0 & 1 \\ -1 & 3 \end{pmatrix}} = \underset{\mathcal{D}^\top \Sigma \mathcal{D}}{\begin{pmatrix} 1 & -\frac{7}{2} \\ -\frac{7}{2} & 13 \end{pmatrix}}.$$

This yields

$$\sqrt{n} \begin{pmatrix} \overline{x}_1^2 - \overline{x}_2 \\ \overline{x}_1 + 3\overline{x}_2 \end{pmatrix} \xrightarrow{\mathcal{L}} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\frac{7}{2} \\ -\frac{7}{2} & 13 \end{pmatrix} \right).$$

Limit Theorems

- If $X_1, \ldots, X_n$ are i.i.d. random vectors with $X_i \sim (\mu, \Sigma)$ then the distribution of $\sqrt{n}(\overline{x} - \mu)$ is asymptotically $N(0, \Sigma)$ (Central Limit Theorem).
- If $X_1, \ldots, X_n$ are i.i.d. random variables with $X_i \sim (\mu, \sigma)$ then an asymptotic confidence interval can be constructed by the CLT: $\bar{x} \pm \frac{\hat{\sigma}}{\sqrt{n}} u_{1-\alpha/2}$.
- For small sample sizes the Bootstrap improves the precision of this confidence interval.
- If t is a statistic that is asymptotically normal, i.e., $\sqrt{n}(t - \mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma)$, then this holds also for a function $f(t)$, i.e., $\sqrt{n}\{f(t) - f(\mu)\}$ is asymptotically normal.

Týden 4

Datové matice, testování mnohorozměrné střední hodnoty:

- Wishartovo a Hotellingovo rozdělení,
- testy vícerozměrné střední hodnoty.

Further matrix algebra: Kronecker product

Let $\mathcal{A} \otimes \mathcal{B}$ denote the Kronecker product of matrices $\mathcal{A}$ and $\mathcal{B}$ and $\text{vec}(\mathcal{A})$ denote the vector obtained by stacking the columns of $\mathcal{A}$. Kronecker product and vectorization are useful tools for working with (random) matrices:

- $\alpha(\mathcal{A} \otimes \mathcal{B}) = (\alpha\mathcal{A}) \otimes \mathcal{B} = \mathcal{A} \otimes (\alpha\mathcal{B})$
- $\mathcal{A} \otimes (\mathcal{B} \otimes \mathcal{C}) = (\mathcal{A} \otimes \mathcal{B}) \otimes \mathcal{C}$
- $(\mathcal{A} \otimes \mathcal{B})^\top = \mathcal{A}^\top \otimes \mathcal{B}^\top$
- $(\mathcal{A} \otimes \mathcal{B})(\mathcal{C} \otimes \mathcal{D}) = (\mathcal{A}\mathcal{C}) \otimes (\mathcal{B}\mathcal{D})$
- $(\mathcal{A} \otimes \mathcal{B})^{-1} = (\mathcal{A}^{-1} \otimes \mathcal{B}^{-1})$
- $(\mathcal{A} + \mathcal{B}) \otimes \mathcal{C} = \mathcal{A} \otimes \mathcal{C} + \mathcal{B} \otimes \mathcal{C}$
- $\mathcal{A} \otimes (\mathcal{B} + \mathcal{C}) = \mathcal{A} \otimes \mathcal{B} + \mathcal{A} \otimes \mathcal{C}$
- $\text{vec}(\mathcal{A}\mathcal{X}\mathcal{B}) = (\mathcal{B}^\top \otimes \mathcal{A})\text{vec}(\mathcal{X})$
- $\text{tr}(\mathcal{A} \otimes \mathcal{B}) = \text{tr}(\mathcal{A})\,\text{tr}(\mathcal{B})$

Normal data matrices

Definition: Let $X_1, \ldots, X_n$ be a random sample from $N_p(\mu, \Sigma)$. Then $\mathcal{X} = (X_1, \ldots, X_n)^\top$ is called a data matrix from $N_p(\mu, \Sigma)$.

Clearly, if $\mathcal{X}$ is a data matrix from $N_p(\mu, \Sigma)$ then $\bar{x}_n = \mathcal{X}^\top 1_n/n \sim N_p(\mu, \Sigma/n)$.

Theorem: If $\mathcal{X}$ is a data matrix from $N_p(\mu, \Sigma)$ then $\mathcal{Y} = \mathcal{A}\mathcal{X}\mathcal{B} \sim N_q(\alpha\mathcal{B}^\top\mu, \beta\mathcal{B}^\top\Sigma\mathcal{B})$ if and only if:

- $\mathcal{A}1_n = \alpha 1_m$ for some $\alpha \in \mathbb{R}$, or $\mathcal{B}^\top\mu = 0$, and
- $\mathcal{A}\mathcal{A}^\top = \beta\mathcal{I}_m$ for some $\beta \in \mathbb{R}$, or $\mathcal{B}^\top\Sigma\mathcal{B} = 0_q 0_q^\top$.

The proof is based on vectorization of $\mathcal{X}$ (i.e., column stacking): $\text{vec}(\mathcal{X}) \sim N_{np}(\mu \otimes 1_n, \Sigma \otimes \mathcal{I}_n)$ and $\text{vec}(\mathcal{A}\mathcal{X}\mathcal{B}) = (\mathcal{B}^\top \otimes \mathcal{A})\text{vec}(\mathcal{X})$.

Normal data matrices: independence

Theorem: If $\mathcal{X}$ is a data matrix from $N_p(\mu, \Sigma)$ then $\mathcal{Y} = \mathcal{A}\mathcal{X}\mathcal{B}$ and $\mathcal{Z} = \mathcal{C}\mathcal{X}\mathcal{D}$ are independent if and only if $\mathcal{B}^\top\Sigma\mathcal{D} = 0$ or $\mathcal{A}\mathcal{C}^\top = 0$.

Proof: assume (WLOG) that $\mu = 0$, then $\text{vec}(\mathcal{Y}) = (\mathcal{B}^\top \otimes \mathcal{A})\text{vec}(\mathcal{X})$, and the covariance matrix between $\text{vec}(\mathcal{Y})$ and $\text{vec}(\mathcal{Z})$ is

$$\begin{aligned} E\text{vec}(\mathcal{Y})\text{vec}(\mathcal{Z})^\top &= (\mathcal{B}^\top \otimes \mathcal{A})E\text{vec}(\mathcal{X})\text{vec}(\mathcal{X})^\top(\mathcal{D}^\top \otimes \mathcal{C})^\top \\ &= (\mathcal{B}^\top \otimes \mathcal{A})(\Sigma \otimes \mathcal{I}_n)(\mathcal{D} \otimes \mathcal{C}^\top) \\ &= \mathcal{B}^\top\Sigma\mathcal{D} \otimes \mathcal{A}\mathcal{C}^\top. \end{aligned}$$

Wishart distribution

Definition: Assuming that $\mathcal{X}$ is a data matrix from $N_p(0_p, \Sigma)$, the random matrix

$$\mathcal{M}(p \times p) = \mathcal{X}^\top\mathcal{X} \sim W_p(\Sigma, n),$$

where $W_p(\Sigma, n)$ denotes Wishart distribution with parameters Σ and n.

Example:

$$p = 1, \qquad X \sim N_1(0, \sigma^2)$$

$$\mathcal{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \mathcal{M} = \mathcal{X}^\top\mathcal{X} = \sum_{i=1}^n x_i^2 \sim \sigma^2\chi_n^2$$

It follows that Wishart distribution is generalisation of χ_n^2

Theorem:

$$\mathcal{M} \sim W_p(\Sigma, n) \text{ and } \mathcal{B}(p \times q) \Rightarrow \mathcal{B}^\top \mathcal{M} \mathcal{B} \sim W_q(\mathcal{B}^\top \Sigma \mathcal{B}, n)$$

Theorem:(Cochran) $\mathcal{X}(n \times p)$ is data matrix with $N_p(\mu, \Sigma)$. Then:

- $\mathcal{X}^\top \mathcal{C} \mathcal{X}$, where $\mathcal{C}$ is symmetric, has the same distribution as a weighted sum of independent $W_p(\Sigma, 1)$ matrices, where the weights are eigenvalues of $\mathcal{C}$.
- $n\mathcal{S} = \mathcal{X}^\top \mathcal{H} \mathcal{X} \sim W_p(\Sigma, n-1)$,
- $\bar{x}$ and $\mathcal{S}$ are independent.

Proof: see Theorem 3.4.4. (page 68) in MKB (using spectral decomposition of $\mathcal{C}$ or $\mathcal{H}$).

Wilks' Λ-distribution

Definition: Assume that $\mathcal{A} \sim W_p(\mathcal{I}, m)$ and $\mathcal{B} \sim W_p(\mathcal{I}, n)$ are indepependent, $m \geq p$, we say that the random variable

$$\Lambda = |\mathcal{A}|/|\mathcal{A} + \mathcal{B}|$$

has Wilks' lambda distribution with parameters p, m, and n, i.e., $\Lambda \sim \Lambda(p, m, n)$

This distribution occurs frequently in likelihood ratio tests.

The random variable Λ is basically a ratio of two 'generalized variances'—therefore, Wilks' Λ distribution can be seen as a multivariate generalization of F distribution.

Hotelling's T^2-distribution

Definition: Assume that random vector $Y \sim N_p(0, \mathcal{I})$ is independent of random matrix $\mathcal{M} \sim W_p(\mathcal{I}, n)$. Then

$$n \; Y^\top \mathcal{M}^{-1} Y \sim T^2(p, n),$$

where $T^2(p, n)$ denotes Hotelling's distribution with parameters p and n.

Hotelling's T^2 generalizes Student's t-distribution

The critical values of Hotelling's T^2 can be calculated using F-distribution:

$$T^2(p, n) = \frac{np}{n-p+1} \; F_{p,n-p+1}$$

Distributions related to multinormal

- The Wishart distribution is a generalization of the χ^2-distribution.
- Assuming normality, the empirical covariance matrix $\mathcal{S}$ has a $\frac{1}{n} W_p(\Sigma, n-1)$ distribution.
- In the normal case, $\bar{x}$ and $\mathcal{S}$ are independent.
- Hotelling's T^2-distribution is a generalization of the t-distribution.
- $(n-1)(\bar{x}-\mu)^\top \mathcal{S}^{-1}(\bar{x}-\mu)$ has a $T^2(p, n-1)$ distribution.
- The relation between Hotelling's T^2- and Fisher's F-distribution is given by $T^2(p, n) = \frac{np}{n-p+1} \; F_{p,n-p+1}$.
- Wilks' Λ-distribution can be seen as a multivariate generalization of F distribution (ratio of two variances).

Testing the multivariate mean

$X_i \sim N_p(\mu, \Sigma)$ i.i.d.

$$H_0 : \mu = \mu_0, \quad \Sigma \text{ unknown}, \qquad H_1 : \quad \text{no constraints.}$$

Under H_0: $(n-1)(\bar{x}-\mu_0)^\top \mathcal{S}^{-1}(\bar{x}-\mu_0) \sim T^2(p, n-1)$.

Equivalently:

$$\left(\frac{n-p}{p}\right)(\bar{x}-\mu_0)^\top \mathcal{S}^{-1}(\bar{x}-\mu_0) \sim F_{p,n-p}$$

The rejection region may be defined as

$$\left(\frac{n-p}{p}\right)(\bar{x}-\mu_0)^\top \mathcal{S}^{-1}(\bar{x}-\mu_0) > F_{1-\alpha;p,n-p}.$$

R: `library(DescTools); help(HotellingsT2Test)`

Confidence region for μ

$$\left(\tfrac{n-p}{p}\right)(\bar{x}-\mu)^\top \mathcal{S}^{-1}(\bar{x}-\mu) \sim F_{p,n-p}$$

$$\left\{\mu \in I\!R^p \mid (\mu-\bar{x})^\top \mathcal{S}^{-1}(\mu-\bar{x}) \le \frac{p}{n-p} F_{1-\alpha;p,n-p}\right\}$$

is a confidence region at level (1-α) for μ; it is the interior of an iso-distance ellipsoid in $I\!R^p$.

When p is large, ellipsoids are not easy to handle for practical purposes. One is thus interested in finding confidence intervals for $\mu_1, \mu_2, \ldots, \mu_p$ so that simultaneous confidence on all the intervals reaches the desired level say, $1-\alpha$.


```
library(mvtnorm)
s=matrix(c(1,-0.5,-0.5,1),2);x=seq(-3,3,by=0.015)
contour(x,x,outer(x,x,
        function(x,y){dmvnorm(cbind(x,y),sigma=s)}))

n=20;X=rmvnorm(n,sigma=s);m=apply(X,2,mean);S=cov(X)
points(m[1],m[2],pch=8,col="red",cex=2)

#contour(x,x,outer(x,x,function(x,y){(n-2)*
#  diag(t(t(cbind(x,y))-m)%*%solve(S)%*%(t(cbind(x,y))-m))<
#  2*qf(0.95,2,n-2)}),col="red",add=TRUE)
S1=solve(S)
contour(x,x,outer(x,x,function(x,y){(n-2)*
  apply(t(t(cbind(x,y))-m),1,function(x){t(x)%*%S1%*%x})<
  2*qf(0.95,2,n-2)}),col="red",add=TRUE)

bodyx=m[1]+c(-1,1)*sqrt(S[1,1]*2*qf (0.95,2,n-2) /(n-2))
bodyy=m[2]+c(-1,1)*sqrt(S[2,2]*2*qf (0.95,2,n-2) /(n-2))
polygon(x=bodyx[c(1,1,2,2,1)],y=bodyy[c(1,2,2,1,1)],border="blue")
```


Simultaneous Confidence Intervals for $a^\top \mu$

Obvious confidence interval for certain $a^\top \mu$ is given by:

$$\left|\frac{\sqrt{n-1}(a^\top \mu - a^\top \bar{x})}{\sqrt{a^\top \mathcal{S} a}}\right| \le t_{1-\frac{\alpha}{2};n-1}$$

or equivalently

$$t^2(a) = \frac{(n-1)\left\{a^\top(\mu-\bar{x})\right\}^2}{a^\top \mathcal{S} a} \le F_{1-\alpha;1,n-1}$$

which provides the $(1-\alpha)$ confidence interval for $a^\top \mu$:

$$\left(a^\top \bar{x} - \sqrt{F_{1-\alpha;1,n-1}\frac{a^\top \mathcal{S} a}{n-1}} \le a^\top \mu \le a^\top \bar{x} + \sqrt{F_{1-\alpha;1,n-1}\frac{a^\top \mathcal{S} a}{n-1}}\right).$$

Using Theorem on maximum of quadratic forms we see that:

$$\max_a t^2(a) = (n-1)(\bar{x}-\mu)^\top \mathcal{S}^{-1}(\bar{x}-\mu) \sim T^2(p, n-1).$$

$$\max_a t^2(a) = (n-1)(\bar{x}-\mu)^\top \mathcal{S}^{-1}(\bar{x}-\mu) \sim T^2(p, n-1)$$

implies that the **simultaneous confidence intervals** for all possible linear combinations $a^\top\mu$, $a \in \mathbb{R}^p$ of the elements of μ is given by:

$$\left(a^\top\bar{x} - \sqrt{K_\alpha a^\top \mathcal{S} a},\ a^\top\bar{x} + \sqrt{K_\alpha a^\top \mathcal{S} a}\right),$$

where $K_\alpha = \frac{p}{n-p}F_{1-\alpha;p,n-p}$.

Example:

95% confidence region for μ_f, the mean of the forged banknotes, is given by the ellipsoid:

$$\left\{\mu \in \mathbb{R}^6 \left| (\mu-\bar{x}_f)^\top S_f^{-1}(\mu-\bar{x}_f) \le \frac{6}{94}F_{0.95;6,94}\right.\right\}$$

95% simultaneous c.i. are given by (using $F_{0.95;6,94} = 2.1966$)

$$\begin{array}{rcccr} 214.692 & \le & \mu_1 & \le & 214.954 \\ 130.205 & \le & \mu_2 & \le & 130.395 \\ 130.082 & \le & \mu_3 & \le & 130.304 \\ 10.108 & \le & \mu_4 & \le & 10.952 \\ 10.896 & \le & \mu_5 & \le & 11.370 \\ 139.242 & \le & \mu_6 & \le & 139.658 \end{array}$$

Comparison with $\mu_0 = (214.9, 129.9, 129.7, 8.3, 10.1, 141.5)^\top$ shows that almost all components (except the first one) are responsible for the rejection of μ_0.

In addition, choosing e.g. $a^\top = (0,\ 0,\ 0,\ 1,\ -1,\ 0)$ gives c.i. $-1.211 \le \mu_4 - \mu_5 \le 0.005$ shows that for the forged bills, the lower border is essentially smaller than the upper border.

Testing the difference of two multivariate means

Suppose $X_{i1} \sim N_p(\mu_1, \Sigma), i = 1, \cdots, n_1$ and $X_{j2} \sim N_p(\mu_2, \Sigma), j = 1, \cdots, n_2$, all the random vectors being independent.

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \text{ no constraints.}$$

Both samples provide the statistics $\bar{x}_k$ and $\mathcal{S}_k$, k=1,2.

Let $\delta = \mu_1 - \mu_2$,we have

$$(\bar{x}_1 - \bar{x}_2) \sim N_p\left(\delta, \frac{n_1+n_2}{n_1 n_2}\Sigma\right)$$

$$n_1 S_1 + n_2 S_2 \sim W_p(\Sigma, n_1 + n_2 - 2).$$

The rejection region is:

$$\frac{n_1 n_2 (n_1+n_2-p-1)}{p(n_1+n_2)^2}\left((\bar{x}_1-\bar{x}_2)\right)^\top \mathcal{S}^{-1}\left((\bar{x}_1-\bar{x}_2)\right)$$

$$\ge F_{1-\alpha;p,n_1+n_2-p-1}$$

A $(1-\alpha)*100\%$ confidence region for δ is given by the ellipsoid centered at $(\bar{x}_1 - \bar{x}_2)$

$$(\delta - (\bar{x}_1-\bar{x}_2))^\top \mathcal{S}^{-1}(\delta - (\bar{x}_1-\bar{x}_2))$$

$$\le \frac{p(n_1+n_2)^2}{(n_1+n_2-p-1)(n_1 n_2)}F_{1-\alpha;p,n_1+n_2-p-1},$$

and the simultaneous confidence intervals for all linear combinations of the elements of δ : $a^\top\delta$ are given by

$$a^\top\delta \in a^\top(\bar{x}_1-\bar{x}_2) \pm \sqrt{\frac{p(n_1+n_2)^2}{(n_1+n_2-p-1)(n_1 n_2)}F_{1-\alpha;p,n_1+n_2-p-1}a^\top\mathcal{S}a}.$$

Example: We want to compare the mean of the assets (X_1) and of the sales (X_2) of the two sectors energy (group 1) and manufacturing (group 2).

We have the following statistics $n_1 = 15,\ n_2 = 10,\ p = 2$,
$\bar{x}_1 = \begin{pmatrix} 4084 \\ 2580.5 \end{pmatrix}$, $\bar{x}_1 = \begin{pmatrix} 4307.2 \\ 4925.2 \end{pmatrix}$,
$\mathcal{S}_1 = 10^7 * \begin{pmatrix} 1.6635 & 1.2410 \\ 1.2410 & 1.3747 \end{pmatrix}$,
$\mathcal{S}_2 = 10^7 * \begin{pmatrix} 1.2248 & 1.1425 \\ 1.1425 & 1.5112 \end{pmatrix}$,
$\mathcal{S} = 10^7 * \begin{pmatrix} 1.4880 & 1.2016 \\ 1.2016 & 1.4293 \end{pmatrix}$.

The observed value of the test statistic is $F_{\text{obs}} = 2.7036$.

Since $F_{0.95;2,22} = 3.4434$ the hypothesis of equal means of the two groups is not rejected although it would be rejected at a less severe level ($p - \text{value} = 0.0892$).

The 95% simultaneous confidence intervals for the differences are given by

$$\begin{array}{rcrcr} -4628.6 & \leq & \mu_{1a} - \mu_{2a} & \leq & 4182.2 \\ -6662.4 & \leq & \mu_{1s} - \mu_{2s} & \leq & 1973.0. \end{array}$$

Testing means with unequal covariance matrices I

Suppose $X_{i1} \sim N_p(\mu_1, \Sigma_1), i = 1, \cdots, n_1$ and
$X_{j2} \sim N_p(\mu_2, \Sigma_2), j = 1, \cdots, n_2$, all the variables being independent.

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \quad \text{no constraints.}$$

$$(\bar{x}_1 - \bar{x}_2) \sim N_p\left(\delta, \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right).$$

Therefore,

$$(\bar{x}_1 - \bar{x}_2)^\top \left(\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right)^{-1} (\bar{x}_1 - \bar{x}_2) \sim \chi^2_p$$

Since $\mathcal{S}_i$ is a consistent estimator of $\Sigma_i,\ i = 1, 2$ we have

$$(\bar{x}_1 - \bar{x}_2)^\top \left(\frac{\mathcal{S}_1}{n_1} + \frac{\mathcal{S}_2}{n_2}\right)^{-1} (\bar{x}_1 - \bar{x}_2) \to \chi^2_p$$

Example: Let us compare the forged and the genuine bank notes again (n_1 and n_2 are both large). The test statistic turns out to be 2436.8 which is again highly significant. The 95% simultaneous confidence intervals are now:

$$\begin{array}{rcrcr} -0.0389 & \leq & \delta_1 & \leq & 0.3309 \\ -0.5140 & \leq & \delta_2 & \leq & -0.2000 \\ -0.6368 & \leq & \delta_3 & \leq & -0.3092 \\ -2.6846 & \leq & \delta_4 & \leq & -1.7654 \\ -1.2858 & \leq & \delta_5 & \leq & -0.6442 \\ 1.8146 & \leq & \delta_6 & \leq & 2.3194 \end{array}$$

showing that all the components except the first are different from zero, the larger difference coming from X_6 (length of the diagonal) and X_4 (lower border).

Testing means with unequal covariance matrices II

Clearly, the χ^2 approximation to the distribution of the test statistic

$$(\bar{x}_1 - \bar{x}_2)^\top \left(\frac{\mathcal{S}_1}{n_1} + \frac{\mathcal{S}_2}{n_2} \right)^{-1} (\bar{x}_1 - \bar{x}_2)$$

is usable only for sufficiently large sample sizes.

For smaller sample sizes, one can use approximate likelihood ratio tests (Mardia et al, Section 5.4.1) or Welch approximation to degrees of freedom (Mardia et al, Section 5.4.2).

Note: the problem of testing equality of means without equality of variances is known as the Behrens-Fisher problem (at least in the univariate case).

Týden 5

Odhadování a testování:

- odhady metodou maximální věrohodnosti,
- testování poměrem věrohodností,
- příklady.

Estimation

The aim is to estimate vector of parameters θ from a sample $\mathcal{X}$ through estimators $\widehat{\theta}(\mathcal{X})$.

Most common approaches:

- maximum likelihood,
- Bayesian approach,
- robust methods (M-estimation).

In the following, we shortly discuss maximum likelihood theory.

The Likelihood Function

$X \sim f(x, \theta)$ pdf. parameter θ

Likelihood function

$$L(\mathcal{X}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

MLE

$$\widehat{\theta} = \arg\max_{\theta} L(\mathcal{X}; \theta)$$

log-likelihood

$$\ell(\mathcal{X}; \theta) = \log L(\mathcal{X}; \theta)$$

Derivatives

Function $f : I\!R^p \to I\!R$

$\frac{\partial f(x)}{\partial x}$ is the gradient, i.e., column vector of partial derivatives $\left\{\frac{\partial f(x)}{\partial x_j}\right\}, j = 1, \ldots, p$

$\frac{\partial f(x)}{\partial x^\top}$ row vector of the same derivative

$\frac{\partial^2 f(x)}{\partial x \partial x^\top}$ is the $(p \times p)$ Hessian matrix of second derivatives $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}, i = 1, \ldots, p,\ j = 1, \ldots, p.$

Some useful formulae

Linear transformations:

$\mathcal{A}(p \times p), x \in I\!R^p$

$$\frac{\partial a^\top x}{\partial x} = \frac{\partial x^\top a}{\partial x} = a$$

Quadratic form (i.e., $\mathcal{A}$ is symmetric):

$$\frac{\partial x^\top \mathcal{A} x}{\partial x} = (\mathcal{A} + \mathcal{A}^\top)x = 2\mathcal{A}x$$

$$\frac{\partial^2 x^\top \mathcal{A} x}{\partial x \partial x^\top} = 2\mathcal{A}$$

Derivative of trace and determinant

This is useful for derivation of MLEs for multivariate normal distribution:

$$\frac{\partial \mathrm{tr}\, \mathcal{X}\mathcal{A}}{\partial \mathcal{X}} = \begin{cases} \mathcal{A}^\top & \text{if elements of } \mathcal{A} \text{ are distinct,} \\ \mathcal{A} + \mathcal{A}^\top - \mathrm{diag}(\mathcal{A}) & \text{for } \mathcal{A} \text{ symmetric.} \end{cases}$$

$$\frac{\partial |\mathcal{X}|}{\partial x_{ij}} = x_{ij} \text{ if elements of } \mathcal{X} \text{ are distinct.}$$

$$\frac{\partial |\mathcal{X}|}{\partial x_{ij}} = \begin{matrix} x_{ij} & \text{for } i = j \\ 2x_{ij} & \text{for } i \neq j \end{matrix} \quad \text{for } \mathcal{X} \text{ symmetric.}$$

For $\mathcal{V} = \Sigma^{-1}$ symmetric it follows that:

$$\frac{\partial \log |\mathcal{V}|}{\partial \mathcal{V}} = 2\Sigma - \mathrm{diag}(\Sigma).$$

Derivatives

- The column vector $\frac{\partial f(x)}{\partial x}$ is called the gradient.
- The gradient of $\frac{\partial a^\top x}{\partial x} = \frac{\partial x^\top a}{\partial x}$ equals a.
- The derivative of the quadratic form $\frac{\partial x^\top \mathcal{A} x}{\partial x}$ equals $2\mathcal{A}x$.
- The Hessian of $f : I\!R^p \to I\!R$ is the $(p \times p)$ matrix of second derivatives $\frac{\partial f(x)}{\partial x_i \partial x_j}$.
- The Hessian of the quadratic form $x^T \mathcal{A} x$ equals $2\mathcal{A}$.

