Multivariate Statistics

Course Notes

Univ.-Prof. Dipl.-Ing. Dr.techn.
Peter Filzmoser

TU Wien
Institute of Statistics and Mathematical Methods in Economics

Vienna, October 2024

Preface

Multivariate statistical analysis is referring to the analysis of observations which have been observed simultaneously on several variables. This is in contrast to univariate statistical analysis, where we only have observations of a single variable, and we analyze the statistical behavior of these univariate data. In order to analyze multivariate data, we need appropriate tools, called multivariate statistical methods. Multivariate statistics is thus the extension of univariate statistics to more than one dimension.

An example of multivariate data are the results of student exams in different subjects. The observations are the students, and the variables are the different subjects. The essential difference to univariate analysis is that we are not (only) interested in the distribution of the single variables, but we rather want to analyze the joint distribution of all the investigated variables. This means that we could be interested in correlations among the variables, or in a grouping structure of the students in the mutivariate sense.

Most phenomena in real world can be better characterized by using multivariate rather than only univariate information. Statistical models for regression or classification will in general be more accurate if multivariate information is considered. This refers to the complexity of problems, which is characterized by a multitude of underlying processes.

In general, the goals of multivariate statistical analyses are:

- simplifying complex phenomena to interpretable relationships
- sorting and grouping
- investigation of dependencies among the variables
- prediction
- hypothesis testing

The development of the theory of multivariate analysis started in the early 20^{th} century. With the development of computers, the possibilities for multivariate statistical methods increased, as well as the complexity of the models and algorithms. However, even the most complex models are still “wrong”, since natural phenomena are usually much more complex – but hopefully these models are useful in practice.

Contents

Chapter 1

Introduction

1.1 Simple graphical techniques

Visualization is usually simple to do and still very helpful in analyzing data. Even when analyzing multivariate data, graphical displays exist which provide insight into the data structure. We would always recommend to first consult graphics and try to visualize the data before statistical models are fit, since usually the models have some prerequisits (normal distribution, etc.). Moreover, visualization might help to discover data artifacts such as outliers, which could spoil the statistical model estimation.

A well known display in two dimensions if the *scatter plot*. Two variables are displayed, and the observations are presented in this coordinate system.

Example 1.1.1 *We consider financial data of the biggest publishing companies, as they have been published in an article of the magazine "Forbes" on April 30, 1990. In particular, the variable x refers to the number of employees, and variable y to the profit per employee. Figure 1.1 shows a scatter plot of the data. We can see that "Dun & Bradstreet", the company with the highest number of employees, shows an average value for profit per employee, and thus it deviates somwhow from the data majority. "Time Warner" has an average value for the number of employees, but a very low profit per employee (even negative).*

The empirical correlation coefficient between x and y is -0.39. If we omit "Dun & Bradstreet", the correlation coefficient descreases to -0.56. By additionally omitting "Time Warner", the correlation coefficient is -0.50. Thus, this example demonstrates that deviating observations can have a strong influence on statistical estimators.

Visually, one could identify such deviating observations by a two-dimensional extension of the boxplot, a so-called "bagplot" (Rousseeuw et al., 1999). Figure 1.2 shows these data in a bagplot.

In analogy to the univariate boxplot, a robust mean, the region of the innermost 50% of the data (dark gray), a region separating regular data points from outliers (light gray), and outliers are presented. Indeed, the two atypical data points are identified as outliers. The shape of the dark gray region ("bag") also represents the relationship between the two variables.