Example: $\{x_i\}_{i=1}^n$ is a sample from a normal distribution $N_p(\mu, \Sigma)$

Due to the symmetry of Σ, the unknown parameter θ is in fact $\{p + \frac{1}{2}p(p+1)\}$-dimensional.

$$L(\mathcal{X}; \theta) = |2\pi\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu)\right\}$$

$$\ell(\mathcal{X}; \theta) = -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu).$$

After some calculations, the log-likelihood function for $N_p(\mu, \Sigma)$ is:

$$\ell(\mathcal{X}; \theta) = -\frac{n}{2}\log|2\pi\Sigma| - \frac{n}{2}\mathrm{tr}\{\Sigma^{-1}\mathcal{S}\} - \frac{n}{2}(\overline{x} - \mu)^\top \Sigma^{-1}(\overline{x} - \mu)$$

Reparametrizing $\mathcal{V} = \Sigma^{-1}$, we obtain:

$$\frac{\partial \ell(\mathcal{X}; \theta)}{\partial \mu} = n\mathcal{V}(\overline{x} - \mu)$$

and

$$\frac{\partial \ell(\mathcal{X}; \theta)}{\partial \mathcal{V}} = n\{2\mathcal{M} - \mathrm{diag}(\mathcal{M})\}/2,$$

where $\mathcal{M} = \Sigma - \mathcal{S} - (\overline{x} - \mu)(\overline{x} - \mu)^\top$.

Score and Fisher information

The score function $s(\mathcal{X}; \theta)$ is the derivative of the log-likelihood function w.r.t. $\theta \in \mathbb{R}^k$

$$s(\mathcal{X}; \theta) = \frac{\partial}{\partial\theta}\ell(\mathcal{X}; \theta) = \frac{1}{L(\mathcal{X}; \theta)}\frac{\partial}{\partial\theta}L(\mathcal{X}; \theta).$$

The covariance matrix

$$\mathcal{F}_n = E\{s(\mathcal{X}; \theta)s(\mathcal{X}; \theta)^\top\} = \mathrm{Var}\{s(\mathcal{X}; \theta)\} = -E\left\{\frac{\partial^2}{\partial\theta\partial\theta^\top}\ell(\mathcal{X}; \theta)\right\}$$

is called the Fisher information matrix.

Cramer-Rao theorem

The importance of the Fisher information matrix is explained by the Cramer-Rao theorem, which gives the lower bound for the variance matrix for any unbiased estimator of θ.

Theorem: If $\widehat{\theta} = t = t(\mathcal{X})$ is an unbiased estimator for θ, then under regularity conditions

$$\mathrm{Var}(t) \geq \mathcal{F}_n^{-1}.$$

The proof can be based on some special properties of the score function.

An unbiased estimator with the variance equal to $\mathcal{F}_n^{-1}$ is called a minimum variance unbiased estimator.

Asymptotic normality of MLEs

Another important result says that the MLE is asymptotically unbiased, efficient (minimum variance), and normally distributed.

Theorem: Suppose that the sample $\{x_i\}_{i=1}^n$ is i.i.d. If $\widehat{\theta}$ is the MLE for $\theta \in \mathbb{R}^k$ then under some regularity conditions, as $n \to \infty$:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{L}} N_k(0, \mathcal{F}_1^{-1}),$$

where $\mathcal{F}_1$ denotes the Fisher information for sample size $n = 1$.

This result gives us a very useful and simple approximation whenever we are not able to calculate the exact distribution of the MLE $\widehat{\theta}$.

Even in very complicated situations, the Fisher information matrix can be approximated numerically.

Hypothesis Testing

- Maximum likelihood estimators are easy to calculate but we have to know the true distribution.
- MLEs have asymptotically normal distribution.
- The asymptotic normality of transformed MLEs can be derived by using Delta theorem.
- MLEs are asymptotically optimal.

Likelihood ratio tests (LRTs)

Consider hypotheses:

$$\begin{aligned} H_0 &: \theta \in \Omega_0, \\ H_1 &: \theta \in \Omega_1, \end{aligned}$$

where θ is a parameter of the distribution of $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$.

Wilks' Theorem says:

Theorem: If $\Omega_1 \subset \mathbb{R}^q$ is a q-dimensional space and if $\Omega_0 \subset \Omega_1$ is an r-dimensional subspace, then under regularity conditions:

$$\forall\ \theta \in \Omega_0 : -2\log\lambda = 2(\ell_1^* - \ell_0^*) \xrightarrow{\mathcal{L}} \chi^2_{q-r} \qquad \text{as} \quad n \to \infty,$$

where ℓ_j^*, $j = 1, 2$ are the maxima of the log-likelihood for each hypothesis.

Testing the multivariate mean

$X_i \sim N_p(\mu, \Sigma)$ i.i.d.

$$H_0 : \mu = \mu_0, \quad \Sigma \text{ unknown}, \qquad H_1 : \quad \text{no constraints.}$$

Under H_0 it can be shown that

$$\ell_0^* = \ell(\mu_0, \mathcal{S} + dd^\top), \quad d = (\overline{x} - \mu_0)$$

and under H_1 we have

$$\ell_1^* = \ell(\overline{x}, \mathcal{S}).$$

This leads to

$$-2\log\lambda = 2(\ell_1^* - \ell_0^*) = n\log(1 + d^\top \mathcal{S}^{-1} d).$$

Note that this statistic depends on $(n-1)d^\top \mathcal{S}^{-1} d$ which has, under H_0, a Hotelling's T^2-distribution.

Testing the variance matrix

$X_i \sim N_p(\mu, \Sigma)$ i.i.d.

$$H_0 : \Sigma = \Sigma_0, \quad \mu \text{ unknown}, \qquad H_1 : \quad \text{no constraints.}$$

Looking at the log-likelihood function, we observe that the MLEs are $\overline{x}$ and Σ_0 under H_0 and $\overline{x}$ and $\mathcal{S}$ under the alternative.

Therefore:

$$l_0^* = -\frac{1}{2} n \log |2\pi\Sigma_0| - \frac{1}{2} n \operatorname{tr} \Sigma_0^{-1}\mathcal{S}, \text{ and } l_1^* = -\frac{1}{2} n \log |2\pi\mathcal{S}| - \frac{1}{2} np.$$

This leads to

$$-2 \log \lambda = n \operatorname{tr} \Sigma_0^{-1}\mathcal{S} - n \log |\Sigma_0^{-1}\mathcal{S}| - np = np(a - \log g - 1),$$

where a and g denote respectively the arithmetic and geometric mean of eigenvalues of $\Sigma_0^{-1}\mathcal{S}$.

Test of homogeneity of covariances

Let $X_{ih} \sim N_p(\mu_h, \Sigma_h)$, $i = 1 \cdots, n_h$; $h = 1, \cdots, k$
all variables being independent,

$$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k, \qquad H_1 : \quad \text{no constraints.}$$

$\mathcal{S}_h$ is the MLE estimator of Σ_h under the alternative and the weighted average $\mathcal{S} = \frac{n_1\mathcal{S}_1 + \cdots + n_k\mathcal{S}_k}{n}$ is the MLE of Σ under the null (H_0).

The likelihood ratio test leads to the statistic

$$-2 \log \lambda = n \log |\mathcal{S}| - \sum_{h=1}^{k} n_h \log |\mathcal{S}_h|$$

which under H_0 is approximately distributed as a $\mathcal{X}^2_m$ where $m = \frac{1}{2}(k-1)p(p+1)$.

Test of independence

Let $X_i \sim N_p(\mu, \Sigma)$, $i = 1 \cdots, n$ be independent,

$$H_0 : \Sigma_{12} = 0, \qquad H_1 : \quad \text{no constraints.}$$

We partition the variables into two sets with dimensions p_1 and p_2. The estimators under H_0 is

$$\hat{\mu} = \overline{x}_n$$

and

$$\hat{\Sigma} = \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix}$$

It follows that the LRT test statistics for $H_0 : \Sigma_{12} = 0$ is:

$$\begin{aligned} -2 \log \lambda &= -n \log |\hat{\Sigma}^{-1} S| = -n \log |S_{22} - S_{21} S_{11}^{-1} S_{12}| / |S_{22}| \\ &= -n \log |\mathcal{I} - S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}| = -n \log \prod_{i=1}^{k} (1 - \lambda_i), \end{aligned}$$

where λ_i are non-zero eigenvalues of $S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$.

It can be shown that the test statistics follows the so-called *Wilks' lambda* distribution (distribution of a ratio of determinants of independent Wishart matrices).

This test is applicable in canonical correlation analysis (investigating correlations between two sets of variables).

For $p_1 = 1$, the LRT test statistics simplifies to a function of multiple correlation coefficient.

Tests of multivariate normality

Multivariate skewness:

$$\beta_{1,p} = E\{(X-\mu)^\top \Sigma^{-1}(Y-\mu)\}^3,$$

where X and Y are iid.

Multivariate kurtosis:

$$\beta_{2,p} = E\{(X-\mu)^\top \Sigma^{-1}(X-\mu)\}^2$$

It can be shown that $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$ for $X \sim N_p(\mu, \Sigma)$. This easily follows from the symmetry of $V = (X-\mu)^\top \Sigma^{-1}(Y-\mu)$ and $(X-\mu)^\top \Sigma^{-1}(X-\mu) \sim \chi^2_p$.

Tests of multivariate normality

Assuming normality, the distribution of $b_{1,p}$ and $b_{2,p}$ (the sample counterparts of $\beta_{1,p}$ and $\beta_{2,p}$) is:

$$\frac{1}{6} n b_{1,p} \sim \chi^2_{p(p+1)(p+1)/6}$$

and

$$\sqrt{n}\frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)}} \sim N(0,1).$$

Note: QQ diagram can be plotted using quantiles of χ^2_p distribution and ordered values of $n(X_i - \overline{x})^\top S^{-1}(X_i - \overline{x})$.

Týden 6–7

Metoda hlavních komponent:

- definice a interpretace,
- standardizace,
- asymptotické vlastnosti,
- použití.

Principal Components

Principal components are (orthogonal) linear combinations maximizing the variance of standardized linear combinations (SLC):

$$\delta^\top X = \sum_{j=1}^{p} \delta_j X_j \text{ such that } ||\delta|| = \delta^\top \delta = 1.$$

Maximizing:

$$\max_{\{\delta: \|\delta\|=1\}} \text{var}(\delta^\top X) = \max_{\{\delta: \|\delta\|=1\}} \delta^\top \text{Var}(X)\delta$$

is easy using the spectral decomposition $\text{Var}(X) = \Gamma\Lambda\Gamma^\top$.

Example:

Bivariate normal distribution $N(0, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \rho > 0.$

Eigenvalues of this matrix are $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$ with corresponding eigenvectors

$$\gamma_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \gamma_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The PC transformation is thus

$$Y = \Gamma^\top (X - \mu) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} X$$

or

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}.$$

The first principal component is

$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$$

and the second is

$$Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2).$$

Let us compute the variances of these PCs:

$$\begin{aligned} \mathrm{var}(Y_1) &= \mathrm{var}\left\{\frac{1}{\sqrt{2}}(X_1 + X_2)\right\} = \frac{1}{2}\mathrm{var}(X_1 + X_2) \\ &= \frac{1}{2}\{\mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2)\} \\ &= \frac{1}{2}(1 + 1 + 2\rho) = 1 + \rho \\ &= \lambda_1. \end{aligned}$$

Similarly we find that: $\mathrm{var}(Y_2) = \lambda_2$.

Properties of PCs

Let $X \sim (\mu, \Sigma)$ and let Y be the PC transformation $Y = \Gamma^\top (X - \mu)$.

Then

$$\begin{aligned} EY &= 0_p \\ \mathrm{var}(Y) &= \Lambda \\ \mathrm{var}(Y_1) &\geq \cdots \geq \mathrm{var}(Y_p) \geq 0 \\ \sum_j \mathrm{var}(Y_j) &= \sum_j \lambda_j = \mathrm{tr}(\Sigma) = \sum_j \mathrm{var}(X_j) \\ \prod \mathrm{var}(Y_j) &= |\Sigma|. \end{aligned}$$

Note: $|\Sigma|$ is called the (population) generalized variance and $\mathrm{tr}(\Sigma)$ the (population) total variation.

Example: In practice, the *sample* principal components are calculated from the sample variance matrix:

$$\begin{aligned} \mathcal{S} &= \mathcal{G}\mathcal{L}\mathcal{G}^\top \\ \mathcal{Y} &= (\mathcal{X} - 1_n \bar{x}^\top)\mathcal{G} \end{aligned}$$

```
data(bank2)
eigen(var(bank2))

pcb=prcomp(bank2)
pcb
plot(pcb)
pcb$x
```

PCA stopping rules

For dimension reduction, the number of PCs is usually chosen by simple ad-hoc rules:

- scree-plot (of eigenvalues),
- log-eigenvalue diagram (LEV),
- percentage of total variation (explain 80 or 90% of total variation),
- Kaiser criterion (choose PCs with higher than the "average variance").

The interpretation of the Kaiser criterion simplifies for standardized data sets: $\operatorname{tr}\Sigma = p$ implies that the average variance is 1.

Interpretation

The interpretation of PCs is based on its variances (eigenvalues) and its coefficients (eigenvectors).

Warning: rescaling can change everything.

Example: bank2[,1]=bank2[,1]*1000; prcomp(bank2)

$$\operatorname{Cov}(X, Y) = \Sigma\Gamma = \Gamma\Lambda\Gamma^\top\Gamma = \Gamma\Lambda$$

$$\rho_{X_i Y_j} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_i X_i}}\right)^{1/2}$$

Interestingly $\sum_j \rho^2_{X_i Y_j} = \cdots = 1$.

In practice, one should consider standardization of variables before running PCA.

```
prcomp(bank2,scale.=TRUE)
```

Example: Some examples:

- bank2,
- athletic records,
- geopol,
- timebudget.

Asymptotic properties

Theorem: For normal data and Σ with distinct eigenvalues, the sample principal components and sample eigenvalues are the maximum likelihood estimators of the (true) principal components and eigenvalues.

Proof: The theorem follows from the inveriance of maximum likelihood estimators (and because $\mathcal{S}$ is MLE of Σ).

Theorem: Assume that $\Sigma = \Gamma\Lambda\Gamma^\top > 0$ with with distinct eigenvalues and $\mathcal{U} = \mathcal{G}\mathcal{L}\mathcal{G}^\top \sim m^{-1}W_p(\Sigma, m)$. Then

$$\sqrt{m}(\ell - \lambda) \xrightarrow{\mathcal{L}} N_p(0, 2\Lambda^2)$$

and

$$\sqrt{n-1}(g_j - \gamma_j) \xrightarrow{\mathcal{L}} N_p\left(0, \lambda_j \sum_{k\neq j} \lambda_k \gamma_k \gamma_k^\top / (\lambda_k - \lambda_j)^2\right).$$

The proof uses transformation $m\Gamma^\top \mathcal{U}\Gamma \sim W_p(\Lambda, m)$, see MKB, p. 231.

Example:

Assuming normality:

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

$\sqrt{n-1}(\ell_j - \lambda_j) \xrightarrow{\mathcal{L}} N(0, 2\lambda_j^2), \quad j = 1, \ldots, p,$

using log transformation:

$$\sqrt{\frac{n-1}{2}}(\log \ell_j - \log \lambda_j) \xrightarrow{\mathcal{L}} N(0,1)$$

Clearly, the estimator of variance explained by first q PCs $\widehat{\psi} = (\ell_1 + \cdots + \ell_q)/\sum_{j=1}^{p} \ell_j$ is a nonlinear transformation of ℓ.

Therefore,

$$\sqrt{n-1}(\widehat{\psi} - \psi) \xrightarrow{\mathcal{L}} N(0, \omega^2),$$

where

$$\begin{aligned} \omega^2 &= \frac{2}{\{\mathrm{tr}(\Sigma)\}^2}\left\{(1-\psi)^2(\lambda_1^2 + \cdots + \lambda_q^2) + \psi^2(\lambda_{q+1}^2 + \cdots + \lambda_p^2)\right\} \\ &= \frac{2\,\mathrm{tr}(\Sigma^2)}{\{\mathrm{tr}(\Sigma)\}^2}(\psi^2 - 2\beta\psi + \beta), \end{aligned}$$

where $\beta = (\lambda_1^2 + \cdots + \lambda_q^2)/(\lambda_1^2 + \cdots + \lambda_p^2)$.

Remark: use $\mathrm{tr}(\Lambda) = \mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Lambda^2) = \mathrm{tr}(\Sigma^2)$ to simplify the calculation!

Example: The first PC for Swiss bank notes resolves 67% of the variation. Let us test whether the true proportion could be 75%.

The 95% confidence interval for the true proportion is

$$0.668 \pm 1.96\sqrt{\frac{0.142}{199}} = (0.615, 0.720).$$

We reject the hypothesis that $\psi = 75\%$!

Application of PCA

The usual flow of PCA:

1. Is it necessary to standardize the data set?
2. How many PCs?
3. Interpretation!

Usual applications:

- dimension reduction,
- visualization (plotting) of high-dimensional datasets,
- regression on PCs (removes multicollinearity).

Týden 7–8

Faktorová analýza:

- model faktorové analýzy,
- odhadování a rotace faktorů,
- interpretace.

Factor analysis

Factor analysis has provoked rather turbulent controversy throughout its history.

. . . each application of the technique must be examined on its own merits to determine its success.

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors.

Factor analysis can be considered as an extension of principal component analysis . . . the approximation based on the factor analysis model is more elaborate.

(Johnson and Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 1992)

Factor Analysis Model

We want to explain p components of X by smaller number of common factors.

$$X = \mathcal{Q}F + U + \mu$$

$$\begin{aligned} \mathcal{Q} &= (p \times k) \text{ loadings} \\ F &= (k \times 1) \text{ common factors} \\ U &= (p \times 1) \text{ specific factors} \end{aligned}$$

where F and U are centered, $\mathrm{Var}(F) = \mathcal{I}_k$, $Var(U) = \Psi = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp})$, and $\mathrm{Cov}(F, U) = 0$.

Estimates of the loadings $\mathcal{Q}$ and specific variances Ψ are deduced from $\mathrm{var}\, X$ (using $\mathrm{var}\, X = \Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$).

Example: Perfect FA is PCA with only k positive eigenvalues:

$$\begin{aligned} \Sigma &= \sum_{\ell=1}^{k} \lambda_\ell \gamma_\ell \gamma_\ell^\top. \\ X &= \mathcal{Q}F + \mu \\ \mathcal{Q} &= \left(\sqrt{\lambda_1}\gamma_1, \ldots, \sqrt{\lambda_k}\gamma_k\right) \\ F &= k - \text{dim vector (random)} \\ EF &= 0 \\ \mathrm{Var}(F) &= \mathcal{I}_k \end{aligned}$$

Clearly, the matrix Q is not unique (because rotation leads to equivalent solution).

Communality and specific variance

Define

$$h_j^2 = \sum_{\ell=1}^{k} q_{j\ell}^2 \quad \text{communality}$$

$$\psi_{jj} \quad \text{specific variance}$$

Notice that $\operatorname{var} X_j = h_j^2 + \psi_{jj}$, i.e., the communality is the part of variance of X_j explained by the common factors. The specific variance is the unexplained part.

Two important properties of FA model are invariance of scale and non-uniqueness (with respect to rotations).

Invariance of scale

Assume that we have the following FA model for X: $\operatorname{var} X = \mathcal{Q}_X \mathcal{Q}_X^\top + \Psi_X$.

What happens if we change the scale of X?

$$\begin{aligned} Y &= \mathcal{C}X,\ \mathcal{C} = \operatorname{diag}(c_1, \ldots, c_p) \\ \operatorname{Var}(Y) &= \mathcal{C}\Sigma\mathcal{C}^\top \\ &= \mathcal{C}\mathcal{Q}_X\mathcal{Q}_X^\top\mathcal{C}^\top + \mathcal{C}\Psi_X\mathcal{C}^\top \end{aligned}$$

Hence the k-factor model is also true for Y with

$$\begin{aligned} \mathcal{Q}_Y &= \mathcal{C}\mathcal{Q}_X \\ \Psi_Y &= \mathcal{C}\Psi_X\mathcal{C}^\top. \end{aligned}$$

Non-Uniqueness of Factor Loadings

For orthogonal matrix $\mathcal{G}$ we get:

$$X = (\mathcal{Q}\mathcal{G})(\mathcal{G}^\top F) + U + \mu.$$

We get a k-factor model with factor loadings $\mathcal{Q}\mathcal{G}$ and common factors $\mathcal{G}^\top F$. In practical analysis, we will choose the rotation which gives "desirable" interpretation.

For the purpose of evaluation, the non-uniqueness can be solved by imposing additional constraints, e.g.,

$$\mathcal{Q}^\top\Psi^{-1}\mathcal{Q} \text{ is diagonal.}$$

Interpretation of the Factors

Interpretation of unobserved latent factors F is based on covariances and correlations:

$$\begin{aligned} \Sigma_{XF} &= E\{(\mathcal{Q}F + U)F^\top\} = \mathcal{Q} \\ \mathcal{P}_{XF} &= D^{-1/2}\mathcal{Q}, \end{aligned}$$

where $D = diag(\sigma_{X_1X_1}, \ldots . \sigma_{X_pX_p})$.

Correlations $\mathcal{P}_{XF}$ show the relationship between the original variables $X_1, \ldots, X_p$ and the common factors $F_1, \ldots, F_k$.

Number of parameters in the model

We have $p(p+1)/2$ equations and $pk+p$ parameters (pk parameters from $\mathcal{Q}$ and p parameters from Ψ) with $\frac{1}{2}\{(k(k-1)\}$ constraints (e.g. $\mathcal{Q}^\top \Psi^{-1}\mathcal{Q}$ is diagonal):

$$\begin{aligned} \Rightarrow \quad d &= \#\text{ pars for } \Sigma \text{ unconstrained} \\ &\quad -\#\text{ pars for } \Sigma \text{ constrained} \\ &= \tfrac{1}{2}(p-k)^2 - \tfrac{1}{2}(p+k). \end{aligned}$$

$$\begin{aligned} d &< 0 \text{ infinity of exact solutions} \\ d &> 0 \text{ look for approximate solutions} \end{aligned}$$

Example: $p=3,\ k=1 \Rightarrow d=0$

$$\Sigma = \begin{pmatrix} q_1^2+\psi_{11} & & \\ q_1q_2 & q_2^2+\psi_{22} & \\ q_1q_3 & q_2q_3 & q_3^2+\psi_{33} \end{pmatrix}$$

$d=0$ yields only a unique numerical solution! It need not be consistent with statistical thinking

The solution in the case $d=0$ might be numerically correct but inconsistent with statistical interpretation.

Example:

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix}$$

$[\Psi_{11} = -0.575!,\ q_{11} = 1.255]$.

Example: Suppose now $p=2$ and $k=1$, then $d<0$.

$$\Sigma = \begin{pmatrix} 1 & \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} q_1^2+\Psi_1 & \\ q_1q_2 & q_2^2+\Psi_2 \end{pmatrix}$$

We have an infinity of solutions: for any $\alpha(\rho<\alpha<1)$ a solution is provided by:

$$q_1=\alpha;\ q_2=\rho/\alpha;\ \Psi_1=1-\alpha^2;\ \Psi_2=1-(\rho/\alpha)^2$$

Factor Analysis Model

* The factor analysis model aims to describe the dependencies between the p variables in a data set by a lower number $k < p$ of latent factors, i.e. it assumes $X = \mathcal{Q}F + U + \mu$. The random vector F (k-dimensional) contains the common factors, U (p-dimensional) the specific factors, $\mathcal{Q}(p \times k)$ the loadings matrix.
* It is supposed that F and U are uncorrelated and have mean zero and uncorrelated components, i.e., $F \sim (0, \mathcal{I})$, $U \sim (0, \Psi)$ with a diagonal Ψ, $\mathrm{Cov}(F, U) = 0$.
  This leads to the covariance structure $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$.
* The interpretation of the factor F is obtained through the correlation $\mathcal{P}_{XF} = D^{-1/2}\mathcal{Q}$.

Factor Analysis Model

* A normalized analysis is obtained by the model $\mathcal{P} = \mathcal{Q}\mathcal{Q}^\top + \Psi$ The interpretation of the factors is given directly by the loadings $\mathcal{Q}$: $\mathcal{P}_{XF} = \mathcal{Q}$.
* The factor analysis model is scale invariant. The loadings are not unique (only up to multiplication by an orthogonal matrix).
* The non-uniqueness of the model is determined through the degrees of freedom $d = 1/2(p-k)^2 - 1/2(p+k)$

Estimation of the Factor Model

It is often easier to make the calculations for the standardized model (recall that FA is scale invariant).

Define:

$$\mathcal{Y} = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2}$$

$\uparrow$ centering matrix

Find a decomposition of the correlation matrix $\mathcal{P}$:

$$\mathcal{P} = \widehat{\mathcal{Q}}_Y\widehat{\mathcal{Q}}_Y^\top + \widehat{\Psi}_Y$$

$\widehat{\mathcal{Q}}_Y\widehat{\mathcal{Q}}_Y^\top$ common factors
$\widehat{\Psi}_Y$ specific factors

Example: Data set `carmean2` consists of the averaged marks (from 1 low to 7 high) for 31 car types.

We consider price, security and easy handling.

$$\mathcal{R} = \begin{pmatrix} 1 & 0.975 & 0.613 \\ & 1 & 0.620 \\ & & 1 \end{pmatrix}.$$

We look for one factor, i.e. $k = 1$. (# number of parameters of Σ unconstrained − # parameters of Σ constrained) equals here $\frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) = \frac{1}{2}(3-1)^2 - \frac{1}{2}(3+1) = 0$.

So there is an exact solution!

The equation

$$\begin{pmatrix} 1 & r_{X_1X_2} & r_{X_1X_2} \\ & 1 & r_{X_1X_3} \\ & & 1 \end{pmatrix} = \mathcal{R} = \begin{pmatrix} \widehat{q}_1^2 + \widehat{\psi}_{11} & \widehat{q}_1\widehat{q}_2 & \widehat{q}_1\widehat{q}_3 \\ & \widehat{q}_2^2 + \widehat{\psi}_{22} & \widehat{q}_2\widehat{q}_3 \\ & & \widehat{q}_3^2 + \widehat{\psi}_{33} \end{pmatrix}$$

yields the communalities $\widehat{h}_i^2 = \widehat{q}_i^2$

$$\widehat{q}_1^2 = \frac{r_{X_1X_2} r_{X_1X_3}}{r_{X_2X_3}} \qquad \widehat{q}_2^2 = \frac{r_{X_1X_2} r_{X_2X_3}}{r_{X_1X_3}} \qquad \widehat{q}_3^2 = \frac{r_{X_1X_3} r_{X_2X_3}}{r_{X_1X_2}} .$$

Together with $\widehat{\psi}_{11} = 1 - \widehat{q}_1^2$, $\widehat{\psi}_{22} = 1 - \widehat{q}_2^2$ and $\widehat{\psi}_{33} = 1 - \widehat{q}_3^2$ we get the solution

$$\begin{array}{lllllllll} \widehat{q}_1 & = & 0.982 & & \widehat{q}_2 & = & 0.993 & & \widehat{q}_3 & = & 0.624 \\ \widehat{\psi}_{11} & = & 0.035 & & \widehat{\psi}_{22} & = & 0.014 & & \widehat{\psi}_{33} & = & 0.610. \end{array}$$

Since the first two communalities are close to one, we conclude that the first two variables, namely price and security, are explained by the factor very well.

This factor might be interpreted as a "price+security" factor.

The Principal Component Method

Decompose $\text{var}(X) = \mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^\top$.

Retain the first k eigenvectors to build

$$\hat{\mathcal{Q}} = [\sqrt{\ell_1} g_1, \ldots, \sqrt{\ell_k} g_k].$$

Omitting $p - k$ eigenvectors shouldn't cause big error if the corresponding eigenvalues λ_i, $i = k+1, \ldots, p$ are small.

Specific variance are estimated by diagonal elements of

$$\mathcal{S} - \hat{\mathcal{Q}}\hat{\mathcal{Q}}^\top.$$

This gives $\hat{\Psi}$.

Error of approximation

Residual matrix $\mathcal{S} - (\hat{\mathcal{Q}}\hat{\mathcal{Q}}^\top + \hat{\Psi})$

[diag is 0 but off-diag not]

Analytically:

$$\sum_{i,j} \left[\mathcal{S} - (\hat{\mathcal{Q}}\hat{\mathcal{Q}}^\top + \hat{\Psi})\right]_{i,j}^2 \leq \hat{\ell}_{k+1}^2 + \ldots + \hat{\ell}_p^2$$

gives an estimate of error of the approximation (using Frobenius norm).

This gives simple criterion for the choice of number of the factors.

Method of Principal Factors

We start with an estimate of the communality:

1) $\widetilde{h}_j^2$ = the square of the multiple correlation coefficient, i.e. $\rho^2(V, W\widehat{\beta})$ with $V = X_j$
   W = $(X_\ell)_{\ell \neq j}$
   $\widehat{\beta}$ = OLS of regression of V on W
2) $\widetilde{h}_j^2$ = $\max_{\ell \neq j} |r_{X_j X_\ell}|$
   $\mathcal{R}$ = correlation matrix.

Algorithm of Principal Factors Method

$$
\begin{aligned}
\tilde{\psi}_{jj} &= 1 - \tilde{h}_j^2 \\
\text{Construct} & \quad \mathcal{R} - \widetilde{\Psi} \\
\mathcal{R} - \widetilde{\Psi} &= \sum_{\ell=1}^{p} \lambda_\ell \gamma_\ell \gamma_\ell^\top \\
\widehat{q}_\ell &= \sqrt{\lambda_\ell}\, \gamma_\ell \,, \quad \ell = 1, \ldots k \\
\widehat{\mathcal{Q}} &= \Gamma_1 \Lambda_1^{1/2} \\
\Gamma_1 &= (\gamma_1, \ldots, \gamma_k) \\
\Lambda_1 &= \mathrm{diag}(\lambda_1, \ldots, \lambda_k) \\
\widehat{\psi_{jj}} &= 1 - \sum_{\ell=1}^{k} \widehat{q}_{j\ell}^2
\end{aligned}
$$

The Maximum Likelihood Method

Log-likelihood function ℓ for a data matrix $\mathcal{X}$ of observations for $X \sim N_p(\mu, \Sigma)$:

$$
\begin{aligned}
\ell(\mathcal{X}; \mu, \Sigma) &= -\frac{n}{2} \log \mid 2\pi\Sigma \mid -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)\Sigma^{-1}(x_i - \mu)^\top \\
&= -\frac{n}{2} \log \mid 2\pi\Sigma \mid -\frac{n}{2} \mathrm{tr}(\Sigma^{-1}\mathcal{S}) - \frac{n}{2}(\overline{x} - \mu)\Sigma^{-1}(\overline{x} - \mu)^\top .
\end{aligned}
$$

Evaluated at its maximum $\widehat{\mu} = \overline{x}$:

$$
\ell(\mathcal{X}; \widehat{\mu}, \Sigma) = -\frac{n}{2} \left\{ \log(\mid 2\pi\Sigma \mid) - \mathrm{tr}(\Sigma^{-1}\mathcal{S}) \right\} .
$$

By substituting $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$

$$
\ell(\mathcal{X}; \widehat{\mu}, \mathcal{Q}, \Psi) = -\frac{n}{2} \left[\log\{\mid 2\pi(\mathcal{Q}\mathcal{Q}^\top + \Psi) \mid\} - \mathrm{tr}\{(\mathcal{Q}\mathcal{Q}^\top + \Psi)^{-1}\mathcal{S}\} \right] .
$$

This model is not well defined.

Therefore, we require that $\mathcal{Q}^\top \Psi^{-1} \mathcal{Q}$ is diagonal matrix.

The maximum likelihood estimates of $\mathcal{Q}$ and Ψ are obtained using an iterative numerical algorithm (function `factanal()` in R library `MASS`).

LR test for the Number of Common Factors

The test follows directly from the assumption of normality. We test
$H_0 : \Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$
$H_1 : \Sigma$ arbitrary (positive definite) matrix

See the chapter on Likelihood Ratio tests.

The likelihood ratio statistic is

$$-2\Lambda = -2\log\left[\frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood}}\right]$$

$$= -2\log\left(\frac{|\hat{\mathcal{Q}}\hat{\mathcal{Q}}^\top + \hat{\Psi}|}{|S_n|}\right)^{-n/2} + n\{tr[(\hat{\mathcal{Q}}\hat{\mathcal{Q}}^\top + \hat{\Psi})^{-1}S_n] - p\}$$

with $(1/2)[(p-m)^2 - p - m]$ degrees of freedom.

Rotation

The factor analysis model is not uniquely defined and the factors can be rotated without any loss of information.

We are free to rotate the estimated factors rather arbitrary. This feature of factor analysis is rather controversial.

Usually, we rotate the factors in a way which provides reasonable interpretation which is consistent with the measured variables.

In the most simple case of $k = 2$ factors a rotation matrix $\mathcal{G}$ is given by

$$\mathcal{G}(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

which represents a clockwise rotation of the coordinate axes by the angle θ (then $\hat{\mathcal{Q}}^* = \hat{\mathcal{Q}}\mathcal{G}(\theta)$).

Varimax

The varimax method tries to find "reasonable rotation" automatically.

The interpretation of the loadings would be simple if the variables split into disjoint sets, each of which is associated with one factor. A well known analytical algorithm which tries to rotate the loadings in this way is the *varimax rotation method*.

Varimax method tries to find the rotation which maximizes the sum of the variances of the squared loadings $\hat{q}^*_{ij}$ within each column of $\hat{\mathcal{Q}}^*$ (this should lead to q_{ij}s close to 0 or 1):

$$\max_{\text{rotations } Q^*} \sum_{j=1}^{k}\left\{\frac{1}{p}\sum_i (q^*_{ij})^4 - \left[\frac{1}{p}\sum_i (q^*_{ij})^2\right]^2\right\}.$$

Promax

The promax rotation is similar to varimax but it works without the condition of orthogonality (so-called oblique rotation).

The resulting correlated(!) factors are typically easier to interpret.

Strategy for Factor Analysis

1. Perform a principal component factor analysis, look for suspicious observations, try varimax rotation
2. Perform maximum likelihood factor analysis including varimax rotation
3. Compare the factor analyses: do the loadings group in the same manner?
4. Repeat the previous steps for other number of common factors
5. For large data sets, split them in half and perform a factor analysis on each part. Compare the solutions.

Estimation of the Factor Model

- In practice $\mathcal{Q}$ and Ψ have to be estimated from $\mathcal{S} = \widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top + \widehat{\Psi}$. The number of free parameters is $d = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$.
- The maximum-likelihood method supposes a normal distribution for the data, a solution can be found by numerical algorithms.
- The method of principal factors is a two-stage method which calculates $\widehat{\mathcal{Q}}$ from the reduced correlation matrix $\mathcal{R} - \widetilde{\Psi}$, $\widetilde{\Psi}$ a pre-estimate for Ψ. The final estimate for Ψ is found by $\widehat{\psi_{ii}} = 1 - \sum_{j=1}^{k} \widehat{q_{ij}^2}$.
- Principal components can be interpreted as a simple factor analysis model with loadings $\mathcal{Q} = \Gamma_k \Lambda_k^{1/2}$.
- A better interpretation can be found by rotating the loadings $\mathcal{Q}$.

Factor scores

Factor scores are estimates of the unobserved random vectors F_l, $l = 1, \ldots, k$, for each individual x_i, $i = 1, \ldots, n$.

Factor scores may be useful for interpretation as well as in the diagnostic analysis

The idea of the *regression method* (or Thomson method) is to consider the joint distribution of $(X - \mu)$ and F (assuming multivariate normality) and then derive the conditional distribution $F|X$.

The joint covariance matrix of $(X - \mu)$ and F is:

$$\text{var}\begin{pmatrix} X-\mu \\ F \end{pmatrix} = \begin{pmatrix} \mathcal{Q}\mathcal{Q}^\top + \Psi & \mathcal{Q} \\ \mathcal{Q}^\top & \mathcal{I}_k \end{pmatrix}.$$

Note that the upper left entry of this matrix equals Σ and that the matrix has size $(p+k) \times (p+k)$.

Assuming joint normality, the conditional distribution of $F|X$ is multinormal with $E(F|X = x) = \mathcal{Q}^\top \Sigma^{-1}(X - \mu)$ and the covariance matrix $\text{var}(F|X = x) = \mathcal{I}_k - \mathcal{Q}^\top \Sigma^{-1} \mathcal{Q}$.

In practice, we replace the unknown $\mathcal{Q}$, Σ and μ by corresponding estimators, leading to the estimated individual factor scores:

$$\widehat{f}_i = \widehat{\mathcal{Q}}^\top \mathcal{S}^{-1}(x_i - \overline{x}).$$

We prefer to use the original sample covariance matrix $\mathcal{S}$ as an estimator of Σ, instead of the factor analysis approximation $\widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top + \widehat{\Psi}$, in order to be more robust against incorrect determination of the number of factors.

Notes

1. The same rule can be followed when using $\mathcal{R}$ instead of $\mathcal{S}$. In this case the factors are given by

$$\widehat{f}_i = \widehat{\mathcal{Q}}^\top \mathcal{R}^{-1}(z_i),$$

where $z_i = \mathcal{D}_S^{-1/2}(x_i - \bar{x})$, $\widehat{\mathcal{Q}}$ is the loading obtained with the matrix $\mathcal{R}$, and $\mathcal{D}_S = \text{diag}(s_{11}, \ldots, s_{pp})$.

2. Using MLE (treating F as unknown parameters), one arrives to Bartlett's scores.
3. Clearly, if the factors are rotated by the orthogonal matrix $\mathcal{G}$, the factor scores have to be rotated accordingly, that is

$$\widehat{f}_i^* = \mathcal{G}^\top \widehat{f}_i.$$

Concluding remarks

In practice, this technique is also called Exploratory Factor Analysis.

After exploring the factors, one can perform the so-called Confirmatory Factor Analysis allowing more detailed investigation of the underlying factors (one can imagine that oblique factors could be explained by another factor analysis leading to a hiearchical model).

Relationship between the unobserved factors can be investigated using Structural Equation Models (R library `sem`, M-plus, LISREL).

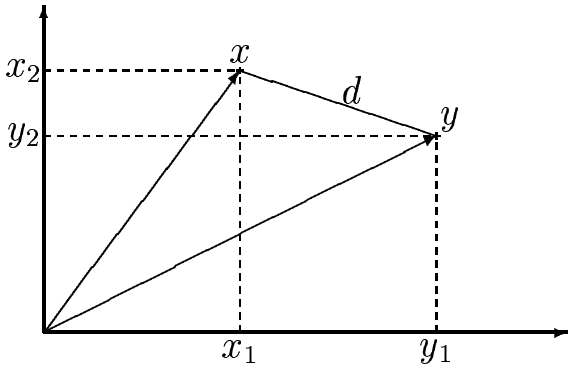
Factor analysis models are popular mainly in psychology and behavioral science.

Týden 9

Mnohorozměrné škálování:

- matice vzdáleností,
- metrické řešení,
- nemetrické řešení (PAVA a STRESS).

Euclidean distance



$$d(x, y) = \sqrt{(x-y)^\top (x-y)}$$

Distance matrix

$\mathcal{X}(n \times p)$ with n measurements (objects) of p variables.

The distance matrix $\mathcal{D}(n \times n)$ is a matrix of all distances between all pairs of observations:

$$\mathcal{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & \dots & \dots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & \dots & d_{nn} \end{pmatrix}$$

Example: L_2-norm: $d_{ij} = \|x_i - x_j\|_2$, where x_i and x_j denote the rows of the data matrix $\mathcal{X}$.

Distance and similarity

Distance can be easily calculated for *numerical measurements* (Euclidean distance, L_1 distance, Mahalanobis distance, etc.)

Example:

```
data(bank2)
dist(bank2,method="euclidean")
dist(scale(bank2),method="euclidean")
```

For *nominal (or binary) variables* it is easier to define a measure of similarity (e.g., various ratios of "number of concordances" such as Jaccard, Tanimoto, Simple Matching, etc.)

Example: Let us consider data set on songs in 20 medieval songbooks.

	spev	P	U	K	E	G	Kl	SMF	Pa	S	M	W	Lc	B	F	SG	To	C	Me	Q	R
1	HOMO QUIDAM	x	x	x	x	x	x	x	_	_	x	x	x	_	_	x	x	x	x	x	x
2	FACTUM EST	x	x	x	x	x	x	x	x	x	x	_	_	_	_	x	x	x	x	x	x
3	ELEVANS AUTEM	x	x	x	x	x	x	_	_	_	_	_	_	_	_	_	_	_	x	x	x
4	ROGO ERGO	x	x	x	x	x	x	x	x	x	_	x	x	_	_	x	x	_	x	x	x
5	DIVES ILLE	x	x	x	x	x	x	x	_	_	_	_	_	_	_	x	x	x	x	_	_
6	DEUS CARITAS	x	x	_	_	_	x	_	_	_	_	_	_	_	_	_	x	_	x	x	_
7	HOMO QUIDAM	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	_
8	EXI CITO	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	_
9	DOMINE FACTUM	x	x	x	x	x	x	x	_	_	_	_	_	_	_	x	_	_	x	x	x
10	DICO AUTEM	x	x	x	x	x	x	x	_	_	x	_	_	_	_	_	_	x	x	x	x
11	QUIS EX	x	x	x	x	x	x	x	x	x	x	_	x	x	x	x	x	x	x	x	x
12	CONGRATULAMINI	x	x	x	x	x	x	_	_	_	x	x	_	_	_	x	x	x	x	x	x

Considering observations x_i and x_j and denoting

$$\begin{aligned} a_1 &= \sum_{k=1}^{p} I(x_{ik} = x_{jk} = "x"), \\ a_2 &= \sum_{k=1}^{p} I(x_{ik} = "\_", x_{jk} = "x"), \\ a_3 &= \sum_{k=1}^{p} I(x_{ik} = "x", x_{jk} = "\_"), \\ a_4 &= \sum_{k=1}^{p} I(x_{ik} = x_{jk} = "\_"). \end{aligned}$$

we can define a proximity measure as

$$s_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}$$

by choosing some δ and λ.

Name	δ	λ	Definition
Jaccard	0	1	$\frac{a_1}{a_1+a_2+a_3}$
Tanimoto	1	2	$\frac{a_1+a_4}{a_1+2(a_2+a_3)+a_4}$
Simple Matching (M)	1	1	$\frac{a_1+a_4}{p}$
Dice	0	0.5	$\frac{2a_1}{2a_1+(a_2+a_3)}$

In the songbooks example, the Jaccard measure seems to be reasonable. Calculating the similarity measure for all pairs of songbooks, we obtain a similarity matrix $S = s_{ij}$.

Distance and similarity

Distances and similarities are closely related. It may be useful to "transform" similarity to distance because some methods require distances.

Denoting similarities as s_{ij}, distances can be defined as $d_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$ or $d_{ij} = \max_{i,j}\{s_{ij}\} - s_{ij}$ (we want symmetry and $d_{ii} = 0$).

In the songbook example, we can define

$$d_{ij} = \frac{a_2 + a_3}{a_1 + a_2 + a_3} = 1 - s_{ij}$$

as a ratio of "common songs" (from songs that are contained in songbooks i and j).

The Proximity between Objects

- The proximity between data points is measured by a distance or similarity matrix $\mathcal{D}$ whose components d_{ij} give the similarity coefficient or the distance between two points x_i, x_j.
- There exists a variety of similarity (distance) measures for binary data (e.g., Jaccard, Tanimoto, Simple Matching coefficients) and for continuous data (e.g., L_r-distances).
- The nature of the data could impose to choose a particular metric $\mathcal{A}$ for defining the distance (standardization, χ^2-metric etc.).

Euclidean matrix

It is easy to calculate distance matrix $\mathcal{D}$ from the data matrix $\mathcal{X}$ but is it possible to calculate the data matrix $\mathcal{X}$ from a distance matrix $\mathcal{D}$?

Definition: We say that $\mathcal{D} = (d_{ij})$ is a distance matrix if $d_{ij} = d_{ji} \geq 0$ and $d_{ii} = 0$, for $i, j = 1, \ldots, n$.

The first step would be to verify that the matrix $\mathcal{D}$ is Euclidean (i.e., that it contains Euclidean distances).

Definition: We say that a matrix $\mathcal{D} = (d_{ij})$ is Euclidean if for some points $x_1, \ldots, x_n \in \mathbb{R}^p$; $d_{ij}^2 = (x_i - x_j)^\top (x_i - x_j)$.

Theorem: Define $\mathcal{A} = (a_{ij}), a_{ij} = -\frac{1}{2}d_{ij}^2$, $\mathcal{B} = \mathcal{H}\mathcal{A}\mathcal{H}$, $\mathcal{H}$ being the centering matrix. Then the matrix $\mathcal{D}$ is Euclidean if and only if $\mathcal{B}$ is positive semidefinite.

Idea of the proof:

1/ Assuming that $\mathcal{D}$ is Euclidean for centered data matrix $\mathcal{X}$, we have $d_{ij}^2 = x_i^\top x_i + x_j^\top x_j - 2x_i^\top x_j$.

Writing $\mathcal{B} = \mathcal{H}\mathcal{A}\mathcal{H}$ implies that $b_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..} = \cdots = x_i^\top x_j$. Therefore $\mathcal{B} = \mathcal{X}\mathcal{X}^\top \geq 0$.

2/ Assuming that $\mathcal{B} \geq 0$ and $\text{rank}(B) = p$, we can write $\mathcal{B} = \Gamma_p \Lambda_p \Gamma_p^\top$ and it follows (similarly as above) that $\mathcal{D}$ is matrix of Euclidean distances of points in $\mathcal{X} = \Gamma_p \Lambda_p^{1/2}$.

Note that the matrix $\mathcal{X} = \Gamma_p \Lambda_p^{1/2}$ is centered.

Multidimensional scaling

MDS uses proximities (distances) between objects to produce a spatial representation of these items.

In contrast to the techniques considered so far MDS does not start from the raw multivariate data matrix, $\mathcal{X}$, but from a $(n \times n)$ dissimilarity or a distance matrix $\mathcal{D}$. Hence, the underlying dimensionality of the data under investigation is not known.

More precisely: MDS searches for a "configuration" of points in $\mathbb{R}^k$ that "preserves" the distances of objects in $\mathbb{R}^p$ (where p is not known).

MDS-techniques can help to understand how people perceive and evaluate certain items:

metric MDS is based on Euclidean distances,

non-metric MDS assumes that distances are on ordinal scale.

MDS solution (metric MDS)

Recall that $\mathcal{B} = \mathcal{H}\mathcal{A}\mathcal{H}$, where $a_{ij} = -\frac{1}{2}d_{ij}^2$.

Assuming that $\text{rank}(B) = p$ and writing $\mathcal{B} = \Gamma_p \Lambda_p \Gamma_p^\top$, we obtain data matrix $\mathcal{X} = \Gamma_p \Lambda_p^{1/2}$ that preserves the observed distances in p-dimensional space.y

If some of the eigenvalues are small, we can obtain a good representation (of the distances) in k-dimensional space by $\mathcal{X} = \Gamma_k \Lambda_k^{1/2}$.

Note that the final configuration of points in $\mathbb{R}^k$ can be arbitrarily rotated and shifted without changing the distances.

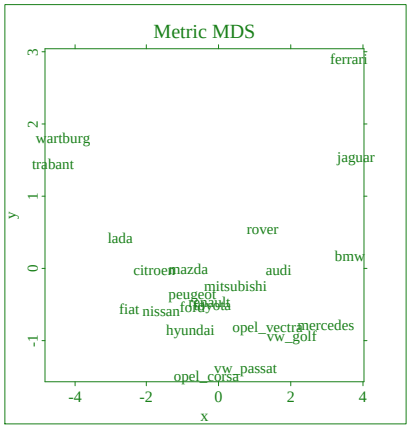
Example: Consumers' impressions of the dissimilarity of certain cars.