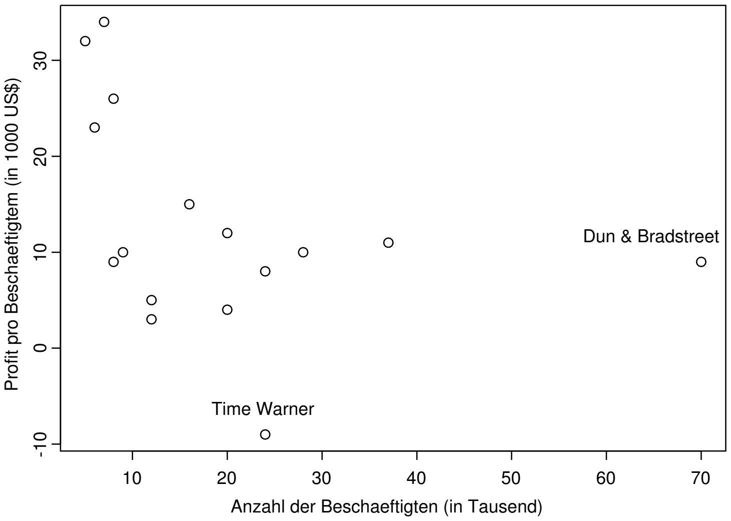


Figure 1.1: Scatter plot of the 2-dimensional financial data

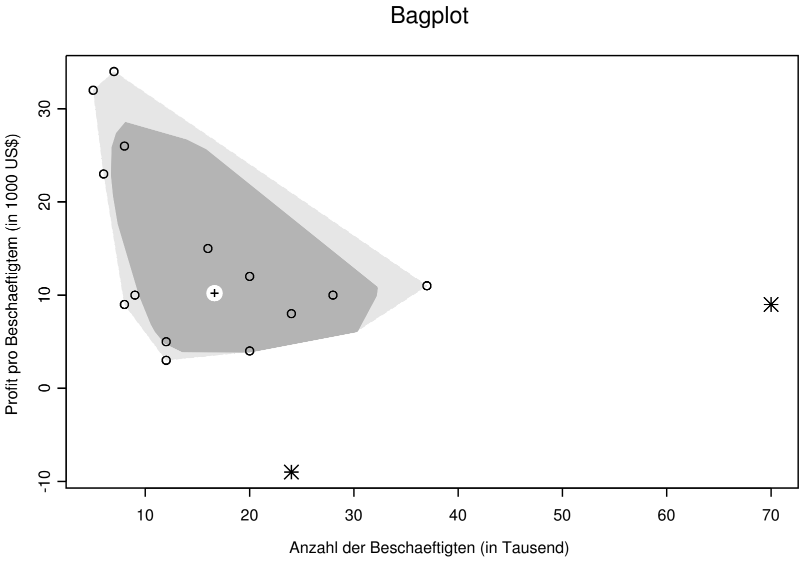


Figure 1.2: Bagplot of the financial data (`bagplot()` in the R package `aplpack`)

The scatter plot can also be used in matrix form for more than two dimensions, as shown with the next example.

Example 1.1.2 *Table 1.1 shows data of paper quality measurements. Since the fibers in paper have a distinct direction, the tearing strength can be measured in different directions during production. Here we have measurements along (x_2) and orthogonal (x_3) to the production direction. Variable x_1 specifies the denseness of the paper in g/cm^3.*

These 3-dimensional data can be visualized in scatter plots by plotting all variable pairs against each other, see Figure 1.3.

Table 1.1: Paper quality measurements; x_1 = denseness, x_2 = tearing strength along production, x_3 = tearing strength orthogonal to production.

x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
0.801	121.41	70.42	0.832	117.51	71.62	0.822	130.50	80.33
0.824	127.70	72.47	0.796	109.81	53.10	0.822	127.90	75.68
0.841	129.20	78.20	0.759	109.10	50.85	0.843	123.90	78.54
0.816	131.80	74.89	0.770	115.10	51.68	0.824	124.10	71.91
0.840	135.10	71.21	0.759	118.31	50.60	0.788	120.80	68.22
0.842	131.50	78.39	0.772	112.60	53.51	0.782	107.40	54.42
0.820	126.70	69.02	0.806	116.20	56.53	0.795	120.70	70.41
0.802	115.10	73.10	0.803	118.00	70.70	0.805	121.91	73.68
0.828	130.80	79.28	0.845	131.00	74.35	0.836	122.31	74.93
0.819	124.60	76.48	0.822	125.70	68.29	0.788	110.60	53.52
0.826	118.31	70.25	0.971	126.10	72.10	0.772	103.51	48.93
0.802	114.20	72.88	0.816	125.80	70.64	0.776	110.71	53.67
0.810	120.30	68.23	0.836	125.50	76.33	0.758	113.80	52.42
0.802	115.70	68.12	0.815	127.80	76.75			