	Audi 100	BMW 5	Citroen AX	Ferrari	...
Audi 100	0	2.232	3.451	3.689	...
BMW 5	2.232	0	5.513	3.167	...
Citroen AX	3.451	5.513	0	6.202	...
Ferrari	3.689	3.167	6.202	0	...
⋮	⋮	⋮	⋮	⋮	⋱

```
library(SMDdata);library(MASS);data(carmean2)

X=cmdscale(dist(carmean2))

plot(X,type="n")
text(X,labels=row.names(carmean2))
```



MDS solution.

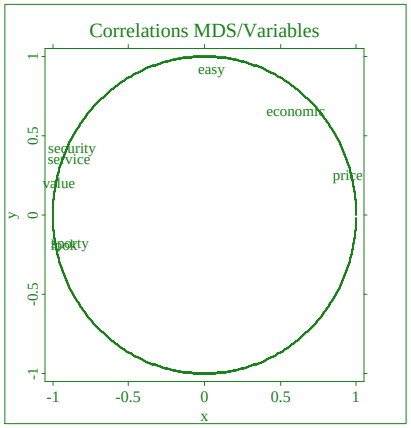
The dissimilarities were in fact computed as Euclidean distances from the original data containing car marks data on economy, price, security, ...

Therefore, we can plot the correlation between the MDS projection and the original variables (see next slide).

The first MDS direction is highly correlated with service(-), value(-), design(-), sportiness(-), safety(-) and price(+). We can interpret the first direction as the price direction since a bad mark in price ("high price") obviously corresponds with a good mark, say, in sportiness ("very sportive").

The second MDS direction is highly positively correlated with practicability.

We see that we have a non-linear relationship between price and practicability.



Correlations between the MDS direction and the original variables.

Relation to principal components

Let $\mathcal{X}$ be a data matrix. Assume (WLOG) that $\mathcal{X}$ is centered.

Notice that $n\mathcal{S} = \mathcal{X}^\top\mathcal{X}$ and $\mathcal{B} = \mathcal{X}\mathcal{X}^\top$ have the same non-zero eigenvalues.

The SVD decomposition $\mathcal{X} = \mathcal{U}\mathcal{L}\mathcal{V}^\top$ implies that the spectral decomposition of $\mathcal{X}^\top\mathcal{X}$ is $\mathcal{V}\mathcal{L}^2\mathcal{V}^\top$. Therefore, the principal compoments $\mathcal{X}\mathcal{V} = \mathcal{U}\mathcal{L}\mathcal{V}^\top\mathcal{V} = \mathcal{U}\mathcal{L} = \Gamma\Lambda^{1/2}$ (where $\mathcal{X}\mathcal{X}^\top = \Gamma\Lambda\Gamma^\top (= \mathcal{U}\mathcal{L}^2\mathcal{U}^\top)$).

Therefore, the metric MDS solution recovers the first k principal components (of the original data set).

Multidimensional Scaling

- MDS uses distances between n items to project high-dimensional data in a low-dimensional space.
- MDS (using Euclidean distances in p dimensions) leads to first k principal components of the original data set.
- It can be shown that the metric solution to MDS leads to optimal representation of the original data set in k dimensional space (from the point of view of $\sum(d_{ij}^2 - \hat{d}_{ij}^2)$, where d_{ij} are the original distances and $\hat{d}_{ij}$ are the projections in $\mathbb{R}^k$—note also that $\hat{d}_{ij} \leq d_{ij}$).

Nonmetric MDS

Nonmetric MDS is based on a "loose" relationship between dissimilarities and distances.

The distance is defined as an arbitrary monotone function of the dissimilarities (i.e., nonmetric MDS is based on the rank order of the dissimilarities).

The most common approach is to determine (some) (non-Euclidean) distances and then obtain the coordinates of the objects by using the iterative Shepard-Kruskal algorithm.

Shepard-Kruskal algorithm

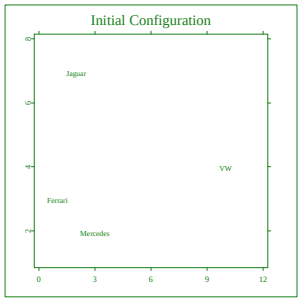
1. calculate Euclidean distances from arbitrarily chosen inital configuration $\mathcal{X}$ (or use metric MDS to obtain the initial coordinates),
2. define new distances so that they are monotone function of the original dissimilarities (using monotone regression),
3. calculate new configuration of the data which is more closely related to the distances obtained in step 2 (minimize STRESS, numerical approximation on a computer is needed),
4. check the change of STRESS, if it isn't small enough, iterate the algorithm.

We demonstrate each step of the algorithm using a simple example.

Example: Consider a small example with 4 objects based on the car marks data set.

	j	1	2	3	4
i		Mercedes	Jaguar	Ferrari	VW
1	Mercedes	-			
2	Jaguar	3	-		
3	Ferrari	2	1	-	
4	VW	5	4	6	-

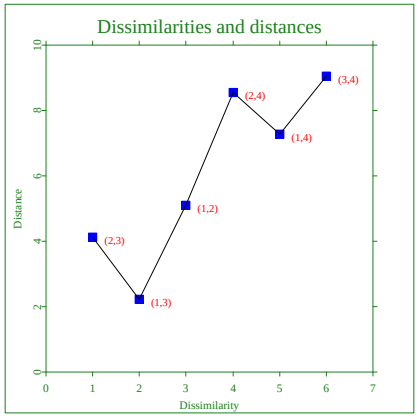
Our aim is to find a $p^* = 2$ dimensional representation via MDS. Suppose that we choose as initial configuration $\mathcal{X}_0$ the coordinates as:



i		x_{i1}	x_{i2}
1	Mercedes	3	2
2	Jaguar	2	7
3	Ferrari	1	3
4	VW	10	4

The corresponding distances $d_{ij} = \sqrt{(x_i - x_j)^\top (x_i - x_j)}$ are

i, j	d_{ij}	$rank(d_{ij})$	δ_{ij}
1,2	5.1	3	3
1,3	2.2	1	2
1,4	7.3	4	5
2,3	4.1	2	1
2,4	8.5	5	4
3,4	9.1	6	6



A plot of the dissimilarities is not satisfactory since the ranking of the δ_{ij} did not result in a monotone relation of the corresponding distances d_{ij}. We apply therefore the PAV algorithm.

PAVA = "pool adjacent violators" algorithm (= algoritmus "zprůměrování sousedních narušitelů") is used to calculate the LS estimator under assumption of monotonicity.

The first violator of monotonicity is the second point $(1, 3)$ we therefore average the distances d_{13} and d_{23} to obtain the disparities

$$\hat{d}_{13} = \hat{d}_{23} = \frac{d_{13} + d_{23}}{2} = \frac{2.2 + 4.1}{2} = 3.17.$$

We apply the same procedure to the pair $(2, 4)$ and $(1, 4)$ to yield $\hat{d}_{24} = \hat{d}_{14} = 7.9$. The plot od δ_{ij} versus the disparities $\hat{d}_{ij}$ represents a monotone regression relationship.

In the initial configuration, the point 3 (Ferrari) could be moved so that the distance to object 2 (Jaguar) is smaller. This procedure however also alters the distance between objects 3 and 4. More care has therefore to be taken for an establishment of a monotone relation between δ_{ij} and d_{ij}.

STRESS

In order to assess how well the derived configuration fits the given dissimilarities Kruskal suggests a measure called STRESS1 that is given by

$$STRESS1 = \left(\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right)^{\frac{1}{2}}.$$

An alternative measure of STRESS1 is given by

$$STRESS2 = \left(\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}(d_{ij} - \overline{d})^2} \right)^{\frac{1}{2}},$$

where $\overline{d}$ denotes the average distance.

STRESS calculations for the car example:

(i,j)	δ_{ij}	d_{ij}	$\hat{d}_{ij}$	$(d_{ij} - \hat{d}_{ij})^2$	d_{ij}^2	$(d_{ij} - \overline{d})^2$
(2,3)	1	4.1	3.15	0.9	16.8	3.8
(1,3)	2	2.2	3.15	0.9	4.8	14.8
(1,2)	3	5.1	5.1	0	26.0	0.9
(2,4)	4	8.5	7.9	0.4	72.3	6.0
(1,4)	5	7.3	7.9	0.4	53.3	1.6
(3,4)	6	9.1	9.1	0	82.8	9.3
Σ		36.3		2.6	256.0	36.4

The average distance is $\overline{d} = 36.4/6 = 6.1$. The corresponding STRESS measures are: $STRESS1 = \sqrt{2.6/256} = 0.1$,
$STRESS2 = \sqrt{2.6/36.4} = -.27$

The aim is a point configuration that balances the effects STRESS and non monotonicity. This is achieved by an iterative procedure defining new position of object i relative to object j by

$$x_{il}^{NEW} = x_{il} + \alpha \left(1 - \frac{\hat{d}_{ij}}{d_{ij}}\right)(x_{jl} - x_{il}), \qquad l = 1, \ldots, p^*.$$

Here α denotes the step width of the iteration.

The configuration of object i is improved relative to object j. In order to obtain an overall improvement relative to all remaining points one uses:

$$x_{il}^{NEW} = x_{il} + \frac{\alpha}{n-1} \sum_{j=1, j\neq i}^{n} \left(1 - \frac{\hat{d}_{ij}}{d_{ij}}\right)(x_{jl} - x_{il}), \qquad l = 1, \ldots, p^*.$$

The choice of step width α is crucial. Kruskal proposes a starting value of $\alpha = 0.2$. The iteration is continued by a numerical approximation procedure.

In a fourth step, the evaluation phase, the STRESS measure is used to evaluate if its change as a result of the last iteration is sufficiently small to terminate the procedure, or not. At this stage the optimal fit has been obtained for a given dimension. Hence, the whole procedure needs to be carried out for a several dimensions.

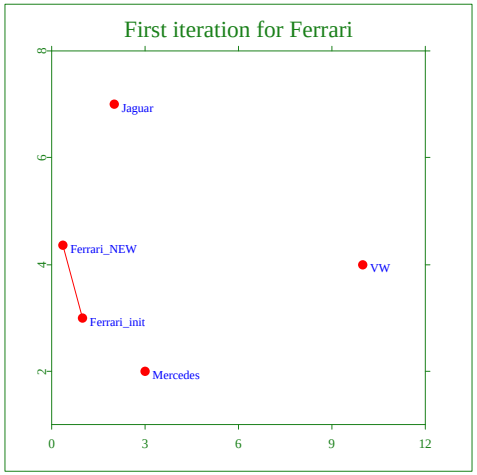
Let us compute the new point configuration for $i = 3$ (Ferrari). The initial coordinates are $x_{31} = 1$, $x_{32} = 3$. Applying the above formula yields:

$$x_{31}^{NEW} = 1 + \frac{3}{4-1} \sum_{j=1, j\neq 3}^{4} \left(1 - \frac{\hat{d}_{31}}{d_{31}}\right)(x_{j1} - 1)$$

$$= 1 + \left(1 - \frac{3.15}{2.2}\right)(3-1) + \left(1 - \frac{3.15}{2.2}\right)(2-1) + \left(1 - \frac{9.1}{9.1}\right)(10-1)$$

$$= 1 - 0.86 + 0.23 + 0 = 0.37$$

Similarly we obtain $x_{32}^{NEW} = 4.36$.



First iteration for Ferrari.

Similarity of cars in R:

```
d=matrix(c(0,3,2,5,3,0,1,4,2,1,0,6,5,4,6,0),nrow=4)
row.names(d)=c("Mercedes","Jaguar","Ferrari","VW")
colnames(d)=row.names(d)
### the initial configuration
init=matrix(c(3,2,1,10,2,7,3,4),ncol=2)
par(mfrow=c(1,2))
plot(init,xlim=1.1*range(init),ylim=1.1*range(init),
     xlab="",ylab="")
text(init,labels=row.names(d))
### the non-metric solution
mds=isoMDS(d,trace=TRUE,y=init)
plot(NULL,xlim=1.1*range(mds$points),
     ylim=1.1*range(mds$points),xlab="",ylab="")
text(mds$points,labels=row.names(d))
```

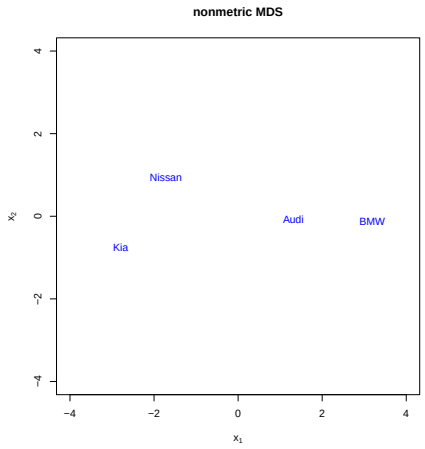

Example: Inter-city distances in Czech Republic.

```
d=matrix(c(0,3,1,5,  3,0,4,2,  1,4,0,6,  5,2,6,0),nrow=4)
row.names(d)=c("Praha","Brno","Plzen","Ostrava")
colnames(d)=row.names(d)
### the metric solution
body=cmdscale(d)
plot(body,xlim=1.1*range(body),ylim=1.1*range(body),
     xlab="",ylab="")
text(body,labels=c("Praha","Brno","Plzen","Ostrava"))
### the non-metric solution
mds=isoMDS(d,trace=TRUE)
plot(NULL,xlim=1.1*range(mds$points),
     ylim=1.1*range(mds$points),xlab="",ylab="")
text(mds$points,labels=c("Praha","Brno","Plzen","Ostrava"))
```


Example: Dissimilarity matrix for car marks data set:

	j	1	2	3	4
i		Nissan	Kia	BMW	Audi
1	Nissan	-			
2	Kia	2	-		
3	BMW	5	6	-	
4	Audi	3	4	1	-

The dissimilarity matrix contains obviously only ranks of dissimilarity. Applying metric MDS may not be appropriate in this situation.



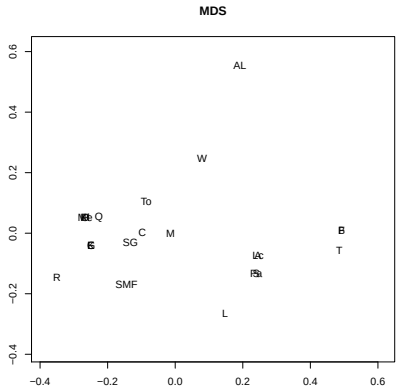
The outcome of the Shepard-Kruskal algorithm. It is important that both axes have the same scale, different scales could lead to wrong interpretations.

The Euclidean distances between the points are:

i	j	1 Nissan	2 Kia	3 BMW	4 Audi
1	Nissan	-			
2	Kia	2.00	-		
3	BMW	5.02	6.02	-	
4	Audi	3.20	4.16	1.88	-

These distances are different from the original dissimilarities but their order is the same, i.e., the STRESS measure is equal to 0.

Example: The nonmetric MDS solution for the songbooks example (→ `SMSclussong`):



Nonmetric Multidimensional Scaling

- Nonmetric MDS is based only upon the rank order of dissimilarities.
- The object of nonmetric MDS is to create a spatial representation of the objects with low dimensionality.
- A practical algorithm is given as:
  1. Choose an initial configuration
  2. Normalize the configuration.
  3. Find d_{ij} from the normalized configuration
  4. Fit $\hat{d}_{ij}$, the disparities by the PAV algorithm
  5. Find the new configuration $\mathcal{X}_{n+1}$ by using steepest descent.
  6. Go to 2 and interate until STRESS is small enough.

Týden 10

Shluková analýza:

- shlukovací algoritmy,
- hierarchické aglomerativní algoritmy,
- dendrogram.

Cluster analysis

Cluster analysis is a set of tools and methods for building groups (clusters) from multivariate data objects. The aim is to find groups with homogeneous properties out of heterogeneous large samples.

The algorithm is usually divided into two fundamental steps:

1. the choice of a proximity measure,
2. the choice of a group-building algorithm.

We have already discussed several distance and proximity measures. The choice of proximity measure typically follows from the type of measurements in the data set.

In the following, we assume that we have $(n \times n)$ distance matrix $\mathcal{D}$ (calculated from p-dimensional data set $\mathcal{X}$).

Group-building algorithms

Two types of clustering methods:

- partioning algorithms (typically computationally intensive optimization of a given criterion),
- hierarchical algorithms:
  - agglomerative,
  - partioning.

In partitioning techniques the assignment of objects into groups may change during the (iterative) algorithm.

In hierarchical clustering this assignment cannot be changed (the algorithm produces a sequence of clusters by "splitting" or "joining").

In the following, we look at agglomerative techniques.

Agglomerative algorithms

The agglomerative algorithm consists of the following steps:

1. Construct the finest partition.
2. Compute the distance matrix $\mathcal{D}$.

DO

3. Find the clusters with the closest distance.
4. Put those two clusters into one cluster.
5. Compute the distance between the new groups and obtain a reduced distance matrix $\mathcal{D}$.

UNTIL all clusters are agglomerated into $\mathcal{X}$.

Agglomerative techniques are computationaly simple because the distances between clusters can be easily calculated from the distance matrix $\mathcal{D}$.

If two objects or groups P and Q are to be united one obtains the distance to another group (object) R by the following distance function

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|$$

δ_j weighting factors

	δ_1	δ_2	δ_3	δ_4
Single linkage	$1/2$	$1/2$	0	-1/2
Complete linkage	$1/2$	$1/2$	0	$1/2$
Average linkage (unweighted)	$1/2$	$1/2$	0	0
Average linkage (weighted)	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroid	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Median	$1/2$	$1/2$	-1/4	0
Ward	$\frac{n_R+n_P}{n_.}$	$\frac{n_R+n_Q}{n_.}$	$-\frac{n_R}{n_.}$	0

$n_P = \sum_{i=1}^{n} I(x_i \in P)$ denotes the number of objects in group P

$n_. = n_R + n_P + n_Q$

Example: $x_1 = (0,0), x_2 = (1,0), x_3 = (5,5)$ and the squared Euclidean distance matrix with single linkage weighting.

The algorithm starts with $N = 3$ clusters $P = \{x_1\}, Q = \{x_2\}, R = \{x_3\}$.

The single linkage distance between the remaining two clusters:

$$\begin{aligned} d(R, P+Q) &= \frac{1}{2}d(R,P) + \frac{1}{2}d(R,Q) - \frac{1}{2}|d(R,P) - d(R,Q)| \\ &= \min(d(R,P), d(R,Q)) \\ &= \min(d(R,P), d(R,Q)) \\ &= 41 \end{aligned}$$

The reduced distance matrix is then $\begin{pmatrix} 0 & 41 \\ 41 & 0 \end{pmatrix}$.

Single linkage = nearest neighbor!

Dendrogram

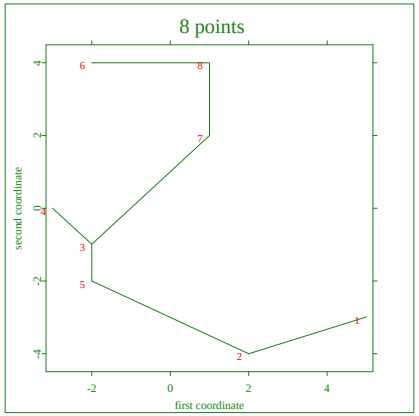
Dendrogram:

- a graphical representation of the sequence of clustering,
- displays the observations, the sequence of clusters and the distances between the clusters.

Construction of dendrogram:

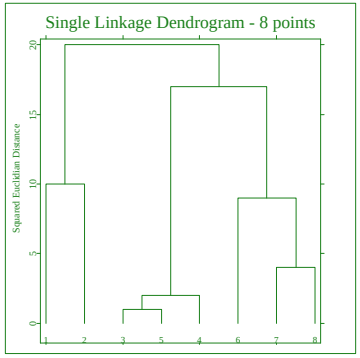
- tree displaying the progress of the agglomerative clustering algorithm,
- the row name (or row number) is given on the horizontal axis.
- the vertical axis gives the distance between clusters.

Example:



The 8 points example:

```
eight=cbind(c(4,2,-2,-3,-2,-2,1,1),c(-3,-4,-1,0,-2, 4,2,4))
```

The dendrogram for the 8 points example (single linkage algorithm with squared Euclidean distances).

```
plot(hclust(dist(eight)^2,method="single"))
```


Cutting the tree

If we decide to cut the tree at the level 10 we define three clusters: $\{1,2\}$, $\{3,4,5\}$ and $\{6,7,8\}$.

```
gr=cutree(hclust(dist(eight)^2,method="single"),k=3)
```

In practice, it is important to interpret the resulting clusters using tables of means and (multivariate) graphics:

```
sapply(data.frame(eight),tapply,gr,
       function(x)sprintf("%0.1f",mean(x)))

plot(eight,col=as.numeric(gr),pch=as.numeric(gr)+1)
```

In practice, the choice of the number of clusters is usually based on the visual inspection of the dendrogram.

Group building algorithms

Single linkage nearest neighbor, tends to build "chains".

Complete linkage furthes neighbor, creates groups where all points are close.

Average linkage computes average distance (compromise between single and complete linkage).

Centroid uses geometrical distance.

Ward joins groups that do not increase too much a given measure of heterogeneity (and creates nice looking homogeneous groups).

In practice, most "usable" results are typically obtained by Ward algorithm.

Ward algorithm

The measure of heterogeneity for a group R is the inertia inside the group:

$$I_R = \frac{1}{n_R} \sum_{i=1}^{n_R} d^2(x_i, \overline{x}_R),$$

where $\overline{x}_R$ is the mean (center of gravity) of the group R.

When two objects or groups P and Q will be joined, the new group $P + Q$ will have a larger inertia I_{P+Q}. The increase of inertia is given by

$$\Delta(P, Q) = \frac{n_P n_Q}{n_P + n_Q} d^2(P, Q).$$

The Ward algorithm joins groups P and Q that give the smallest increase of $\Delta(P, Q)$.

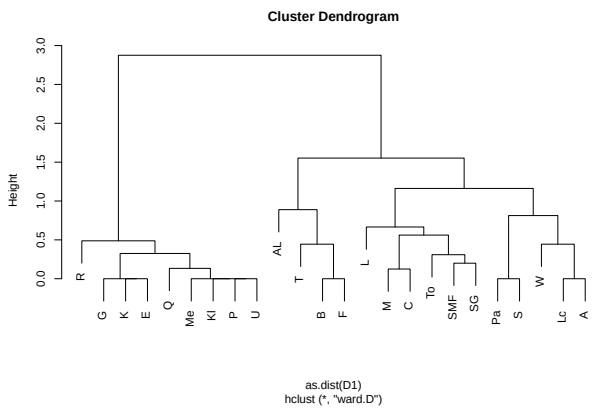
This interpretation is correct if we are working with squared Euclidean distances (note that `hclust()` contains two versions of Ward algorithm).

Example: US companies data set.

```
data(uscomp)
uscomp$Sales=as.numeric(as.character(uscomp$Sales))
uscomp$Sales[65]= 1601
d=dist(scale(uscomp[,c(-7)]))
plot(cluscomp.c<-hclust(d)
plot(cluscomp.s<-hclust(d,method="single"))
plot(cluscomp.w<-hclust(d,method="ward"))

gr3=cutree(cluscomp.w,k=3)
sapply(uscomp[,-7],tapply,gr3,function(x)round(mean(x)))
parcoord(uscomp[,-7],col=as.numeric(gr3))
table(gr3,uscomp[,7])
```

It seems that better results could be obtained by logarithmic transformation of the data set.



The dendrogram for the songbooks example (Ward algorithm based on Jaccard measure): two cluster solution corresponds to the division of Francia into West Francia (more-or-less current France) and East Francia (more-or-less current Germany) in the 9th century (after the death of Charlemagne) → `SMSclussong`

Cluster analysis

- The class of clustering algorithms can be divided into two types: hierarchical and partitioning algorithms. Partitioning algorithms start from a preliminary clustering and optimize given criterion by exchanging group elements.
- Hierarchical agglomerative techniques start from the finest possible structure, compute the distance matrix, and join clusters with the smallest distance. This step is repeated until all points are united in one cluster.
- The agglomerative procedure depends on the definition of the distance between two clusters. Often used distances are single linkage, complete linkage, Ward distance.
- The process of the unification of clusters can be graphically represented by a dendrogram.

Týden 10–11

Diskriminační analýza:

- motivace a maximální věrohodnost,
- lineární a kvadratická diskriminační analýza,
- pravděpodobnost chybné klasifikace,
- Fisherův přístup.

Discriminant analysis

The aim of discriminant analysis is to construct discriminant rules allowing classification of new items (subjects) into known populations Π_j, $j = 1, \ldots, J$.

Discriminant rule is a partition of the sample space:

$$\bigcup_{j=1}^{J} R_j = I\!R^p$$

↑ partition

The new observation is classified into population Π_j if it falls in R_j.