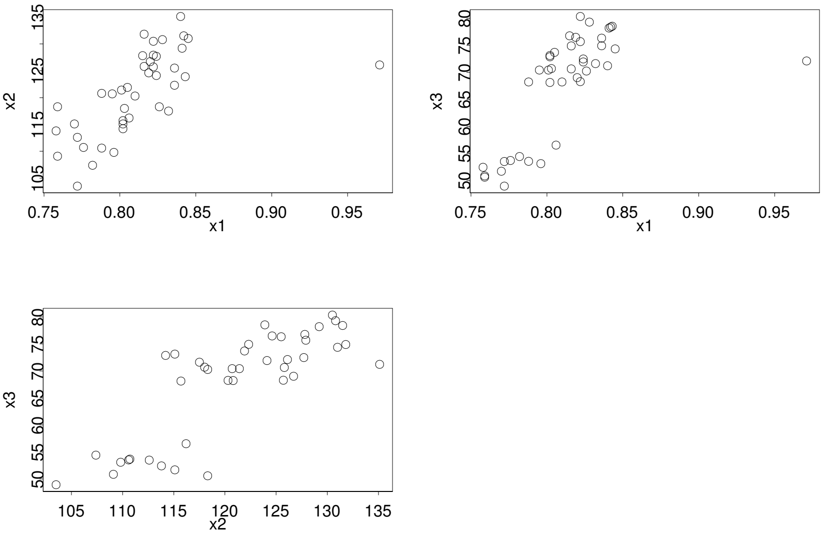


Figure 1.3: Scatter plot of the paper quality data set.

So far we have just used bivariate scatter plots to visualize the data. With interactive visualization it is also possible to show dynamical graphics of 3-dimensional scatter plots. One can even go further and visualize higher-dimensional data. One approach for this is the *Grand Tour* introduced by Asimov (1985), where 2-dimensional projections of p-dimensional data are considered, and continuously changing the projection directions. This is implemented in the software package *GGobi* (`http://www.ggobi.org`).

Table 1.2: Measurements of the strength of plates; x_1: measurement from shock wave, x_2: measurement from vibration, x_3, x_4: from statistical tests.

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
1889	1651	1561	1778	1954	2149	1180	1281
2403	2048	2087	2197	1325	1170	1002	1176
2119	1700	1815	2222	1419	1371	1252	1308
1645	1627	1110	1533	1828	1634	1602	1755
1976	1916	1614	1883	1725	1594	1313	1646
1712	1712	1439	1546	2276	2189	1547	2111
1943	1685	1271	1671	1899	1614	1422	1477
2104	1820	1717	1874	1633	1513	1290	1516
2983	2794	2412	2581	2061	1867	1646	2037
1745	1600	1384	1508	1856	1493	1356	1533
1710	1591	1518	1667	1727	1412	1238	1469
2046	1907	1627	1898	2168	1896	1701	1834
1840	1841	1595	1741	1655	1675	1414	1597
1867	1685	1493	1678	2326	2301	2065	2234
1859	1649	1389	1714	1490	1382	1214	1284

Example 1.1.3 *Table 1.2 shows 4 measurements of the strength of 30 plates. The first measurement was obtained by a shock wave onto the plate, the second from a vibration, and the remaining from statistical tests.*

Figure 1.4 shows different perspectives of the data from Table 1.2 in a 3-dimensional scatter plot. Figure 1.4a reveals the outliers, while Figure 1.4b rather masks the outliers.

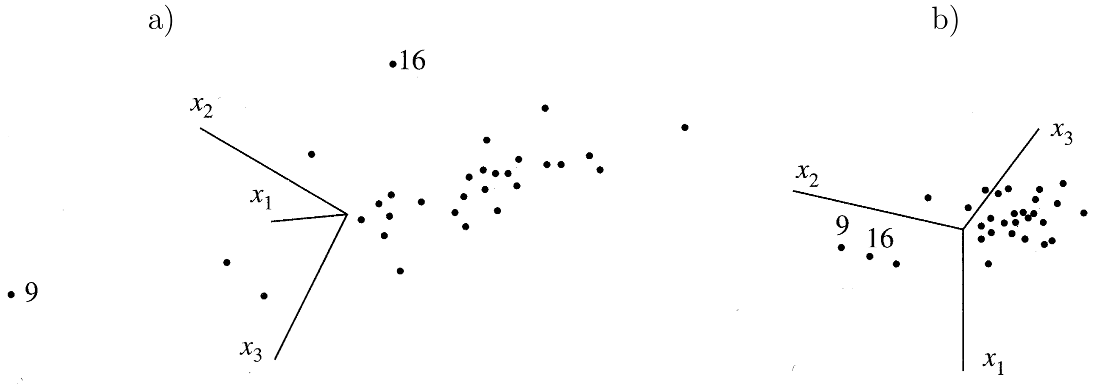


Figure 1.4: Different perspectives of 3-dimensional scatter plots for the plate data set.