Bayes rule

Suppose that observations from Π_j have density $f_j(x)$ and that the π_j is the prior probability of Π_j.

Using Bayes theorem:

$$P(\Pi_j|X = x) = \frac{f_j(x)\pi_j}{\sum_{i=1}^{J} f_i(x)\pi_i}.$$

Interpreting $P(\Pi_j|X = x)$ as the posterior probability of population Π_j (after observing $X = x$), we classify X to $\Pi_{\text{argmax}_j P(\Pi_j|X)}$.

The corresponding discriminant rule R_j is defined as $\{x : f_j(x)\pi_j \geq f_j i(x)\pi_i, i \neq j\}$ (maximum likelihood).

Example: A discrimination rule based on observations of a one-dimensional variable with an exponential distribution.

The pdf is $f(x) = \lambda \exp\{-\lambda x\}$ for $x > 0$. Comparing the likelihoods for two populations $\Pi_1 : \text{Exp}(\lambda_1)$ and $\Pi_2 : \text{Exp}(\lambda_2)$, we allocate the observation x into population Π_1 if

$$\begin{aligned} L_1(x)/L_2(x) &\geq 1 \\ x(\lambda_1 - \lambda_2) &\leq \log \frac{\lambda_1}{\lambda_2}. \end{aligned}$$

Assuming that $\lambda_1 < \lambda_2$, we obtain:

$$R_1 = \left\{ x : x \geq \frac{\log \lambda_1 - \log \lambda_2}{\lambda_1 - \lambda_2} \right\}.$$

The observation x is classified into Π_1 if it is greater than the constant $(\log \lambda_1 - \log \lambda_2)/(\lambda_1 - \lambda_2)$.

Credit scoring

Example: Let γ denote the gain of the bank from a correctly classified good client. Let Π_2 denote the population of good clients.

Π_1 represents the population of bad clients that bring the loss $C(2|1)$ if they are classified as good clients.

$C(1|2)$ denotes the cost of loosing a good client classified as bad.

The gain of the bank as a function of the discriminant rule "client is good if he falls in region R" is:

$$\begin{aligned} G(R) &= \gamma\pi_2 \int I(x \in R) f_2(x)dx - C(2|1)\pi_1 \int I(x \in R) f_1(x)dx \\ &\quad - C(1|2)\pi_2 \int \{1 - I(x \in R)\} f_2(x)dx \end{aligned}$$

Straightforward calculations lead to

$$R = \left\{ x : \frac{f_2(x)}{f_1(x)} \geq \frac{C(2|1)\pi_1}{\{C(1|2) + \gamma\}\pi_2} \right\}.$$

Theorem: The rule minimizing the Ecpected Cost of Misclassification $ECM = C(2|1)p_{21}\pi_1 + C(1|2)p_{12}\pi_2$ (where p_{ij} is the probability that observation from Π_j falls into region R_i) is given by

$$\begin{aligned} R_1 &= \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{\pi_2}{\pi_1}\right) \right\}, \\ R_2 &= \left\{ x : \frac{f_1(x)}{f_2(x)} < \left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{\pi_2}{\pi_1}\right) \right\}. \end{aligned}$$

Clearly, the Bayes rule is a special case of the ECM rule for equal misclassification costs.

One-dimensional normal distributions

Consider two normal populations $\Pi_1 : N(\mu_1, \sigma_1^2)$ and $\Pi_2 : N(\mu_2, \sigma_2^2)$ and $\pi_1 = \pi_2$.

Then

$$L_i(x) = (2\pi\sigma_i^2)^{-1/2} \exp\left\{ -\frac{1}{2}\left(\frac{x - \mu_i}{\sigma_i}\right)^2 \right\}$$

and $L_1(x) > L_2(x)$ (i.e., $x \in R_1$ is classified to Π_1)

$$\begin{aligned} &\iff \frac{\sigma_2}{\sigma_1} \exp\left\{ -\frac{1}{2}\left[\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - \left(\frac{x - \mu_2}{\sigma_2}\right)^2 \right] \right\} > 1 \\ &\iff x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) < 2\log\frac{\sigma_2}{\sigma_1}. \end{aligned}$$

This is quadratic inequality $\iff \sigma_1^2 \neq \sigma_2^2$.

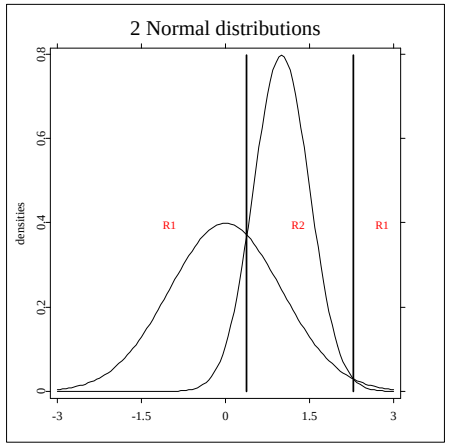
The quadratic rule classifies distant observations into the group with larger variance.

Example: Suppose that $\mu_1 = 0$, $\sigma_1 = 1$ and $\mu_2 = 1$, $\sigma_2 = \frac{1}{2}$:

$$\begin{aligned} R_1 &= \left\{ x : x < \frac{1}{3}\left(4 - \sqrt{4 + 6\log(2)}\right) \text{ or } x > \frac{1}{3}\left(4 + \sqrt{4 + 6\log(2)}\right) \right\}, \\ R_2 &= I\!R \setminus R_1. \end{aligned}$$

If $\sigma_1 = \sigma_2$ then (for $\mu_1 < \mu_2$) we obtain a very simple linear discriminant rule:

$$\begin{aligned} R_1 &= \{x : x \leq \frac{1}{2}(\mu_1 + \mu_2)\}, \\ R_2 &= \{x : x > \frac{1}{2}(\mu_1 + \mu_2)\}. \end{aligned}$$



Maximum likelihood rule for one-dimensional normal distributions with different variances.

Multinormal distribution with common variance matrix

Suppose Π_i : $N_p(\mu_i, \Sigma)$.

The Bayes rule (assuming equal prior probabilities) allocates x to Π_j, where $j \in \{1, \ldots, J\}$ is the value that minimizes the square Mahalanobis distance between x and μ_i:

$$\delta^2(x, \mu_i) = (x - \mu_i)^\top \Sigma^{-1}(x - \mu_i) \ , \ i = 1, \ldots, J.$$

In the case of $J = 2$: x is allocated to Π_1 if

$$(x - \mu_1)^\top \Sigma^{-1}(x - \mu_1) < (x - \mu_2)^\top \Sigma^{-1}(x - \mu_2).$$

Multinormal distribution with common variance matrix

Rearranging terms leads to:

$$\begin{aligned} -2\mu_1^\top \Sigma^{-1} x + 2\mu_2^\top \Sigma^{-1} x + \mu_1^\top \Sigma^{-1}\mu_1 - \mu_2^\top \Sigma^{-1}\mu_2 &< 0 \\ 2(\mu_2 - \mu_1)^\top \Sigma^{-1} x + (\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 + \mu_2) &< 0 \\ (\mu_1 - \mu_2)^\top \Sigma^{-1}\{x - \frac{1}{2}(\mu_1 + \mu_2)\} &> 0 \\ \alpha^\top (x - \mu) &> 0, \end{aligned}$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

The resulting discriminant rule is linear (see also R command `lda()`).

Probability of misclassification

Suppose that Π_i : $N_p(\mu_i, \Sigma)$.

Consider

$$p_{12} = P(x \in R_1 \mid \Pi_2) = P\{\alpha^\top (x - \mu) > 0 \mid \Pi_2\}$$

In Π_2, $\alpha^\top (X - \mu) \sim N\left(-\frac{1}{2}\delta^2, \delta^2\right)$ where $\delta^2 = (\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)$ is the squared Mahalanobis distance between the two populations, we obtain

$$p_{12} = \Phi\left(-\frac{1}{2}\delta\right).$$

Similarly, we obtain the probability of misclassification into population 2 for x from Π_1 as $p_{21}{=}\Phi\left(-\frac{1}{2}\delta\right)$.

Two multinormal distributions

Assuming that $\Pi_i: N_p(\mu_i, \Sigma_i)$, for $i = 1, 2$, the discriminant rule becomes more complicated.

$$R_1 = \left\{ x : -\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}) x - k \right.$$

$$\left. \geq \ln \left[\left\{ \frac{C(1|2)}{C(2|1)} \right\} \left\{ \frac{\pi_2}{\pi_1} \right\} \right] \right\}$$

where $k = \frac{1}{2} \ln \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2} (\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_2^{-1} \mu_2)$.

This is a *quadratic* classification rule (notice that $\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x = 0$ if $\Sigma_1 = \Sigma_2$).

Discriminant Analysis

- Discriminant analysis is a set of methods for distinguishing between groups in data and allocating new observations into groups.
- The Bayes discriminant rule allocates an observation x to the population Π_j that maximizes $\max_j \pi_j f_j(x)$.
- For the ML rule and $J = 2$ multivariate normal populations, the discriminant rule can be derived from ratio of the densities. The discriminant rule is linear for common variance matrices and quadratic if the variance matrices are different.
- For the ML rule and $J = 2$ normal populations with common variance matrix, the probabilities of misclassification are given by $p_{12} = p_{21} = \Phi\left(-\frac{1}{2}\delta\right)$ where δ is the square root of the Mahalanobis distance between the 2 populations.

Discriminant rules in practice

The unknown parameters (μ_j, Σ_j) are estimated by $(\bar{x}_j, S_j)$ in each Π_j.

The common variance matrix Σ can be estimated by the pooled variance matrix $\mathcal{S}_u = \sum_{j=1}^{J} n_j \left(\frac{\mathcal{S}_j}{n-J} \right)$, where $n = \sum_{j=1}^{J} n_j$.

R library `MASS` contains the following simple functions for discriminant analysis:

- `lda()`: linear discriminant analysis (assuming equal variance matrices),
- `qda()`: quadratic discriminant analysis (with possibly different variance matrices).

Example:

```
library(MASS);library(MSES);data(bank2)
lda.b2=lda(bank2,pf<-rep(c("Prave","Fales"),each=100))
lda.b2

table(predict(lda.b2,bank2)$class,pf)

qda.b2=qda(bank2,pf)
qda.b2

table(predict(lda.b2,bank2)$class,pf)

?lda
?qda
```

Note: applying `lda()` with x_i, x_i^2, and $x_i x_j$ is similar (but not equivalent) to `qda()`.

Apparent and actual error rate

The apparent error rate (APER) is defined as the percentage of misclassified observations. APER is based on the observations which were used to construct the discriminant rule and it might be too optimistic.

In order to obtain a more appropriate estimate of the misclassification probability, we may use simple leave-one-out (or cross-validation) algorithm:

1. Calculate the discrimination rule from all but one observation.
2. Allocate the omitted observation according to the rule from step 1.
3. Repeat steps 1 and 2 for all observations and count the number of correct and wrong classifications.

The estimate of the misclassification rate based on this procedure is called the actual error rate (AER).

Example:

```
lda.b2.cv=lda(bank2,pf,CV=TRUE)

table(lda.b2.cv$class,pf)

qda.b2.cv=qda(bank2,pf,CV=TRUE)

table(lda.b2.cv$class,pf)
```


Three (or more) groups

Allocation regions for $J = 3$ groups:

$$\begin{aligned} h_{12}(x) &= (\bar{x}_1 - \bar{x}_2)^\top \mathcal{S}_u^{-1}(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)) \\ h_{13}(x) &= (\bar{x}_1 - \bar{x}_3)^\top \mathcal{S}_u^{-1}(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_3)) \\ h_{23}(x) &= (\bar{x}_2 - \bar{x}_3)^\top \mathcal{S}_u^{-1}(x - \frac{1}{2}(\bar{x}_2 + \bar{x}_3)) . \end{aligned}$$

The ML rule is to allocate x to

$$\begin{cases} \Pi_1 & \text{if } h_{12}(x) > 0 \text{ and } h_{13}(x) > 0 \\ \Pi_2 & \text{if } h_{12}(x) < 0 \text{ and } h_{23}(x) > 0 \\ \Pi_3 & \text{if } h_{13}(x) < 0 \text{ and } h_{23}(x) < 0. \end{cases}$$

In R, discriminant analysis with 3 groups works differently.

Example:

```
data(iris)

## training data set
train=sample(1:150,75);table(iris$Species[train])

z=lda(Species~.,iris,prior=c(1,1,1)/3,subset = train)
table(predict(z, iris[-train,-5])$class,
              iris[-train,"Species"])

## cross-validation
z.cv.cl=lda(Species ~ ., iris, prior = c(1,1,1)/3,
            CV=TRUE)$class
z.al.cl=predict(lda(Species ~ ., iris, prior =
                    c(1,1,1)/3),iris[,-5])$class
```

Fisher's approach

Based on projections $\mathcal{Y} = \mathcal{X}a$ of the original data set $\mathcal{X}$.

Projections leading to a good separation are found by maximizing the ratio of the *between-group-sum of squares* to the *within-group-sum of squares*.

The within-sum-of-squares measures the sum of variations within each group:

$$\sum_{j=1}^{J} \mathcal{Y}_j^\top \mathcal{H}_j \mathcal{Y}_j = \sum_{j=1}^{J} a^\top \mathcal{X}_j^\top \mathcal{H}_j \mathcal{X}_j a = a^\top \mathcal{W} a,$$

where $\mathcal{Y}_j$ denotes the j-th submatrix of $\mathcal{Y}$ corresponding to observations of group j and $\mathcal{H}_j$ denotes the $(n_j \times n_j)$ centering matrix.

The between-sum-of-squares is

$$\sum_{j=1}^{J} n_j(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{J} n_j\{a^\top(\bar{x}_j - \bar{x})\}^2 = a^\top \mathcal{B} a.$$

The total-sum-of-squares $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathcal{Y}^\top \mathcal{H} \mathcal{Y} = a^\top \mathcal{X}^\top \mathcal{H} \mathcal{X} a = a^\top \mathcal{T} a$ can be decomposed as

$$\begin{array}{lcllcl} \text{total SS} & = & \text{within SS} & + & \text{between SS.} \\ a^\top \mathcal{T} a & = & a^\top \mathcal{W} a & + & a^\top \mathcal{B} a \end{array}$$

The idea is to select a maximizing maximizes the ratio

$$\frac{a^\top \mathcal{B} a}{a^\top \mathcal{W} a}.$$

Theorem: The vector a that maximizes $\frac{a^\top \mathcal{B} a}{a^\top \mathcal{W} a}$ is the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ that corresponds to the largest eigenvalue.

Idea of the proof: see Theorem on maximization of quadratic forms.

Discrimination rule: We classify x into the group j for which $a^\top \bar{x}_j$ is closest to $a^\top x$, i.e.,

$$x \to \Pi_j \text{ where } j_0 = \arg\min_j |a^\top(x - \bar{x}_j)|.$$

Example: For two groups of sizes n_1 and n_2, we obtain:

$$\begin{aligned} a^\top \mathcal{B} a &= n_1\{a^\top(\bar{x}_1 - \bar{x})\}^2 + n_2\{a^\top(\bar{x}_2 - \bar{x})\}^2 \\ &= n_1\{a^\top(\bar{x}_1 - \bar{x}_2)/2\}^2 + n_2\{a^\top(\bar{x}_1 - \bar{x}_2)/2\}^2 \\ &= \frac{n_1 + n_2}{4}\{a^\top(\bar{x}_1 - \bar{x}_2)^2\} \end{aligned}$$

Clearly, $\mathcal{B} = \{(n_1 + n_2)/4\} dd^\top$, where $d = (\bar{x}_1 - \bar{x}_2)$ and the largest eigenvalue of $\mathcal{W}^{-1}\mathcal{B}$ is $(n_1 + n_2/4) d^\top \mathcal{W}^{-1} d$.

Therefore, the corresponding eigenvector has to satisfy:

$$\begin{aligned} \mathcal{W}^{-1}\mathcal{B}\gamma &= \{(n_1 + n_2)/4\} d^\top \mathcal{W}^{-1} d\gamma \\ \mathcal{W}^{-1} dd^\top \gamma &= d^\top \mathcal{W}^{-1} d\gamma \end{aligned}$$

leading $\gamma = \mathcal{W}^{-1} d = \mathcal{W}^{-1}(\bar{x}_1 - \bar{x}_2)$.

Proportion of trace and more groups

In this way, we find only one direction maximizing the differences between two groups.

For three groups, $\text{rank}(\mathcal{B}) = 2$, and we obtain two directions (i.e., a linear transformation of the original data set maximizing the between-group differences w.r.t. the within-group variability). The eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$ correspond to the importance of these directions (its percentages can be interpreted as percentages of between-group differences explained by the corresponding directions).

For g groups, $\text{rank}(\mathcal{B}) \leq \min(p, g-1)$. I.e., we obtain at most $g-1$ linear discriminants.

Example:

```
data(iris)

## training data set
train=sample(1:150,75);table(iris$Species[train])

z=lda(Species~.,iris,prior=c(1,1,1)/3,subset = train)
pz<-predict(z, iris[-train,-5])
table(pz$class,iris[-train,"Species"])
eqscplot(pz$x, type="n",xlab="LD1",ylab="LD2")
spec=as.numeric(iris[-train,5],1,1)
text(pz$x,labels=spec,col=spec)
z ## see "proportion of trace"
```


Other usable methods

logistic regression

classification trees

k-nearest neighbors

support vector machine

neural networks

Discrimination Rules in Practice

- Linear discriminant rule allocates x to the population with smallest Mahalanobis distance

$$\delta^2(x;\mu_i) = (x-\mu_i)^\top \Sigma^{-1}(x-\mu_i).$$

- Classification for different covariance structures in the two populations leads to quadratic discrimination rules.
- The probability of misclassification can be estimated by cross-validation.
- Fisher's linear discrimination finds a linear combination $a^\top x$ that maximizes the ratio of the "between-sum-of-squares" and the "within-sum-of-squares". This rule is identical to the (linear) ML rule in the case of $J = 2$ for normal populations.

Týden 11

Kanonické korelace:

- kanonické proměnné, kanonické vektory a kanonické korelace,
- praktické použití a příklad.

Canonical correlations

We have random vectors $X \in \mathbb{R}^q$ and $Y \in \mathbb{R}^p$.

Consider linear combinations:

$$a^\top X \quad \text{and} \quad b^\top Y$$

Correlation of the linear combinations:

$$\rho(a, b) = \rho_{a^\top X\, b^\top Y}.$$

We want to find a, b maximizing the correlation $\rho(a, b)$.

The linear combinations $a^\top X$ and $b^\top Y$ describe the structure of "common variability" of X and Y.

Assuming that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \right),$$

we want to find a, b maximizing the correlation $\rho(a, b) = \rho_{a^\top X\, b^\top Y}$.

Note that $\rho(ca, b) = \rho(a, b)$ for any $c \in \mathbb{R}$. Therefore, we can maximize $a^\top \Sigma_{XY} b$ under the constraints $a^\top \Sigma_{XX} a = b^\top \Sigma_{YY} b = 1$.

And this is the same as maximizing $u^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} v$ under the constraints $\|u\| = \|v\| = 1$.

Denoting $\mathcal{K} = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$, we have $u^\top \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} v = u^\top \mathcal{K} v$.

Clearly, for each v fixed such that $\|v\| = 1$, we have the following

$$\begin{aligned} \max_{u, \|u\|=1} (u^\top \mathcal{K} v)^2 &\leq \max_{u, \|u\|=1} u^\top \mathcal{K} v v^\top \mathcal{K}^\top u \\ &= v^\top \mathcal{K}^\top \mathcal{K} v \\ &\leq \lambda_1, \end{aligned}$$

where λ_1 is the largest eigenvalue of $\mathcal{K}^\top \mathcal{K}$.

The SVD decomposition $\mathcal{K} = \Gamma \Lambda^{1/2} \Delta^\top$ with $k = \text{rank}(\mathcal{K})$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$ then leads

$$\gamma_1^\top \mathcal{K} \delta_1 = \lambda_1^{1/2} (= \rho(\Sigma_{XX}^{-1/2} \gamma_1, \Sigma_{YY}^{-1/2} \delta_1)).$$

Theorem: Define $f_r = \max_{a,b} a^\top \Sigma_{XY} b$ under the constraints $a^\top \Sigma_{XX} a = b^\top \Sigma_{YY} b = 1$ and $a_i^\top \Sigma_{XX} a = b_i^\top \Sigma_{YY} b = 0$ for $i = 1, \ldots, r-1$ (for some $r \in \{1, \ldots, k\}$ fixed).

The maximum of $\rho(a, b)$ under the above constraints is given by f_r and it is attained when $a = a_r = \Sigma_{XX}^{-1/2} \gamma_r$ and $b = b_r = \Sigma_{YY}^{-1/2} \delta_r$.

The correlation $\rho(a, b)$ is maximized for $a = a_1$ and $b = b_1$ and $\rho(a_1, b_1) = \lambda_1^{1/2}$ is the correlation of random variables η_1 and φ_1.

The vectors a_r and b_r maximize the correlation subject to the condition that $a^\top X$ and $b^\top X$ are uncorrelated with the previous canonical variables $a_i^\top X$ and $b_i^\top X$, respectively.

Terminology

Canonical correlation vectors

$$a_i = \Sigma_{XX}^{-1/2} \gamma_i$$
$$b_i = \Sigma_{YY}^{-1/2} \delta_i$$

Canonical variables

$$\eta_i = a_i^\top X$$
$$\varphi_i = b_i^\top Y$$

Canonical correlation coefficients

$$\lambda_1^{1/2}, \ldots, \lambda_k^{1/2}$$

Properties

Theorem: Let η and φ be the canonical variables, i.e., the components of the vector η are

$$\eta_i = \left(\Sigma_{XX}^{-1/2} \gamma_i\right)^\top X,$$

and the components of the vector φ are

$$\varphi_i = \left(\Sigma_{YY}^{-1/2} \delta_i\right)^\top Y,$$

for $1 \le i \le k$. Then

$$\mathrm{Var} \begin{pmatrix} \eta \\ \varphi \end{pmatrix} = \begin{pmatrix} \mathcal{I} & \Lambda^{1/2} \\ \Lambda^{1/2} & \mathcal{I} \end{pmatrix},$$

where $\Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})$.

Relation to principal components

Both PC and CC are calculated using eigenvalues and eigenvectors of some (covariance) matrices.

PC analysis decomposes the "total variability" of one dataset.

CC analysis decomposes the total "common variability" of two datasets. The "common variability" is described in terms of linear combinations (it is common and therefore we get description of the common variability in terms of both datasets).

The canonical variables in both datasets are related: the first canonical variable in the first dataset describes the same part of the common variablity as the first canonical variability in the second dataset.

Canonical correlation analysis

- Canonical correlation analysis aims to identify possible links between two (sub-)sets of variables $X \in \mathbb{R}^q$ and $Y \in \mathbb{R}^p$. The idea is to find indices $a^\top X$ and $b^\top Y$ such that the correlation $\rho(a,b) = \rho_{a^\top X b^\top Y}$ is maximal.
- The maximum correlation is found by $a_i = \Sigma_{XX}^{-1/2}\gamma_i$ and $b_i = \Sigma_{YY}^{-1/2}\delta_i$, where γ_i and δ_i denote the eigenvectors of $\mathcal{K}\mathcal{K}^\top$ and $\mathcal{K}^\top\mathcal{K}$, $\mathcal{K} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$.
- The vectors a_i and b_i are the canonical correlation vectors, $\eta_i = a_i^\top X$ and $\varphi_i = b_i^\top Y$ are the canonical variables.
- The covariance between the canonical variables is $\mathrm{cov}(\eta_i, \varphi_i) = \sqrt{\lambda_i}$, $i = 1, \ldots, k$.
- Canonical correlations are invariant w.r.t. linear transformations of the original variables X and Y.

Canonical Correlations in Practice

In practice, the covariance matrices Σ_{XX}, Σ_{XY}, Σ_{YY} are estimated by sample covariance matrices S_{XX}, S_{XY}, S_{YY}. The canonical correlation analysis is carried out on the estimates.

Before running the analysis, one should test the hypothesis of independence between X and Y (using, e.g., the ML test described previously):

Let $Z_i = (X_i^\top, Y_i^\top)^\top \sim N_{q+p}(\mu, \Sigma)$, $i = 1 \cdots, n$ be independent,

$$H_0 : \Sigma_{XY} = 0, \qquad H_1 : \text{ no constraints.}$$

Test of independence

We have already seen that $-2\log\lambda = -n\log|\mathcal{I} - \mathcal{S}_{22}^{-1}\mathcal{S}_{21}\mathcal{S}_{11}^{-1}\mathcal{S}_{12}|$.

It can be shown that this LRT test statistic is distributed as a ratio of determinants of independent Wishart matrices (this is the *Wilks' lambda distribution*).