Even these simple plots reveal already some structures in the data. For example, Figure 1.3 clearly shows 2 data groups and a single outlier. It would now be intersting to know the reasons for the grouping effect and for the outlier.

There are also *multivariate graphics.* The idea is to represent the multivariate observations by specific symbols. For example, *star plots* represent the variables by rays arranged in a star-shape, and the values for each variable of an observation are represented as the lengths of the corresponding rays. Then the ends of neighboring rays are connected, resulting in a specific star shape for each observartion. Stars which look similar refer to similar multivariate observations.

Figure 1.5 shows a star plot of the election results of the Republican party from the period 1856-1976 (31 values) for 9 of the north-eastern states in the US. The values are placed clockwise, starting from the right-hand side. One can identify states with very similar election structure.

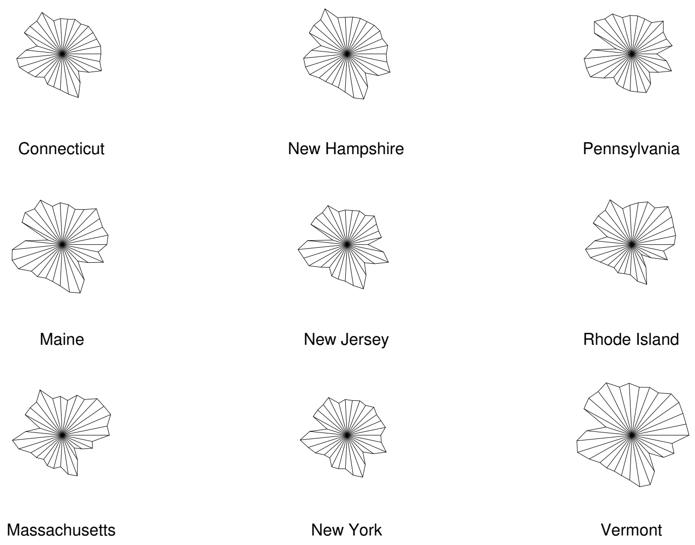


Figure 1.5: Star plot of the election data 1856-1976 for 9 US states.

Of course, there are many more multivariate graphics, but at some point they all have their limitations, and we have to rely on statistical analyses which directly process the multivariate data. Before doing that, we need to introduce a notation.

1.2 Notation

A matrix $\boldsymbol{X}$ of dimensionality $(n \times p)$ is denoted as

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad .$$

Here, one could think in terms of a data matrix, where the n observations are forming the rows, and the p variables the columns of $\boldsymbol{X}$. In these course notes, matrices are always printed by bold-faced letters. A matrix of order 1 is a vector, and again printed bold-faced with lower-case letter. Thus,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

is a vector, here with n components. Note that vectors are always denoted as column vectors! If a vector should be expressed as row vector, we need to transpose the vector:

$$\boldsymbol{x}^\top = (x_1, \dots, x_n) \quad .$$

Now, the columns of our matrix $\boldsymbol{X}$ can be written as

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p$$

and the rows (denoted as column vectors) as

$$\boldsymbol{x}_{1.}, \boldsymbol{x}_{2.}, \dots, \boldsymbol{x}_{n.} \quad ,$$

$$\text{with} \quad \boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad \text{and} \quad \boldsymbol{x}_{i.} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

for $j = 1, \dots, p$ and $i = 1, \dots, n$. This notation might be unusual, but it allows to distinguish observation vectors from variable vectors. Thus we can rewrite the matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p) = \begin{pmatrix} \boldsymbol{x}_{1.}^\top \\ \boldsymbol{x}_{2.}^\top \\ \vdots \\ \boldsymbol{x}_{n.}^\top \end{pmatrix} \quad .$$

1.3 Descriptive statistics

As in the univariate (1-dimensional) case, we are also interested in the multivariate case in simple statistical data descriptions. The location or center of each single variable (data column) thus can be of interest.

Let $x_{11}, x_{21}, \dots, x_{n1}$ be the n observations of the first variable. Then the *arithmetic mean* of these values is

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{i1} \ .$$

Similarly, one can compute the arithmetic mean of the observations of the other variables. This results in $\bar{x}_1, \dots, \bar{x}_p$ for the p variables of our $(n \times p)$ data matrix $\boldsymbol{X}$. The *multivariate (arithmetic) mean* is then given as a p-dimensional vector $\bar{\boldsymbol{x}} = (\bar{x}_1, \dots, \bar{x}_p)^\top$.

A further important statistical characterization is the variance. For our data matrix $\boldsymbol{X}$ we can estimate the variance of the first variable by the empirical variance

$$s_{11} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \ .$$

Generally, the empirical variance of the different variables is defined as

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

for $j = 1, \ldots, p$. Since also the relationship between the different variables is of interest, one can also estimate pairwise covariances by the empirical covariance

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \ .$$

for $j, k = 1, \ldots, p$, $j \neq k$. Note that we have symmetry, since $s_{jk} = s_{kj}$. One can arrange all these values in a matrix – the variances along the main diagonal – resulting in a $p \times p$ matrix which is symmetric. This is the (variance-)covariance matrix denoted by $\boldsymbol{S}$.

Since correlations are normed in $[-1, 1]$, they might be easier to interpret. The sample correlation coefficient is a measure for linear relationship, and it is defined as

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

for $j, k = 1, \ldots, p$. Of course, we have $r_{jj} = 1$. The elements r_{jk} can be arranged in the sample correlation matrix $\boldsymbol{R}$ of dimension $p \times p$.

1.4 Eigenvalues and eigenvectors

Consider a $(p \times p)$ matrix $\boldsymbol{\Sigma}$. Then,

$$q(a) = |\boldsymbol{\Sigma} - a\boldsymbol{I}| \tag{1.1}$$

defines a polynomial of order p in a. Here, $|\cdot|$ denotes the determinant.

Setting this polynomial to zero results in the p-th roots of $q(a)$, i.e. $a_1, \ldots, a_p$, and these are called eigenvalues of $\boldsymbol{\Sigma}$.

Example 1.4.1 *Consider the matrix*

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

Then

$$q(a) = |\boldsymbol{\Sigma} - a\boldsymbol{I}| = \begin{vmatrix} \sigma_{11} - a & \sigma_{12} \\ \sigma_{21} & \sigma_{22} - a \end{vmatrix} = (\sigma_{11} - a)(\sigma_{22} - a) - \sigma_{12}\sigma_{21}.$$

Obviously, this is a polynomial in a of order 2, and setting this to zero,

$$a^2 - a(\sigma_{11} + \sigma_{22}) + \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} = 0 \ ,$$

leads to two solutions for a, called a_1 and a_2.

In general, we have for $i = 1, \dots, p$,

$$|\boldsymbol{\Sigma} - a_i \boldsymbol{I}| = 0 \quad , \tag{1.2}$$

such that $\boldsymbol{\Sigma} - a_i\boldsymbol{I}$ is singular. Thus there exists a vector $\boldsymbol{\gamma}_i = (\gamma_{1i}, \dots, \gamma_{pi})^\top \neq \mathbf{0}$ with

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_i = a_i\boldsymbol{\gamma}_i \quad \text{for } i = 1, \dots, p. \tag{1.3}$$

$\boldsymbol{\gamma}_i$ is called (right-) eigenvector of $\boldsymbol{\Sigma}$ to the eigenvalue a_i. A real-valued eigenvector $\boldsymbol{\gamma}_i$ is called *standardized* if: $\boldsymbol{\gamma}_i^\top\boldsymbol{\gamma}_i = \sum_{j=1}^p \gamma_{ji}^2 = 1$.

If $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are eigenvectors to the same eigenvalue a_i, then also $\boldsymbol{\xi} + \boldsymbol{\zeta}$ and $\alpha\boldsymbol{\xi}$ ($\alpha \in I\!R$) are eigenvectors to a_i.