For large values of n, the Wilks' lambda distribution can be approximated (Bartlett's approximation):

$$-\{n - (p+q+3)/2\}\log|\mathcal{I} - \mathcal{S}_{22}^{-1}\mathcal{S}_{21}\mathcal{S}_{11}^{-1}\mathcal{S}_{12}| \sim \chi^2_{pq},$$

i.e., we reject independence of X and Y if

$$-\{n - (p+q+3)/2\}\log|\mathcal{I} - \mathcal{S}_{22}^{-1}\mathcal{S}_{21}\mathcal{S}_{11}^{-1}\mathcal{S}_{12}| > \chi^2_{qp}(1-\alpha).$$

Example: What is the relationship between the datasets on US crimes (murder, rape, robbery, assault, burglary, larceny, autotheft) and US health (accident, cardiovascular, cancer, pulmonar, pneumonia, diabetis, liver)?

```
data(uscrime)
x=sqrt(as.matrix(uscrime[,3:9]))
x=scale(x)

data(ushealth)
y=sqrt(as.matrix(ushealth[,3:9]))
y=scale(y)

n=nrow(x); p=ncol(x); q=ncol(y)
```

x denotes US crimes
y denotes US health

```
sxx=cov(x);syy=cov(y);sxy=cov(x,y)

t=-(n-(p+q+3)/2)*log(det(diag(1,q)-
  solve(syy)%*%t(sxy)%*%solve(sxx)%*%sxy))
format.pval(1-pchisq(t,p*q)) ## test of independence

e=eigen(sxx)
sxx12=e$vectors%*%(sqrt(diag(1/e$values)))%*%t(e$vectors)
e=eigen(syy)
syy12=e$vectors%*%(sqrt(diag(1/e$values)))%*%t(e$vectors)
kkt=sxx12%*%sxy%*%syy12%*%syy12%*%t(sxy)%*%sxx12
ktk=syy12%*%t(sxy)%*%sxx12%*%sxx12%*%sxy%*%syy12
e1=eigen(kkt)
e2=eigen(ktk)
print(cbind(e1$values,e2$values))
a=sxx12%*%e1$vectors
b=syy12%*%e2$vectors
```



```
## canonical variables
cvx=x%*%a
cvy=y%*%b

## plot of the first pair
plot(cvx[,1],cvy[,1],type="n")
text(cvx[,1],cvy[,1],row.names(ushealth))

## canonical correlation
cor(cvx[,1],cvy[,1])
sqrt(e1$values[1])

## R library stats
cancor(x,y)
## coefficients are divided by sqrt(dim)?
```


Canonical Correlations in Practice

- In practice, we estimate Σ_{XX}, Σ_{XY}, Σ_{YY} by the empirical covariances and to compute estimates ℓ_i, g_i, d_i for λ_i, γ_i, δ_i from the SVD of $\widehat{\mathcal{K}} = \mathcal{S}_{XX}^{-1/2}\mathcal{S}_{XY}\mathcal{S}_{YY}^{-1/2}$.
- The coefficients of the canonical variables (i.e., the canonical vectors) tell us the influence of these variables.
- The independence of the two random vectors can be tested by a likelihood ratio test leading to Wilks' lambda distribution.
- Barlett's test of the null hypothesis *that only s population canonical correlation coefficients are non zero* is based on the statistic $-\{n-(p+q+3)/2\}\log\prod_{i=s+1}^{\min(p,q)}(1-r_i) \sim \chi^2_{(p-s)(q-s)}$, where r_i are the sample canonical correlation coefficients.

Týden 12

Korespondenční analýza:

- testy nezávislosti v kontingenční tabulce,
- reprezentace řádků a sloupců.

Correspondence analysis

Categorical scales are pervasive in the social sciences for measuring attitudes and opinions on various issues and demographic characteristics such as gender, race, and social class.

Categorical scales (...) occur frequently in the behavioral sciences, public health, ecology, education, and marketing. They even occur in highly quantitative fields such as engineering sciences and industrial quality control. Such applications often involve subjective evaluation of some characteristic—how soft to the touch a certain fabric is, how good a particular food product tastes, or how easy a worker finds a certain task to be.

(Alan Agresti, *Categorical Data Analysis*, Wiley, 1990)

Two-way contingency table

Variable Z has I levels

Variable Y has J levels

This gives IJ combinations of levels of Z and Y

We count the responses (Z, Y) in our sample and display this information in rectangular table which has I rows and J columns.

In each *cell* we give the number of subjects in our sample having the corresponding combination of responses on Z and Y.

The entry x_{ij} in the *contingency table* $\mathcal{X}(n \times p)$ is the number of observations in a sample that simultaneously fall in the ith row category and the jth column category.

Example:

$$\mathcal{X} = \left(\begin{array}{ccc|c} 8 & 4 & 3 & \\ 2 & 1 & 6 & \\ 3 & 6 & 2 & \\ \hline 13 & 11 & 11 & 35 \end{array}\right)$$

← Beer
← Wine
← Spirit

↑ Czechia
↑ Russia
↑ GB

Joint distribution: $\pi_{ij} = P(Z = i, Y = j)$ is the probability that Z is equal to i and at the same time Y is j.

Marginal distribution of Z: $\pi_{i\cdot}$ is the probability that Z is equal to i

Marginal distribution of Y: $\pi_{\cdot j}$ is the probability that Y is equal to j

Sampling Distributions

This is the way in which the table was created. It is important for understanding the table correctly.

The likelihoods depend on the sampling distribution.

Poisson sampling: everything is random,
Multinomial sampling: total number of observed subjects is fixed,
Independent multinomial sampling: number of subject in each row or column is fixed.

Estimators and likelihood ratio tests are often identical for all types of sampling (NMST432 Advanced Regression Models).

Maximum Likelihood Estimates

By maximizing the likelihood function we obtain the ML estimator

$$\hat{\pi}_{ij} = p_{ij} = x_{ij}/x_{\bullet\bullet},$$

where $x_{\bullet\bullet} = \sum_{i=1}^{n} x_{i\bullet} = \sum_{j=1}^{n} x_{\bullet j}$ is the total number of observations.

Notice that Z and Y are independent if for all i and j: $\pi_{i|j} = \pi_{ij}/\pi_{\cdot j} = \pi_{i\cdot}$ or $\pi_{j|i} = \pi_{ji}/\pi_{i\cdot} = \pi_{\cdot j}$ or or $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$.

The ML estimators of cell probabilities π_{ij} under independence are

$$\hat{\pi}_{ij} = p_{i\bullet}p_{\bullet j} = (x_{i\bullet}x_{\bullet j})/x_{\bullet\bullet}^2,$$

$x_{i\bullet} = \sum_{j=1}^{n} x_{ij}$ is the number of observations falling into the ith row category.

Test of independence

Likelihood-Ratio Test of Independence can be derived by following standard arguments.

The χ^2 *test of independence* is more popular. It is based on differences between the observed frequencies x_{ij} and E_{ij}, the estimated expected values under the assumption of independence, i.e.,

$$E_{ij} = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}.$$

Under the hypothesis of independence of the row and column categories, the statistic

$$t = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - E_{ij})^2/E_{ij}$$

has a $\chi^2_{(n-1)(p-1)}$ distribution.

Example: Alcohol consumption in three countries.

```
alc=matrix(c(8, 4, 3, 2, 1, 6, 3, 6, 2),3,byrow=T)
row.names(alc)=c("Beer","Wine","Spirit")
colnames(alc)=c("Czechia","Russia","GB")

chisq.test(alc)
```

```
        Pearson's Chi-squared test

data:  alc
X-squared = 9.8406, df = 4, p-value = 0.0432

Warning message: Chi-squared approximation may be incorrect
```


Individual contributions

The correspondence analysis is targeted toward the analysis of the individual contributions to the χ^2-statistic:

$$c_{ij} = (x_{ij} - E_{ij})/E_{ij}^{1/2}, \qquad (1)$$

which may be viewed as a measure of the departure of the observed x_{ij} from independence.

Example:

```
a=apply(alc,1,sum); b=apply(alc,2,sum); n=sum(alc)
E=a%*%t(as.matrix(b))/n
round(C<-(alc-E)/sqrt(E),2)
```

Decomposition of χ^2-statistic

The SVD of $\mathcal{C} = (c_{ij})_{i=1,\ldots,n;j=1,\ldots,p}$ yields

$$\mathcal{C} = \Gamma \Lambda^{1/2} \Delta^\top$$

with $\Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_R^{1/2})$, where $\lambda_1, \ldots, \lambda_R$ are the nonzero eigenvalues of both $\mathcal{C}^\top\mathcal{C}$ and $\mathcal{C}\mathcal{C}^\top$.

Now, it is easy to see that

$$t = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - E_{ij})^2/E_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{p} c_{ij}^2 = \mathrm{tr}(\mathcal{C}\mathcal{C}^\top) = \sum_{k=1}^{R}\lambda_k.$$

Hence, the SVD of the matrix $\mathcal{C}$ decomposes the χ^2-statistic t.

Example:

```
decomp=svd(C)

gamma1=decomp$u[,1]
delta1=decomp$v[,1]

lambda=decomp$d

sum(lambda^2)  ## chi2 statistika
```


Marginal frequencies

Defining $\mathcal{A} = \mathrm{diag}(x_{i\bullet})$ and $\mathcal{B} = \mathrm{diag}(x_{\bullet j})$ leads the vectors of marginal row and column frequencies:

$$a = \mathcal{A}1_n \quad \text{and} \quad b = \mathcal{B}1_p.$$

This allows to write $E = ab^\top x_{\bullet\bullet}^{-1}$ and $\mathcal{C} = \mathcal{A}^{-1/2}(\mathcal{X} - E)\mathcal{B}^{-1/2}\sqrt{x_{\bullet\bullet}}$.

It is easy to verify that

$$\begin{aligned} \mathcal{C}\sqrt{b} = 0 \quad &\text{and} \quad \mathcal{C}^\top\sqrt{a} = 0, \\ \delta_k^\top\sqrt{b} = 0 \quad &\text{and} \quad \gamma_k^\top\sqrt{a} = 0. \end{aligned}$$

Row and column coordinates

The *row coordinates* $r_k = \mathcal{A}^{-\frac{1}{2}}\mathcal{C}\delta_k$ and *column coordinates* $s_k = \mathcal{B}^{-\frac{1}{2}}\mathcal{C}^\top\gamma_k$ satisfy

$$r_k^\top a = \delta_k^\top\mathcal{C}^\top\mathcal{A}^{-\frac{1}{2}}a = \delta_k^\top\mathcal{C}^\top\sqrt{a} = \delta_k^\top 0 = 0$$

and

$$s_k^\top b = \gamma_k^\top\mathcal{C}\mathcal{B}^{-\frac{1}{2}}b = \gamma_k^\top\mathcal{C}\sqrt{b} = \gamma_k^\top 0 = 0.$$

The *true meaning* of relations $r_k^\top = 0$ and $s_k^\top b = 0$ is

$$\bar{r}_k = \frac{1}{x_{\bullet\bullet}}r_k^\top a = 0 \qquad \text{and} \qquad \bar{s}_k = \frac{1}{x_{\bullet\bullet}}s_k^\top b = 0,$$

where means are (of course) weighted by the row and column marginal frequencies. Hence, both row and column factors are centered.

Example:

```
decomp=svd(C); gamma1=decomp$u[,1]; delta1=decomp$v[,1]
A=diag(a); B=diag(b)

r1=diag(1/sqrt(a))%*%C%*%delta1
s1=diag(1/sqrt(b))%*%t(C)%*%gamma1

row.names(r1)=row.names(C)
row.names(s1)=colnames(C)
```


Variance of row and column factors

For the sample variances of r_k and s_k we have the following:

$$\widehat{\mathrm{Var}}(r_k) = \frac{1}{x_{\bullet\bullet}} \sum_{i=1}^{n} x_{i\bullet} r_{ki}^2 = r_k^\top \mathcal{A} r_k / x_{\bullet\bullet} = \delta_k^\top \mathcal{C}^\top \mathcal{C} \delta_k / x_{\bullet\bullet} = \frac{\lambda_k}{x_{\bullet\bullet}},$$

$$\widehat{\mathrm{Var}}(s_k) = \frac{1}{x_{\bullet\bullet}} \sum_{j=1}^{p} x_{\bullet j} s_{kj}^2 = s_k^\top \mathcal{B} s_k / x_{\bullet\bullet} = \gamma^\top \mathcal{C} \mathcal{C}^\top \gamma_k / x_{\bullet\bullet} = \frac{\lambda_k}{x_{\bullet\bullet}}.$$

In practice, statistical software may return differently scaled values (than r_k nad s_k). Functions `corresp()` and `ca()` in R libraries `MASS` and `ca` standardize row and column factors by $\rho_k = (\lambda_k / x_{\bullet\bullet})^{1/2}$.

This means that the row and column factors given by standard software are standardized.

Example:

```
## prumery
sum(r1*a)
sum(s1*b)

## rozptyly
sum((r1^2)*a)/n
sum((s1^2)*b)/n

##
(lambda[1]^2)/n
```


Proportion of explained variance

Hence, the proportion of the variance explained by the kth factor is

$$\widehat{\mathrm{Var}}(r_k) / \sum_{i=1}^{R} \widehat{\mathrm{Var}}(r_k) = \lambda_k / \sum_{i=1}^{R} \lambda_i.$$

The variance of the kth row factor, $\widehat{\mathrm{Var}}(r_k)$, can be further decomposed into the absolute single row contributions defined as

$$C_a(i, r_k) = \frac{x_{i\bullet} r_{ki}^2}{\lambda_k}, \text{ for } i = 1, \ldots, n, \ k = 1, \ldots, R.$$

Similarly $C_a(j, s_k) = x_{\bullet j} s_{kj}^2 / \lambda_k$ for $j = 1, \ldots, p, \ k = 1, \ldots, R$ are the absolute contributions of column j to the variance of the column factor s_k.

These absolute contributions may help to interpret the row and column factors obtained by the correspondence analysis.

Relation between row and column coordinates

From the properties of SVD we know the relationship between δ_k and γ_k:

$$\delta_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{C}^\top\gamma_k \quad \text{and} \quad \gamma_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{C}\delta_k.$$

Therefore

$$s_k = \mathcal{B}^{-1/2}\mathcal{C}^\top\gamma_k = \sqrt{\lambda_k}\mathcal{B}^{-1/2}\delta_k.$$

Using the definition of r_k, we have

$$\begin{aligned} r_k &= \mathcal{A}^{-1/2}\mathcal{C}\delta_k = \sqrt{x_{\bullet\bullet}}\mathcal{A}^{-1/2}\mathcal{A}^{-1/2}(\mathcal{X}-E)\mathcal{B}^{-1/2}\delta_k \\ &= \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}}\mathcal{A}^{-1}(\mathcal{X}-E)s_k = \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}}\mathcal{A}^{-1}\left(\mathcal{X}s_k - \frac{ab^\top s_k}{x_{\bullet\bullet}}\right) \\ &= \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}}\mathcal{A}^{-1}\mathcal{X}s_k. \end{aligned}$$

Example: Plot of two pairs of indices.

```
gamma2=decomp$u[,2]; delta2=decomp$v[,2]
r2=diag(1/sqrt(a))%*%C%*%delta2
s2=diag(1/sqrt(b))%*%t(C)%*%gamma2

cumsum(lambda^2)/sum(lambda^2)

plot(NULL,xlim=c(-1,1),ylim=c(-1,1),xlab="",ylab="")
text(r1,r2,labels=row.names(C))
text(s1,s2,labels=colnames(C),col="blue")

library(ca)
calc=ca(alc)
plot(calc)
```


Covariance

$$\begin{aligned} \widehat{\mathrm{Cov}}(r_k, s_k) &= \frac{1}{x_{\bullet\bullet}}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}s_{kj} \\ &= r_k^\top\mathcal{X}s_k/x_{\bullet\bullet} = \sqrt{\frac{\lambda_k}{x_{\bullet\bullet}}}r_k^\top\mathcal{A}r_k/x_{\bullet\bullet} \\ &= \sqrt{\frac{\lambda_k}{x_{\bullet\bullet}}}\widehat{\mathrm{Var}}(r_k) \\ &= \sqrt{\frac{\lambda_k}{x_{\bullet\bullet}}}\frac{\lambda_k}{x_{\bullet\bullet}} \end{aligned}$$

$$\widehat{\mathrm{Cov}}(r_1, r_2) = \frac{1}{x_{\bullet\bullet}}\sum_{i=1}^{n} x_{i\bullet}r_{1i}r_{2i} = r_1^\top\mathcal{A}r_2/x_{\bullet\bullet} = \delta_1^\top\mathcal{C}^\top\mathcal{C}\delta_2/x_{\bullet\bullet} = 0$$

Correlation

It follows that the sample correlation coefficient of r_k and s_k is:

$$\rho_k = \sqrt{\frac{\lambda_k}{x_{\bullet\bullet}}}.$$

This means that the correlation structure of the row and column coordinates (in correspondence analysis) is similar to the structure of canonical variables (in canonical correlation analysis).

Example: Analysis of Journaux data set (newspapers in Belgium).

```
# load data
data(journaux); x  = journaux;a  = rowSums(x);b  = colSums(x)
e  = matrix(a) %*% b/sum(a)
# chi-matrix
cc = (x - e)/sqrt(e)

# singular value decomposition
sv = svd(cc);g  = sv$u;l  = sv$d;d  = sv$v

# eigenvalues
ll = l * l

# cumulated percentage of the variance
aux  = cumsum(ll)/sum(ll); perc = cbind(ll, aux)
```



```
r1=matrix(l, nrow=nrow(g), ncol=ncol(g), byrow=T) * g
r=r1/matrix(sqrt(a),nrow=nrow(g),ncol=ncol(g),byrow=F)
s1=matrix(l, nrow=nrow(d), ncol=ncol(d), byrow=T) * d
s=s1/matrix(sqrt(b),nrow=nrow(d),ncol=ncol(d),byrow=F)

car=matrix(matrix(a), nrow=nrow(r), ncol=ncol(r), byrow=F
    ) * r^2/matrix(l^2,nrow=nrow(r), ncol=ncol(r), byrow=T)
row.names(car)=row.names(x)
cas=matrix(matrix(b), nrow=nrow(s), ncol=ncol(s), byrow=F
   ) * s^2/matrix(l^2, nrow=nrow(s), ncol=ncol(s), byrow=T)
row.names(cas)=colnames(x)

rr=r[, 1:2]; row.names(rr)=row.names(x)
ss=s[, 1:2]; row.names(ss)=colnames(x)
```



```
# labels for journals
types   =c("va", "vb", "vc", "vd", "ve", "ff", "fg", "fh",
  "fi", "bj", "bk", "bl", "vm", "fn", "fo")

# labels for regions
regions =c("brw", "bxl", "anv", "brf", "foc", "for", "hai",
  "lig", "lim", "lux")

# plot
plot(rr, type="n", xlim=c(-1.1, 1.5), ylim=c(-1.1, 0.6),
    xlab="r_1,s_1", ylab="r_2,s_2", main="Journal Data",
    cex.axis=1.2, cex.lab=1.2, cex.main=1.6)
text(rr, types, cex=1.5, col="blue")
text(ss, regions, col="red"); abline(h=0, v=0, lwd=2)

## library(ca); plot(ca(journaux)); plot3d.ca(ca(journaux))
```


Example:

```
data(food); plot(ca(food))
plot3d.ca(ca(food))
```

Example:

```
data(carmean); plot(ca(carmean2))
plot(ca(5-carmean2))
```

Example:

```
data(uscrime); plot3d.ca(ca(uscrime[,3:9]))
?ca ##
```

Correspondence Analysis

- Correspondence analysis investigates dependencies in contingency tables.
- Correlations between row and column coordinates correspond to contributions to χ^2 statistic.
- The structure of row and column coordinates is similar to canonical variables in canonical correlation analysis.
- Plot of the row and column coordinates displays dependencies in the contingency table.
- The solution allows adding of additional (supplementary) variables that do not influence the calculation of the original coordinates.

Týden 13

Obecnější mnohorozměrná rozdělení:

- sférická a eliptická rozdělení,
- kopule.

Kvantily mnohorozměrných rozdělení:

- hloubka.

Směrová data.

Characteristic function

The characteristic function (CF) of a random vector $X \in \mathbb{R}^p$ is:

$$\varphi_X(t) = E(e^{\mathbf{i}t^\top X}) = \int e^{\mathbf{i}t^\top x} f(x)\, dx, \quad t \in \mathbb{R}^p.$$

The CF has many interesting and useful properties, e.g.:

1. The CF always exists, $\varphi_X(0) = 1$, and $|\varphi_X(t)| \leq 1$.
2. Two random vectors have the same CF if and only if they have the same distribution. If CF $\varphi_X(t)$ is absolutely integrable then $f(x) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} e^{-\mathbf{i}t^\top x} \varphi_X(t)\, dt$.
3. Random vectors X_1 and X_2 are independent if and only if $\varphi_X(t) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2)$, where $X = (X_1^\top, X_2^\top)^\top$.
4. CF of the sum of two independent random vectors X and Y is the product $\varphi_X(t)\varphi_Y(t) = \varphi_{X+Y}(t)$.

Cramér-Wold device

Theorem: (Cramér-Wold) The distribution of $X \in \mathbb{R}^p$ is completely determined by the set of all (one-dimensional) distributions of $t^\top X$ where $t \in \mathbb{R}^p$.

Proof: Let $Y = t^\top X$, then CF $E(e^{\mathbf{i}sY}) = E(e^{\mathbf{i}st^\top X})$ and this becomes the CF $\varphi_X(t) = E(e^{\mathbf{i}t^\top X})$ for $s = 1$.

Corollary: The random vector $X \sim N_p(\mu, \Sigma)$ if and only if the random variable $Y = a^\top X \sim N(a^\top \mu, a^\top \Sigma a)$ for all $a \in I\!R^p$.

Recall that we have already used this characterization to define multivariate normal distribution (see lecture 3).

Spherical distributions

Definition: A $(p \times 1)$ random vector Y is said to have a spherical distribution $S_p(\phi)$ if its characteristic function $\psi_Y(t)$ satisfies: $\psi_Y(t) = \phi(t^\top t)$ for some scalar function $\phi(.)$ (the characteristic generator of the spherical distribution). We will write $Y \sim S_p(\phi)$.

Clearly, $\varphi_{X_1}(t_1) = \varphi_X(t_1, 0, \ldots, 0)$. This implies that all marginal distributions of a spherical distribution are identical (and symmetric).

Example: The multivariate t-distribution. Let $Z \sim N_p(0, \mathcal{I}_p)$ and $S \sim \chi^2_m$ be independent. The random vector

$$Y = \sqrt{m}\,\frac{Z}{S}$$

has a multivariate t-distribution with m degrees of freedom.

Elliptical distributions

Definition: A $(p \times 1)$ random vector X has an elliptical distribution with parameters $\mu(p \times 1)$ and $\Sigma(p \times p)$ if X has the same distribution as $\mu + \mathcal{A}^\top Y$, where $Y \sim S_k(\phi)$ and $\mathcal{A}$ is a $(k \times p)$ matrix such that $\mathcal{A}^\top \mathcal{A} = \Sigma$ with $\mathrm{rank}(\Sigma) = k$. We shall write $X \sim EC_p(\mu, \Sigma, \phi)$

The elliptical distribution can be seen as an extension of $N_p(\mu, \Sigma)$.

Example: The CF of standard multinormal distribution is $\varphi_Y(t) = e^{-t^\top t/2}$ and it is spherically symmetric with the characteristic generator $\exp(-x/2)$. The CF of $X = \mu + \mathcal{A}^\top Y$ is $\varphi_X(t) = e^{it^\top \mu - t^\top \mathcal{A}^\top \mathcal{A} t/2}$ because $t^\top(\mu + \mathcal{A}^\top Y)$ has univariate normal distribution and

$$Ee^{\mathbf{i}t^\top(\mu+\mathcal{A}^\top Y)} = \varphi_{N(t^\top\mu, t^\top\mathcal{A}^\top\mathcal{A}t)}(s)|_{s=1} = e^{it^\top\mu s - t^\top\mathcal{A}^\top\mathcal{A}ts/2}|_{s=1}.$$

Elliptical distributions

The characteristic function of elliptically symmetric X is of the form

$$\psi(t) = e^{\mathbf{i}t^\top \mu}\phi(t^\top \Sigma t)$$

for a scalar function ϕ.

Marginal distributions of elliptically distributed variables are elliptical.

The assumption that the returns on all assets available for portfolio formation are jointly elliptically distributed is used in portfolio theory (multinormal distribution of returns usually does not work).

Clearly, the contours of a spherical distribution are p-dimensional spheres and contours of an elliptical distribution are p-dimensional ellipsoids (if the density exists).