Theorem 1.4.1 *Let $\boldsymbol{C}$ be a non-singular matrix of order $(p \times p)$. Then we have:*

$$|\boldsymbol{\Sigma} - a_i\boldsymbol{I}| = |\boldsymbol{C}|\ |\boldsymbol{\Sigma} - a_i\boldsymbol{C}^{-1}\boldsymbol{C}|\ |\boldsymbol{C}^{-1}| = |\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1} - a_i\boldsymbol{I}| \quad . \tag{1.4}$$

This means that $\boldsymbol{\Sigma}$ and $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}$ have the same eigenvalues. If $\boldsymbol{\gamma}$ is eigenvector of $\boldsymbol{\Sigma}$ to the eigenvalue a_i, then we have

$$a_i\boldsymbol{C}\boldsymbol{\gamma} = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}(\boldsymbol{C}\boldsymbol{\gamma}) \quad . \tag{1.5}$$

It follows that $\boldsymbol{C}\boldsymbol{\gamma}$ is eigenvector of $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}$ to the eigenvalue a_i.

Theorem 1.4.2 *All eigenvalues of a symmetric $(p \times p)$ matrix $\boldsymbol{\Sigma}$ are real-valued.*

Proof: Let $\boldsymbol{\gamma} = \boldsymbol{\xi} + i\boldsymbol{\zeta}$, $a = b + ic$, $\boldsymbol{\gamma} \neq \mathbf{0}$.

By plugging these expressions into $\boldsymbol{\Sigma}\boldsymbol{\gamma} = a\boldsymbol{\gamma}$, we can compute the real and imaginary parts:

$$\boldsymbol{\Sigma}\boldsymbol{\xi} = b\boldsymbol{\xi} - c\boldsymbol{\zeta} \quad , \quad \boldsymbol{\Sigma}\boldsymbol{\zeta} = c\boldsymbol{\xi} + b\boldsymbol{\zeta} \quad . \tag{1.6}$$

Pre-multiplication with $\boldsymbol{\zeta}^\top$ and $\boldsymbol{\xi}^\top$, respectively, and subtraction leads to $c = 0$, i.e. a is a real number. $\boldsymbol{\gamma}$ can also be chosen as real, i.e. $\boldsymbol{\zeta} = \mathbf{0}$. □

Theorem 1.4.3 (Spectral Theorem or eigendecomposition)
Every symmetric matrix $\boldsymbol{\Sigma}$ of order $(p \times p)$ can be decomposed as

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^{\top} = \sum_{i=1}^p a_i\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top \quad , \tag{1.7}$$

where $\boldsymbol{A} = Diag(a_1, \dots, a_p)$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Gamma}$ is an orthogonal matrix (i.e. $\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^{-1}$), where the columns $\boldsymbol{\gamma}_i$ of $\boldsymbol{\Gamma}$, $i = 1, \dots, p$, are standardized eigenvectors of $\boldsymbol{\Sigma}$.

Proof: see, e.g., Johnson and Wichern (2007).

1.5 Expectation and covariance

Let $\boldsymbol{X}$ be a $(p \times q)$ matrix of random variables. We are using the notation $\boldsymbol{X} = [(x_{ij})]$. The expectation of this matrix is defined as

$$E(\boldsymbol{X}) = [(E(x_{ij}))] \quad . \tag{1.8}$$

This means that we obtain a matrix of order $(p \times q)$ with the element-wise expectations $E(x_{ij})$. Since the expectation operator is linear, it follows for data matrices (not random variables!) $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$

$$E(\boldsymbol{AXB} + \boldsymbol{C}) = \boldsymbol{A}\ E(\boldsymbol{X})\ \boldsymbol{B} + \boldsymbol{C} \quad . \tag{1.9}$$

Also the covariance can be transfered to the multivariate case. Let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_q)^\top$ be vectors of random variables. It is clear that we have

$$Cov(x_i, y_j) = E[(x_i - E(x_i))(y_j - E(y_j))]$$

for $i = 1, \ldots, p$ and j=$1, \ldots, q$. Due to the definition and properties of the expectation operator we obtain:

$$Cov(\boldsymbol{x}, \boldsymbol{y}) = [(Cov(x_i, y_j))] = E[(\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{y} - \boldsymbol{\mu_y})^\top] = E(\boldsymbol{x}\boldsymbol{y}^\top) - \boldsymbol{\mu_x}\boldsymbol{\mu_y}^\top \quad , \tag{1.10}$$

where $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ are the expectation vectors of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent, we have $Cov(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{O}$.