Spherical and elliptical distributions

- The characteristic function is always defined and it uniquely determines the probability distribution.
- An arbitrary function $\phi : \mathbb{R}^n \to C$ is the characteristic function of some random variable if and only if ϕ is positive definite, continuous at the origin, and if $\phi(0) = 1$ (Bochner's theorem).
- Spherical distribution can be seen as a generalization of $N_p(0, \mathcal{I}_p)$, elliptical distributions generalize $N_p(\mu, \Sigma)$.
- Elliptical distributions can also be defined in terms of their density functions (if it exists): $f(x) = k \cdot g((x-\mu)'\Sigma^{-1}(x-\mu))$ for some density $g(.)$.

Copula

A copula allows a generalized representation of (complicated) dependencies between random variables (risk factors).

The basic idea is to describe the joint distribution of a random variable $X = (X_1, \ldots, X_p)^\top$ using a function $C : [0,1]^p \to [0,1]$:

$$F(x_1, \ldots, x_p) = C\left(F_1(x_1), \ldots, F_p(x_p)\right),$$

where $F_1, \ldots, F_p$ represent the marginal cumulative distributions function of the variables X_j, $j = 1, \cdots, p$.

A copula C typically depends on some "tuning parameters" determining the dependence.

Two-dimensional copula

Definition: A two-dimensional copula is a function $C : [0,1]^2 \to [0,1]$ with the following properties:

- For every $u \in [0,1]$ $C(0,u) = C(u,0) = 0$ (grounded function).
- For every $u \in [0,1]$: $C(u,1) = u$ and $C(1,u) = u$ (uniform marginals).
- For every $(u_1, u_2), (v_1, v_2) \in [0,1] \times [0,1]$ with $u_1 \le v_1$ and $u_2 \le v_2$: $C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \ge 0$ (two-increasing).

Theorem: (Sklar) Every joint distribution function $H(.)$ with marginal distributions $F_1(.)$ and $F_2(.)$ can be expressed as:

$$H(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

and the copula $C(.)$ is unique when $F_1(.)$ and $F_2(.)$ are continuous.

Example: The product copula Π: two random variables X_1 and X_2 are independent if and only if

$$H(x_1, x_2) = F_1(x_1) \cdot F_2(x_2).$$

Hence, the so-called product copula $C = \Pi$ is given by:

$$\Pi(u_1, \cdots, u_p) = \prod_{j=1}^{p} u_p.$$

Example: Gaussian or normal copula:

$$C_\rho(u_1, u_2) = \int_{-\infty}^{\Phi_1^{-1}(u_1)} \int_{-\infty}^{\Phi_2^{-1}(u_2)} \varphi_\rho(r_1, r_2) dr_2 dr_1 = \Phi_\rho\{\Phi_1^{-1}(u_1), \Phi_2^{-1}(u_2)\},$$

where φ_ρ denotes the bivariate normal density function with correlation ρ and Φ_j, $j = 1, 2$ represent the gaussian marginal distribution (GOOGLE: gaussian copula financial crisis).

Example: An important class of copulas is the Gumbel-Hougaard family:

$$C_\theta(u_1, u_2) \stackrel{\text{def}}{=} \exp\left\{-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}\right\}.$$

For $\theta = 1$ we obtain the product copula: $C_1(u_1, u_2) = \Pi(u_1, u_2) = u_1 u_2$.
For $\theta \to \infty$ we obtain the so-called minimum copula:

$$C_\theta(u_1, u_2) \longrightarrow \min(u_1, u_2) \stackrel{\text{def}}{=} M(u_1, u_2)$$

(that dominates every other copula; $M(.)$ is therefore referred to as the Fréchet-Hoeffding upper bound.

The two-dimensional function $W(u_1, u_2) \stackrel{\text{def}}{=} \max(u_1 + u_2 - 1, 0)$ satisfies $W(u_1, u_2) \le C(u_1, u_2)$ for all copulas and is called the Fréchet-Hoeffding lower bound.

Copulas

- Copulas provide a very flexible way of describing dependencies between random variables. Mathematically, a copula is a multivariate probability distribution function for which the marginal probability distribution of each variable is uniform.
- Copulas are popular in high-dimensional statistical applications as they allow to model and estimate the distribution of random vectors by estimating marginals and copulae separately.
- There are many parametric copula families available, which usually have parameters that control the strength of dependence. Archimedean copulas are defined by $\psi_\theta^{[-1]}\left(\psi_\theta(u_1)+\cdots+\psi_\theta(u_d)\right)$, where $\psi_\theta(.)$ is a generator function. This allows modeling of dependence in high dimensions with only one parameter (θ).
- More information can be found in [Nelsen, R. B. (1999). *An Introduction to Copulas*, Springer, New York.]

Multivariate quantiles

The contours (iso-density regions) for multinormal (and elliptical) distributions are ellipsoids that can be understood as multivariate generalization of quantiles. Unfortunately, defining multivariate quantiles in general is very complicated.

Notice that:

- Median and other quantiles are naturally defined for 1-dim random variable BUT definition of quantile (apart of multinormal or elliptical distribution) is not straigtforward,
- Median: measure of location, "most central" point.
- Quantiles: testing, construction of prediciton regions, boxplots...

Quantiles in more dimensions

THE AIM IS to generalize the definition of "one-dimensional" quantiles to more dimensional data sets.

The definition of quantiles in 1D uses the order of observations (but observations in more dimensions are not clearly ordered).

In order to define some kind of "ordering" we may define depth function—a measure of "how deep in the dataset" is some point.

The p-dimensional data set will be ordered *from outside to inside* instead *from left to right*.

Motivation

Example: boxplot(carmean2)

The standard definition of boxplot is based on sample quantiles (median and quartiles) that are not naturally defined in two and more dimensions.

Boxplot (defined without quantiles, by using the inside×outside ordering):

Central box	contains 1/2 of "most central" observations.
Whiskers	extend to most extreme observations by (at most) 1.5×"central box".
Outliers	are "too far away" from the centre.

Depth

The required inside×outside ordering can be based on values of some *depth function*.

Technically: for a given random vector $X \in \mathbb{R}^p$ (with distribution function F_X) depth is a function $D : \mathbb{R}^p \longrightarrow \mathbb{R}$.

Region $R(a)$ with given depth a in $\mathbb{R}^p$ is $\{x \in \mathbb{R}^p : D(x) \geq a\}$ ... the border of the region $R(a)$ is the a-depth contour (and this is the "multivariate contour").

Example: X is one-dimensional random variable with d.f. $F_X(.)$.

$$D(x) = 1 - 2|F(x) - 1/2|$$

Deepest point: $F(x) = 1/2$ (median).

Point with min. depth: $F(x) = 0$ (extremes).

Note: for a random sample $X_1, \ldots, X_n$, we define sample version of the depth function as $D(x) = 1 - 2|F_n(x) - 1/2|$.

Some desired properties of (sample) depth functions:

- Depth should not depend on the coordinate system (rotation and scale invariance).
- If a distribution is symmetric around s then s is the deepest point.
- Decreasing along rays from the deepest point.
- Vanishing at infinity, i.e., $D(x) \longrightarrow 0$ if $\|x\| \longrightarrow \infty$.
- Quasi-concavity (level sets of depth function are convex).

More details: Liu (1990), Serfling (2000).

Example: `pairs(carmean2)`; How to find deepest point in more dimensions?

Depth functions

Popular depth functions:

- Simplicial depth (Liu depth).
- Halfspace depth (Tukey).

Implementation in R: library(depth), commands: perspdepth, isodepth, depth...

Other approaches: convex hull peeling, zonoids, L1-depth, location-scale depth, and many other.

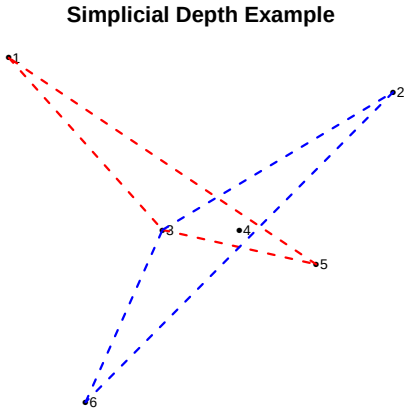
Simplicial depth

The simplicial depth (or Liu depth) of a data point x is defined as the number of convex hulls formed from all possible selections of $p+1$ points covering x (convex hull of $p+1$ points = *simplex*).

The multivariate median (the deepest point) may be defined as the point with the largest simplicial depth, i.e.,

$$x_{\text{med}} = \arg\max_i \#\{k_0, \ldots, k_p \in \{1, \ldots, n\} : x_i \in \text{hull}(x_{k_0}, \ldots, x_{k_p})\}.$$

in 1D: closed intervals given by 2 points $[x_i, x_j]$,
in 2D: triangles given by 3 points,
in 3D: "pyramids" given by 4 points etc.



→ `MVAsimdep1`

Liu depth in R

```
library(depth)
perspdepth(carmean2[,1:2],method="Liu")
d=perspdepth(carmean2[,1:2],method="Liu",output=TRUE)
contour(d)
text(carmean2[,1:2],rownames(carmean2))
```


Halfspace depth

The (sample) halfspace depth of point x (with respect to sample points $x_1, \ldots, x_n$) is defined as the minimum number of sample points on one side of a hyperplane through the point x.

In other words, minimum number of sample points in a halfspace containing the point x.

Example: 1D, points are lying on real line...

Example: 2D

```
isodepth(carmean2[,1:2],mustdith=TRUE)
text(carmean2[,1:2],rownames(carmean2))
```

Halfspace depth

Example: 3D

```
for (i in 1:nrow(carmean2)) {
  print(rownames(carmean2)[i])
  print(depth(carmean2[i,1:3],carmean2[,1:3]))
}
```

⋮

Example: 8D, almost all point are "outside" (see also Ggobi).

Bagplot

Example: Command bagplot in library(aplpack).

```
library(aplpack)
?bagplot  # what is BAG, FENCE, LOOP?

library(SMSdata)
data(carmean2)

bagplot(carmean2[,1:2]) # Service & Value
text(carmean2[,1:2],rownames(carmean2))
```


Depth

- Depth can be seen as a multivariate generalization of (empirical) quantile.
- The most popular depth functions are simplicial (Liu) depth and halfspace (Tukey) depth but many other depth functions have been proposed [D. Hlubinka: Výpravy do hlubin dat, Robust 2008, `http://www.karlin.mff.cuni.cz/~hlubinka/soubory/robust08.pdf`; D. Hlubinka: O kvantilech ve více rozměrech, Robust 2002, `http://www.statspol.cz/oldstat/robust/2002_hlubinka.pdf`].
- Most depth functions are computationally intensive.
- Using depth, it is possible to define bagplot as a two-dimensional generalization of the boxplot.

Directional data

Directional statistics is the analysis of data that are directions: these are unit vectors in a space of any number of dimensions and can be visualized as points on the surface of a hypersphere (in two- or three-dimensional spaces we have points on the circumference of a circle or on the surface of a sphere, i.e. circular and spherical data).

Directional statistics differs from 'usual linear' statistics because of the specific structure of its sample spaces. As hyperspheres have different characteristics than general Euclidean spaces, standard linear methods for analyzing data cannot be used and special directional methods are required.

References: [Mardia, K. V. and Jupp, P. E. (2000). Directional statistics. 2nd ed., Wiley Series in Probability and Statistics. Wiley, Chichester] or [Malá, O. C. (2012) Fisherovo-Binghamovo rozdělení, bakalářská práce, MFF UK.]

Distribution function and density

Let θ be a random angle. Its distribution function F is given by

$$F(\theta) = P(0 < \theta \leq \theta), \quad 0 < \theta \leq 2\pi$$

and

$$F(\theta + 2\pi) - F(\theta) = 1, \quad -\infty < \theta \leq \infty.$$

Let the distribution function F of random angle θ be absolutely continuous. Then for the probability density function f of random angle θ, the following holds:

1. $f(\theta) \geq 0$ almost everywhere on $(-\infty, \infty)$,
2. $f(\theta + 2\pi) = f(\theta)$ almost everywhere on $(-\infty, \infty)$,
3. $\int_0^{0+2\pi} f(\theta)d\theta = 1$ and $\int_x^{x+2\pi} f(\theta)d\theta = 1$.

Summary statistics

Let's have a set of independently observed directions in the plane that are represented by unit vectors $V_1, \ldots, V_n$ (these correspond to unique angles $\theta_1, \ldots, \theta_n$ and to unique points on the unit circle $X_1, \ldots, X_n$.)

By summing these unit vectors and taking their mean, we obtain the mean resultant vector $\overline{V} = \sum V_i/n$ and the endpoint $\overline{X}$ of the vector $\overline{V}$ represents the 'centre of mass' (if points $X_1, \ldots, X_n$ have equal masses).

Points X_j have Cartesian coordinates $(\cos\theta_j, \sin\theta_j)$, i.e., the centre of mass $\overline{X}$ has Cartesian coordinates $(\overline{C}, \overline{S})$, where $\overline{C} = \sum \cos(\theta_j)/n$ and $\overline{S} = \sum \sin(\theta_j)/n$.

Denoting $(\overline{R}, \overline{\theta})$ the polar coordinates of $\overline{X}$, we obtain:

Definition: The sample mean resultant length $\overline{R} \geq 0$ is given by

$$\overline{R} = \|V\| = \sqrt{\overline{C}^2 + \overline{S}^2}.$$

If $\overline{R} > 0$, the sample mean direction $\overline{\theta}$ is defined as follows:

$$\overline{\theta} = \arctan^*(\overline{S}/\overline{C}) = \begin{cases} \arctan(\overline{S}/\overline{C}), & \text{if } \overline{C} > 0, \overline{S} \geq 0, \\ \pi/2 & \text{if } \overline{C} = 0, \overline{S} > 0, \\ \arctan(\overline{S}/\overline{C}) + \pi, & \text{if } \overline{C} < 0, \\ \arctan(\overline{S}/\overline{C}) + 2\pi, & \text{if } \overline{C} \geq 0, \overline{S} < 0, \\ \text{undefined} & \text{if } \overline{C} = 0, \overline{S} = 0. \end{cases}$$

Definition: The sample circular variance V is defined as

$$V = 1 - \overline{R}, \quad 0 \leq V \leq 1.$$

Standard circular distribution

The density of von Mises distribution is:

$$f(x; \mu, \kappa) = c_0(\kappa) \exp\{\kappa \mu^\top x\}, \quad x \in S^1,$$

where $c_0(\kappa)$ is constant.

Notice that

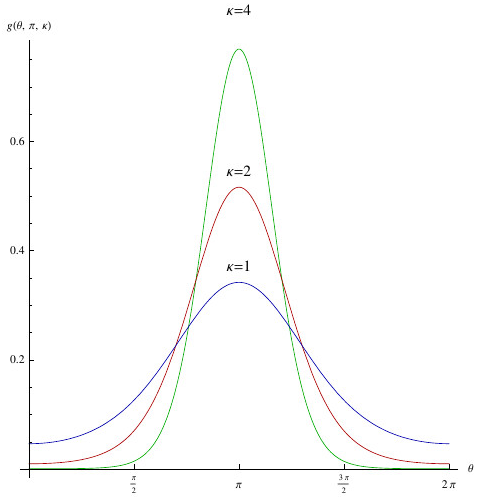
$$\mu^\top x = (\cos\mu, \sin\mu)(\cos\theta, \sin\theta)^\top = (\cos\mu\cos\theta + \sin\mu\sin\theta) = \cos(\theta - \mu).$$

and the probability density function of random angle θ is

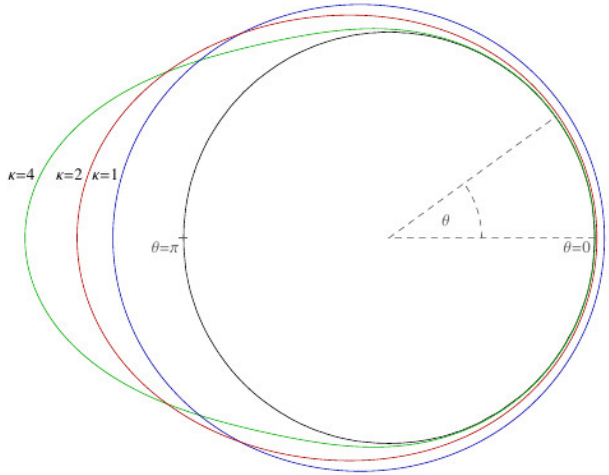
$$g(\theta; \mu, \kappa) = c_0(\kappa) \exp\{\kappa \cos(\theta - \mu)\}, \quad 0 < \theta \leq 2\pi.$$

We only used a different way of representation corresponding to 'unrolling the circle'.

von Mises distribution (for angle θ)



von Mises distribution (on the unit circle)



Directional statistics

- See [Mardia, K. V. and Jupp, P. E. (2000). Directional statistics. 2nd ed., Wiley Series in Probability and Statistics. Wiley, Chichester] or [Malá, O. C. (2012) Fisherovo-Binghamovo rozdělení, bakalářská práce, MFF UK] for more information.
- Directional (or axial) data are encountered in various fields: geology, meteorology, astronomy, geography, medicine and others.
- von Mises distribution can be generalized to more dimensions (Fisher-Bingham distribution, von Mises-Fisher distribution, etc.)
- One can also consider distributions defined on more general 'surfaces' (manifolds).

Týden 13

Další zajímavé metody:

- jádrové odhady hustoty,
- projection pursuit.

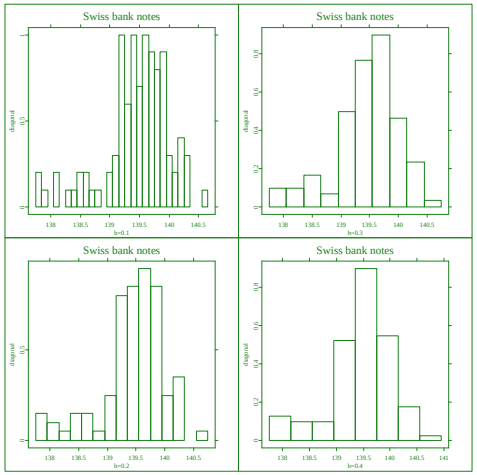
Histograms

The histogram counts relative frequencies of observations x_i falling into predefined bins:

$$\widehat{f}_h(x) = n^{-1}h^{-1}\sum_{j\in\mathbb{Z}}\sum_{i=1}^{n} \mathrm{I}\{x_i \in B_j(x_0,h)\}\,\mathrm{I}\{x \in B_j(x_0,h)\}$$

- the histogram is a simple estimator of a probability density,
- h is a smoothing parameter and controls the width of the histogram bins.

Example: Diagonal of forged bank notes. Histograms with $h = 0.1$ (upper left), $h = 0.2$ (lower left), $h = 0.3$ (upper right), $h = 0.4$ (lower right).



Kernel density estimators

Kernel density estimator is a natural generalization of a histogram (by shifting the "bin", we obtain smooth estimator of the underlying probability density).

Assume we have n independent observations $x_1, \ldots, x_n$ from the random variable X. The kernel density estimator $\widehat{f}_h(x)$ for the estimation of the density value $f(x)$ at point x is defined as

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

where $K(.)$ denotes a kernel function and h the bandwidth.

Multivariate KDEs

The kernel density estimator can be generalized to the multivariate case in a straightforward way.

Suppose we have observations $x_1, \ldots, x_n$ where each of the observations is a d-dimensional vector $x_i = (x_{i1}, \ldots, x_{id})^T$. The multivariate kernel density estimator at a point $x = (x_1, \ldots, x_d)^T$ is defined as

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1 \ldots h_d} K\left(\frac{x_{i1} - x_1}{h_1}, \ldots, \frac{x_{id} - x_d}{h_d}\right),$$

where K is a multivariate kernel funcion and h is a vector of bandwidths $h = (h_1, \ldots, h_d)^\top$.

It can be shown that the optimal MISE is $O(n^{-4/(d+4)})$ (curse of dimensionality).

```
library(sm); library(MSES); data(athletic)

# univariate kernel density estimator
plot(density(athletic[,"100m"]))
plot(density(athletic[,"Marathon"]))

## bivariate kernel density estimator
library(MASS)
plot(athletic[,"Marathon"],athletic[,"100m"])
d1=kde2d(athletic[,"Marathon"],athletic[,"100m"])

image(d1, zlim = c(0, 0.13))
persp(d1, phi = 30, theta = 20, d = 5)
contour(d1)
# add original points
points(athletic[,"Marathon"],athletic[,"100m"])

identify(athletic[,"Marathon"],athletic[,"100m"],
                              label=row.names(athletic))
```


Kernel density estimators

- KDEs are sometimes introduced as "average shifted histograms" (ASH).
- High-dimensional KDEs suffer from "curse of dimensionality" because the optimal MISE is of order $O(n^{-4/(d+4)})$, where d denotes the dimension.
- The various implementations of KDEs in R are not mutually compatible (for example, the bandwidth parameter used by one R function typically does not have exactly the same meaning in other R function).
- One should consider dimension reduction techniques before calculating high-dimensional KDEs.

Projection pursuit

Projection pursuit searches for interesting directions in a p-dimensional data set by maximizing a chosen index.

Exploratory projection pursuit: look for interesting linear combinations—"interestingness" is usually defined by some measure (index) of non-normality.

Projection pursuit regression: the goal is to estimate regression function $m(x) = E(Y|x)$ using approximating function $\widehat{f}(x) = \sum \widehat{g}_k(\Lambda_k^\top x)$ (obviously, lower dimensional projections defined by Λ_k improve statistical properties of the nonparametric regression estimator).

Exploratory projection pursuit

Given p-dimensional random vector X with zero mean (and typically with unit variance, i.e., $\mathrm{Var}(X) = \mathcal{I}_p$), we try to find $\alpha \in I\!R^p$ such that $\alpha^\top X$ is "interesting".

Interestingness of projections $\alpha^\top X$ is measured by index $I(\alpha)$.

Example: PCA: $I(\alpha) = \mathrm{var}(\alpha^\top X)$ works only if the data set is not sphered.

In practice, we have the data matrix $\mathcal{X}$ and we optimize the (sample) projection pursuit index numerically.

Friedman and Tukey index

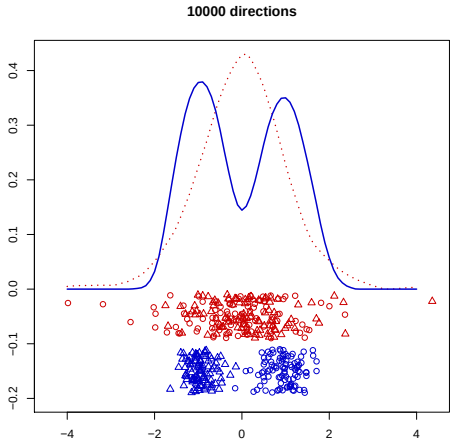
Let $\widehat{f}_{h,\alpha}(z)$ denote the kernel density estimator of the pdf of the projection $Z = \alpha^\top X$, where h denotes the bandwidth.

Friedman and Tukey (1974) proposed the index:

$$I_{FT,h}(\alpha) = n^{-1} \sum_{i=1}^{n} \widehat{f}_{h,\alpha}(\alpha^\top X_i)$$

that can be rewritten as $I_{FT,h}(\alpha) = \int \widehat{f}_{h,\alpha}(z) dF_N(z)$ (i.e., it estimates $\int f(z) dF(z) = \int f^2(z) dz$) leading to the maximization of $\int f^2(z) dz$.

The Friedman-Tukey index is minimal for a parabolic density and, by its maximization, we search for a distribution that is as far from the parabolic density as possible.



The least and the most informative from 10000 randomly chosen directions (FT index) for Swiss bank notes → `SMSeppbank`.

Entropy index

An alternative approach is based on the (minus) entropy measure $\int f(z) \log f(z) dz$ leading to the entropy index:

$$I_{E,h}(\alpha) = n^{-1} \sum_{i=1}^{n} \log\{\widehat{f}_{h,\alpha}(\alpha^\top X_i)\}$$

that can be interpreted as an estimator of minus entropy $\int f(z) \log f(z) dz$.

The index is minimal for normal distribution and maximization of $I_{E,h}(\alpha)$ leads to non-normal projections.

Jones and Sibson index

Jones and Sibson (1987) suggested to approximate the entropy index by a moment-based index:

$$I_{JS}(\alpha) = \{\kappa_3^2(\alpha^\top X) + \kappa_4^2(\alpha^\top X)/4\}/12,$$

where $\kappa_3(\alpha^\top X) = E\{(\alpha^\top X)^3\}$ and $\kappa_4(\alpha^\top X) = E\{(\alpha^\top X)^4\} - 3$ are cumulants of $\alpha^\top X$ (skewness and kurtosis).

The maximization of $I_{JS}(\alpha)$ also leads to the least-normal view of the data set.

Computational aspects

The optimal projection $\alpha \in I\!R^p$ can be found by standard (iterative) optimization routines.

The optimization task is not very simple because the parameter α is p-dimensional and the function $I(\alpha)$ has many local maxima.

In practice, one is interested in finding optimal one- and two-dimensional projections.

It is recommended to use various starting points in order to verify the stability of the result. Often, the optimization of α is used to define a **guided** tour through the data set.

White noise analysis

White noise projections that are most similar to white noise are identified and discarded while the remaining informative projections are used to look for interesting relationships [Hui and Lindsay, 2010, Projection pursuit via white noise matrices, Sankhya B 72(2), 123-–153.]

The White Noise Analysis (WNA) is based on the eigen-analysis of the standardized Fisher information matrix for the square transformed density estimated by the kernel method.

WNA is computationally simpler than the classical Projection Pursuit searching for low-dimensional least-normal projections.

Exercise: Swiss bank notes

1. `library(SMSdata); data(bank2)`
2. sphering (Mahalanobis transformation),
3. generate randomly N directions $\alpha_1, \ldots, \alpha_N$
4. calculate the value $I(\alpha_i)$ for $i = 1, \ldots, N$
5. plot kde of the directions that both maximize and minimize the chosen index,
6. compare the result obtained for PCA index with standard PCA analysis (this will work only without sphering),
7. compare least and most informative projections obtained by JS and FT index,
8. try to find the optimal direction using numerical optimization in R (`optim()`), compare results obtained by different algorithms (Nelder-Mead, BFGS, ...).

Ggobi: 1D and 2D guided tour

Guided tour through a multivariate data set is a sequence of low-dimensional projections that improve the chosen index.

```
library(SMSdata)
data(bank2)

library(rggobi) # using ggobi is easy if this works
ggobi(bank2)

write.csv(bank2,file="bank2.dat")
# START GGOBI AND LOAD DATA SET FROM CSV FILE
```

R: tourr

Guided and grand tours work similarly as in Ggobi (same authors) but R does not allow interaction.

```
library(SMSdata)
data(bank2)

library(tourr)
animate(bank2, guided_tour(index_f=holes), display_xy()
                                          ,sphere=FALSE)
animate(bank2, guided_tour(index_f=cmass), display_xy())
```


R: tourr

```
# tourr SHOULD work also with other type of graphics

animate_dist(bank2[95:106,],guided_tour(index_f=holes))
animate_image(bank2[95:106,],guided_tour(index_f=holes))
animate_pcp(bank2[95:106,],guided_tour(index_f=holes))
animate_scatmat(bank2[95:106,],guided_tour(index_f=holes))
animate_faces(bank2[95:106,],guided_tour(index_f=holes))
animate_stars(bank2[95:106,],guided_tour(index_f=holes))
animate_stereo(bank2[95:106,],guided_tour(index_f=holes))
animate_trails(bank2[95:106,],guided_tour(index_f=holes))
```


Indices in Ggobi and tourr

Holes

$$I_{\text{Holes}}(\alpha) = \frac{1 - \frac{1}{n}\exp(-z_i z_i^\top/2)}{1 - \exp(-p/2)},$$

where z_i is the i-th row of $\mathcal{Z} = \mathcal{X}\alpha$ (the index works also for more dimensional projections).

Central mass

$$I_{\text{CM}}(\alpha) = \frac{\frac{1}{n}\exp(-z_i z_i^\top/2) - \exp(-p/2)}{1 - \exp(-p/2)},$$

is basically the opposite of I_{Holes}.

Both indices are based on $\frac{1}{n}\exp(-z_i z_i^\top/2) = \int \exp(-z_i z_i^\top/2) dF_n(z)$ estimating $E\exp(-Z^\top Z/2)$.

Central mass and Holes

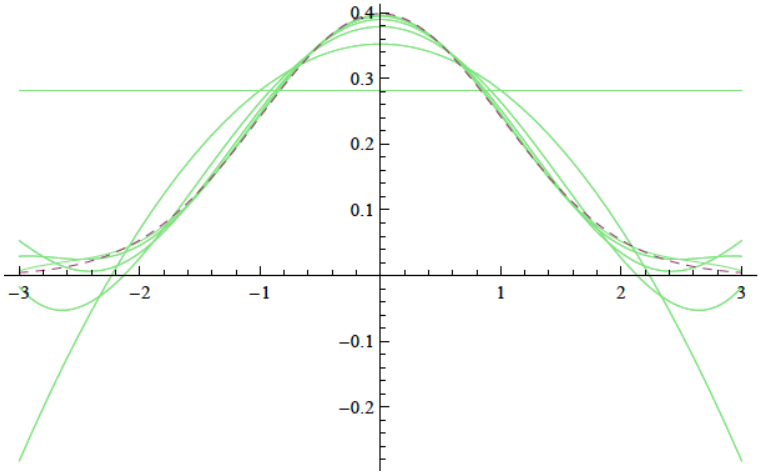
Cook, Buja, Cabrera (1993) Projection Pursuit Indexes Based on Orthonormal Function Expansions. Journal of Computational and Graphical Statistics 2(3), 225–250.

The derivation of both indeces is based on Fourier expansion of density function:

$$f(x) = \sum_{i=0}^{\infty} a_i p_i(x),$$

where $p_i(x)$ are (standardized) orthonormal polynomials with weight function $w(x)$ and $a_i = \langle f, p_i \rangle = \int f(x) p_i(x) w(x) dx$ are Fourier coefficients.

Fourier approximation of normal density



Fourier coefficients can be rewritten as expectation

$$a_i = \langle f, p_i \rangle = \int f(x) p_i(x) w(x) dx = \int p_i(x) w(x) dF(x) = E\{p_i(X) w(X)\}$$

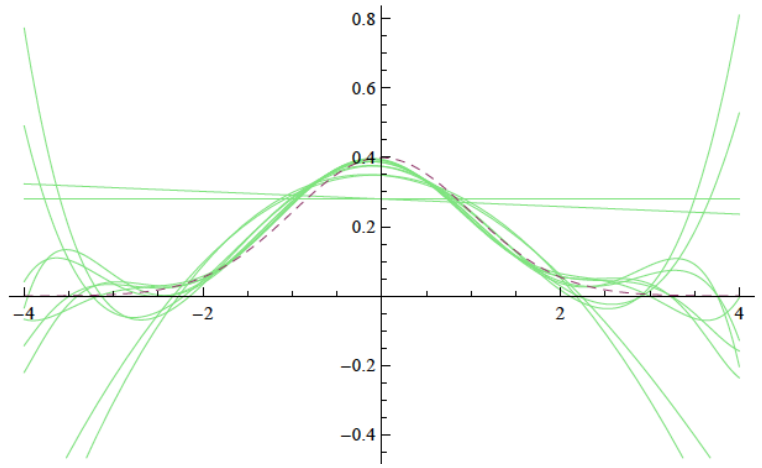
that can be estimated from random sample $X_1, \ldots, X_n$ by sample mean

$$\widehat{a}_i = \frac{1}{n} \sum_{j=1}^{n} p_i(X_j) w(X_j).$$

In practice, the density can be approximated by finite sum

$$\widehat{f}(x) = \sum_{i=0}^{M} \widehat{a}_i p_i(x).$$

Fourier approximation from 100 observations



Distance from Normal distribution

Cook, Buja & Cabrera (1993) define natural Hermite index

$$I_N = \int_{\boldsymbol{R}} \{f(x) - \phi(x)\}^2 \phi(x) dx$$

as a measure of dissimilarity of probability densities $f(x)$ and $\phi(x)$.

It is easy to show that $I_N = \sum (a_i - b_i)^2$, where b_i are (known) Fourier coefficients of $\phi(x)$.

The sample version of I_N is naturally defined as:

$$\widehat{I}_{N,M} = \sum_{i=0}^{M} (\widehat{a}_i - b_i)^2,$$

where $\widehat{a}_i = \sum_{j=1}^{n} p_i(X_j) w(X_j)/n$.

Distance from Normal distribution

Cook, Buja & Cabrera (1993) investigate the sample natural Hermite index for $M = 0$, i.e., $\widehat{I}_{N,0} = (\widehat{a}_0 - b_0)^2$.

Clearly, the quadratic function $(\widehat{a}_0 - b_0)^2$ achieves its minimum $\widehat{a}_0 - b_0$ and is maximized by extreme values of a_0.

Cook, Buja & Cabrera (1993) show that, in a family of distributions with mean zero and variance at most one, a_0 is minimized by the *central hole* distribution:

$$P(X = 1) = 0.5, \quad P(X = -1) = 0.5$$

and a_0 is maximized by the *central mass* distribution:

$$P(X = 0) = 1.$$

Central mass and holes

The central mass index in Ggobi looks for the rotation α maximizing $a_0(\alpha)$.

The holes index in Ggobi looks for the rotation α maximizing $-a_0(\alpha)$.

Distributions with very small or very large a_0 should have large distance (natural Hermite index) from Normal distribution.

Switching repeatedly between maximization of these two indeces leads to informative displays of the data set.

Example: Swiss bank notes in Ggobi and R (tourr).

Interpretation

Original data matrix $\mathcal{X}$.

Sphered data matrix $\mathcal{Y} = (\mathcal{X} - \mathbf{1}_n \bar{x}^\top)\mathcal{S}^{-1/2}$.

Interesting linear combinations are:

$$\mathcal{Y}\alpha = (\mathcal{X} - \mathbf{1}_n \bar{x}^\top)\mathcal{S}^{-1/2}\alpha = \mathcal{X}\mathcal{S}^{-1/2}\alpha + \text{const} = \mathcal{X}\alpha_{\mathcal{X}} + \text{const}.$$

Visual inference

Chowdhury, Cook, Hofmann, Majumder, Lee & Toth (2015) Using visual statistical inference to better understand random class separations in high dimensions, low sample size data, Computational Statistics 30: 293–316.

[The paper can be found using scholar.google.com.]

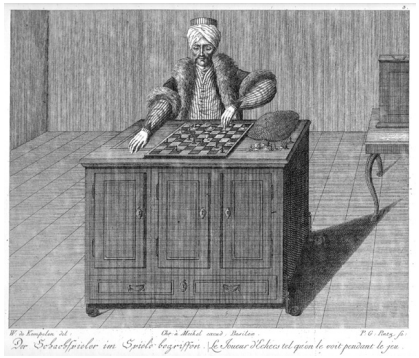
The problem: lower dimensional projections (especially based on LDA) can be misleading (see Figure 1).

Example:

```
d=data.frame(matrix(rnorm(150),ncol=10))
animate(d,guided_tour(index_f=holes),display_xy(),sphere=TRUE)
```

Visual inference

Proposed solution: use visual statistical inference via Amazon's Mechanical Turk (the original "lived" from 1770–1854).



Exercise: simulated data set

1. generate three independent samples with the same p-variate distribution (using, e.g., `rmvnorm(mvtnorm)`),
2. use function `animate(tourr)` to find interesting projections (preferably using the `lda_pp` index),
3. plot the resulting projections (take care about scaling) denoting the three groups by different symbol–can you see some differences?
4. repeat the simulation both for small and high dimension.

Týden 14

Gentle introduction:

- kernel regression estimators,
- additive models,
- projection pursuit regression.

Sliced inverse regression:

- kernel regression estimators,
- additive models and projection pursuit regression,
- inverse regression curve,
- SIR,
- SIR II.

Kernel regression estimators

Suppose that we have independent observations $Y_1, \ldots, Y_n$ and the explanatory variable $X_1, \ldots, X_n$. The Nadaraya-Watson kernel regression estimator is defined as:

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} = \frac{1}{n}\sum_{i=1}^n W_{hi}(x) Y_i.$$

It can be shown that the asymptotic MSE is:

$$\text{AMSE}(n, h) = \frac{1}{nh} C_1 + h^4 C_2,$$

where C_1 and C_2 are constants depending on the kernel function, the (derivatives of) the regression function and the density of X. Using the optimal bandwidth $h = C_3 n^{-1/5}$, AMSE is of order $O(n^{-4/5})$.

Kernel regression estimates in R:

Typically 1 response and 1 or 2 explanatory variables.

1D ksmooth(), locpoly(KernSmooth)
2D sm.regression(sm)

```
library(sm); library(MSES); data(athletic)

# most simple univariate kernel regression estimates
#        (better functions exist in other libraries)
plot(athletic[,"Marathon"],athletic[,"100m"])
lines(ksmooth(athletic[,"Marathon"],athletic[,"100m"],
               kernel="normal", bandwidth=20),col="red",lwd=2)

library(KernSmooth)
plot(athletic[,"Marathon"],athletic[,"100m"])
lines(locpoly(athletic[,"Marathon"],athletic[,"100m"],bandwidth=10),
                                                col="red",lwd=2)
```



```
sm.regression(athletic[,"Marathon"],athletic[,"100m"])

## bivariate kernel density estimator
library(MASS)
plot(athletic[,"Marathon"],athletic[,"100m"])
d1=kde2d(athletic[,"Marathon"],athletic[,"100m"])

image(d1, zlim = c(0, 0.13))
persp(d1, phi = 30, theta = 20, d = 5)
contour(d1)
# add original points
points(athletic[,"Marathon"],athletic[,"100m"])
# add kernel regression line
lines(ksmooth(athletic[,"Marathon"],athletic[,"100m"],
              kernel="normal",bandwidth=20),col="red",lwd=2)
```



```
## bivariate kernel regression

library(sm)

sm.regression(athletic[,c("Marathon","400m")],athletic[,"100m"])

sm.regression(athletic[,c("Marathon","400m")],athletic[,"100m"],
                                              display="image")
```

The asymptotic properties of the kernel regression estimator are bad for high-dimensional explanatory variable (curse of dimensionality). Moreover, it is difficult to plot the resulting estimator for $p > 2$.

Curse of dimensionality (from Wikipedia)

One way to illustrate the "vastness" of high-dimensional Euclidean space is to compare the proportion of an inscribed hypersphere with radius r and dimension d, to that of a hypercube with edges of length $2r$. The volume of such a sphere is: $\frac{2r^d\pi^{d/2}}{d\,\Gamma(d/2)}$. The volume of the cube would be: $(2r)^d$. As the dimension d of the space increases, the hypersphere becomes an insignificant volume relative to that of the hypercube. This can clearly be seen by comparing the proportions as the dimension d goes to infinity: $\frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0$ as $d \to \infty$. Furthermore, the distance between the center and the corners is $r\sqrt{d}$, which increases without bound for fixed r.

In this sense, nearly all of the high-dimensional space is "far away" from the centre. To put it another way, the high-dimensional unit hypercube can be said to consist almost entirely of the "corners" of the hypercube, with almost no "middle".

Additive model

In order to avoid the curse of dimensionality, it can be useful to consider additive model (AM) with response p-dimensional explanatory variable X:

$$E(Y|X=x) = \sum_{j=1}^{p} f_j(x_j) + c,$$

where $c = E(Y)$ and the (univariate) additive components are centered, i.e., $E\{f_j(X_j)\} = 0$ for $1 \leq j \leq p$.

The components of the additive model (and its various generalizations) are usually estimated by iterative algorithms (backfitting).

Projection pursuit regression

Projection pursuit regression [Friedman, J.H. and Stuetzle, W. (1981) Projection Pursuit Regression. Journal of the American Statistical Association, 76, 817–823]:

$$E(Y|X=x) = \sum_{j=1}^{r} f_j(\beta_j^\top x) + c,$$

applies the additive model on projections of explanatory variables, i.e., it reduces the dimensionality of the space of explanatory variables (keeping in mind that we model the conditional expectation of Y).

Implementation in R: function `ppr()` in library `stats`.

Sliced inverse regression

Sliced Inverse Regression (SIR) is a dimension reduction technique that can be described as a generalization of projection pursuit regression.

The idea is to find EDR-directions (i.e., projections of explanatory variables) suitable for nonparametric regression estimator for the response.

Given a response variable Y and a (random) vector $X \in \mathbb{R}^p$ of explanatory variables, SIR is based on the model:

$$Y = m(\beta_1^\top X, \ldots, \beta_k^\top X, \varepsilon),$$

where $\beta_1, \ldots, \beta_k$ are unknown projection vectors

Centered inverse regression curve

Recall that $Y = m(\beta_1^\top X, \ldots, \beta_k^\top X, \varepsilon)$.

According to Theorem 20.1 in [Härdle and Simar, Applied Multivariate Statistical Analysis, 4th edition] we have that: "Under some assumptions, the (p-dimensional) centered inverse regression curve $E(X|Y=y) - EX$ lies in the linear subspace spanned by $\Sigma\beta_i$, $i = 1, \ldots, k$, $\Sigma = \mathrm{Var}\, X$."

It follows that for $Z = \Sigma^{-1/2}(X - EX)$, the standardized inverse regression curve $m_1(y) = E(Z|Y=y)$ lies in a linear subspace spanned by $\eta_i = \Sigma^{1/2}\beta_i$.

The idea of SIR algorithm is to generate points lying on the inverse regression curve and then estimate the linear subspace...

SIR Algorithm (part 1)

The algorithm to estimate the EDR-directions via SIR is as follows:

- Standardize x:
$$z_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x}).$$
- Divide the range of y_i in S non-overlapping intervals (slices) $Sttk_s$, $s = 1, \ldots, S$. n_s denotes the number of observations within slice $Sttk_s$ and I_{Sttk_s} is the indicator function for this slice ($n_s = \sum_{i=1}^{n} I_{Sttk_s}(y_i)$):
- Compute the mean of z_i over all slices. This is a crude estimate $\hat{m}_1$ for the *inverse regression curve* m_1:
$$\bar{z}_s = \frac{1}{n_s} \sum_{i=1}^{n} z_i \, I_{Sttk_s}(y_i).$$

SIR Algorithm (part 2)

- Calculate the estimate for $\mathrm{Var}\{m_1(y)\}$:
$$\widehat{V} = n^{-1} \sum_{s=1}^{S} n_s \bar{z}_s \bar{z}_s^\top.$$
- Identify the eigenvalues $\hat{\lambda}_i$ and eigenvectors $\hat{\eta}_i$ of $\widehat{V}$.
- Transform the standardized EDR-directions $\hat{\eta}_i$ back to the original scale. Now the estimates for the EDR-directions are given by
$$\hat{\beta}_i = \hat{\Sigma}^{-1/2} \hat{\eta}_i.$$

Simulated data set

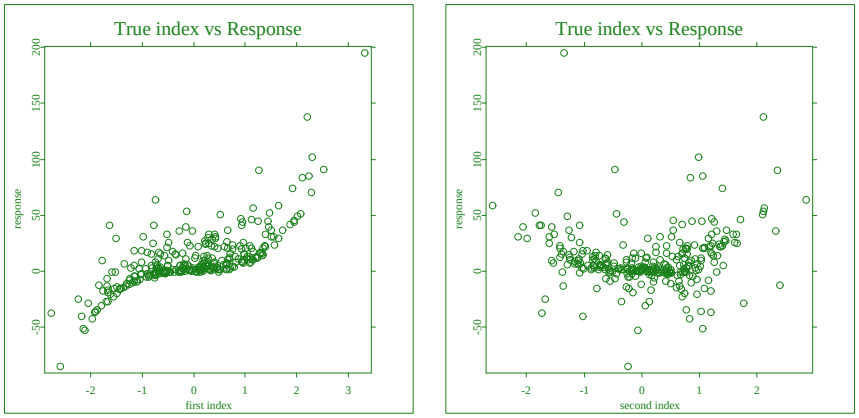
Example: Let us investigate data simulated from the model

$$y_i = \beta_1^\top x_i + (\beta_1^\top x_i)^3 + 4(\beta_2 x_i)^2 + \varepsilon_i$$

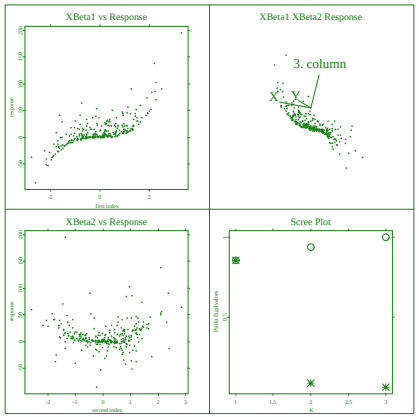
with $\beta_1 = (1, 1, 1)^\top$, $\beta_2 = (1, -1, -1)^\top$.

Looking at the data, it is difficult to find the underlying structure (the surface in 3D plot).

→ `MVAsirdata`

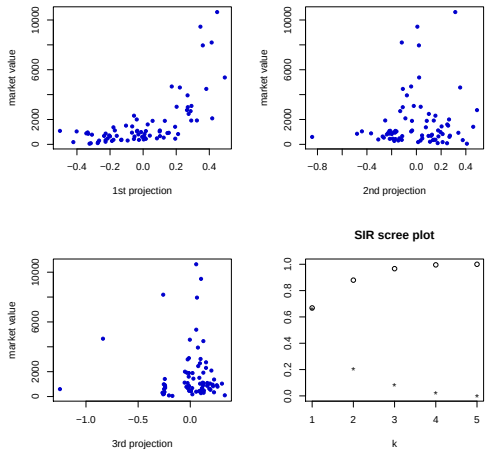


Plot of the true response versus the true indices. The monotonic and the convex shapes can be clearly seen → `MVAsirdata`



SIR algorithm works quite well (although the IR curve may not span the entire EDR space).

US companies



EDR directions for US companies (for market values).

SIR II

In some situations SIR does not find EDR directions because the inverse regression curve does not have to span the entire EDR space.

Example: Suppose that $(X_1, X_2)^\top \sim N(0, \mathcal{I}_2)$ and $Y = X_2^2$.
Notice that the EDR space is spanned by $\beta_1 = (0, 1)^\top$ and the IR curve is $E(X_1|y) = E(X_2|y) = 0$.

SIR II algorithm uses the (inverse) conditional variance $\mathrm{Var}(X|y)$ instead of the inverse regression curve. In practice, it is recommended to use SIR and SIR II jointly.

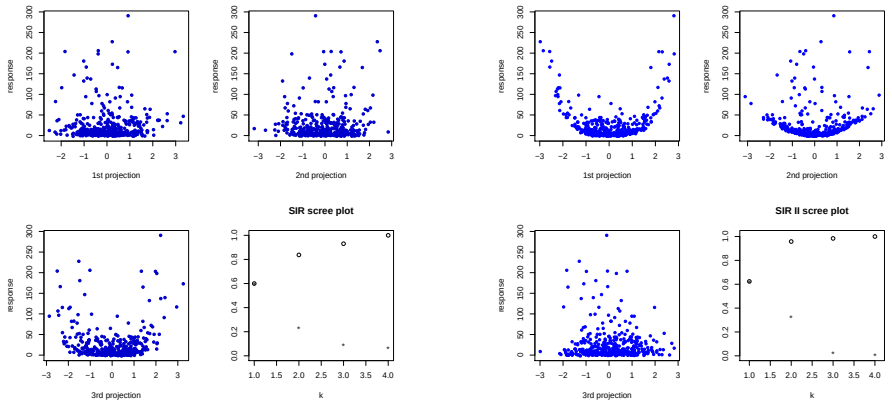
Simulated example

Example: Let us simulate a data set with $X \sim N_4(0, \mathcal{I}_4)$, $Y = (X_1 + 3X_2)^2 + (X_3 - X_4)^4 + \varepsilon$ and $\varepsilon \sim N(0, 1)$ and use the SIR and SIR II technique to find the EDR directions.

The true response variable depends on the explanatory variables nonlinearly through the linear combinations $X\beta_1 = X_1 + 3X_2$ and $X\beta_2 = X_3 - X_4$, where $\beta_1 = (1, 3, 0, 0)^\top$ and $\beta_2 = (0, 0, 3, -4)^\top$.

We simulate altogether 200 observations.

→ `SMSsir2simu`



SIR and SIR II applied on the simulated data set. Screeplot and scatterplots of first three indices against the response. → `SMSsir2simu`

Modifications

SAVE The algorithm *sliced average variance estimates* is based on the conditional variance matrix (similarly as SIR II).

pHd The method of *principal Hessian directions* is based on the Hessian matrix $E\{(Z - EZ)(X - EX)(X - EX)^\top\}$, where the vector Z is given either by the response Y or by the linear model residuals.

R library `dr`:

```
method: This character string specifies the method of
        fitting.  The options include "sir", "save",
        "phdy", "phdres" and "ire".
```


SIR

- SIR serves as dimension reduction tool for regression problems.
- Inverse regression helps to avoid the *curse of dimensionality*.
- The dimension reduction can be conducted without estimation of the regression function $y = m(x)$.
- SIR searches for the eefective dimension reduction (EDR) by computing the inverse regression IR.
- SIR II bases the EDR on computing the inverse conditional variance.
- In certain circumstances, SIR might miss EDR directions that are found by SIR II.

Závěr

Opakování a shrnutí:

- shrnutí,
- informace o zkoušce.

Summary

Multivariate distributions:

- random vector and its characteristics,
- multinormal, spherical and elliptical distributions, copulas.

Estimation and testing: maximum likelihood techniques.

Analysis of multivariate data:

- summary statistics, principal components,
- factor analysis, canonical correlations,
- discriminant analysis, cluster analysis,
- correspondence analysis, projection pursuit,
- projection pursuit regression, SIR.