In the case $\boldsymbol{x} = \boldsymbol{y}$ we use instead of $Cov(\boldsymbol{x}, \boldsymbol{x})$ the notation $Cov(\boldsymbol{x})$, and we obtain

$$Cov(\boldsymbol{x}) = [(Cov(x_i, x_j))] = E[(\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})^\top] \quad . \tag{1.11}$$

The resulting matrix contains in the diagonal $(i = j)$ the variances, and thus it is also called *variance-covariance matrix.*

For data matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we obtain

$$Cov(\boldsymbol{Ax}, \boldsymbol{By}) = \boldsymbol{A}\ Cov(\boldsymbol{x}, \boldsymbol{y})\ \boldsymbol{B}^\top \quad , \tag{1.12}$$

and further

$$Cov(\boldsymbol{Ax}) = \boldsymbol{A}\ Cov(\boldsymbol{x})\ \boldsymbol{A}^\top \quad . \tag{1.13}$$

1.6 The multivariate normal distribution

Recall the density of the univariate normal distribution,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < \infty \ , \tag{1.14}$$

with expectation μ and variance σ^2.

The density of the multivariate normal distribution is obtained by replacing the distance $(x - \mu)/\sigma$ by a multivariate distance

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) := \mathrm{MD}^2(\boldsymbol{x}) \ , \tag{1.15}$$

also called (squared) *Mahalanobis distance.* Moreover, the constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ is replaced by a term such that the volume described by the multivariate density function is standardized to 1.

Thus, let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ be a p-dimensional random variable with expectation $E(\boldsymbol{x}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and covariance $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma} = [(\sigma_{ij})]$. Assume that $\boldsymbol{\Sigma}$ is positive definite (denoted by $\boldsymbol{\Sigma} \geq \boldsymbol{O}$), which is the case if (and only if) all eigenvalues are strictly positive.

Definition 1.6.1 *With the above notation,* $\boldsymbol{x}$ *follows a* p**-dimensional normal distribution** *if the density function of* $\boldsymbol{x}$ *is*

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.16}$$

with $-\infty < x_i < \infty$ *for* $i = 1, \ldots, p$*. We will use the notation* $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*.*

Properties: For $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have:

(a) $\boldsymbol{x} - \boldsymbol{\mu} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$

(b) Let $\boldsymbol{A}$ be a $(q \times p)$ matrix, then we have $\boldsymbol{A}\boldsymbol{x} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$.

(c) Setting $MD(\boldsymbol{x})$ equal to a constant c, i.e.

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) := c^2 \quad , \tag{1.17}$$

defines an ellipsoid in the p-dimensional space, with center $\boldsymbol{\mu}$ (see also Chapter Principal Component Analysis). On this ellipsoid, the density of the p-variate normal distribution is constant. Further, it holds that for normal distributed $\boldsymbol{x}$, the expression

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \sim \chi^2_p \ , \tag{1.18}$$

i.e. we obtain a χ^2 distribution with p degrees of freedom (p is the dimension of $\boldsymbol{x}$).

The expression $\sum_{i=1}^n(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top$ is the multivariate analogon to the univariate sum-of-squares $\sum_{i=1}^n(y_i - \bar{y})^2$. This leads to the so-called **Wishart distribution**, which is defined as follows:

Definition 1.6.2 *A symmetric* $(p \times p)$ *matrix* $\boldsymbol{W}$ *of random variables follows a Wishart distribution, if* $\boldsymbol{W}$ *can be represented by independent* p*-dimensional identically* $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$*-distributed random variables* $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}$ *as*

$$\boldsymbol{W} = \sum_{i=1}^n \boldsymbol{x}_{i.}\boldsymbol{x}_{i.}^\top = \boldsymbol{X}^\top\boldsymbol{X} \quad . \tag{1.19}$$

It is common to use the notation $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$*, or in short* $\boldsymbol{W} \sim W_p$*, and we call this a Wishart distribution with* n *degrees of freedom.*

Properties:

(a) Let $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$, and $\boldsymbol{A}$ is a $(p \times q)$ matrix, then we have that

$$\boldsymbol{A}^\top \boldsymbol{W} \boldsymbol{A} \sim W_q(\boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A}, n) \quad .$$

(b) $\boldsymbol{\Sigma}^{-\mathbf{1/2}} \boldsymbol{W} \boldsymbol{\Sigma}^{-\mathbf{1/2}} \sim W_p(\boldsymbol{I}, n)$

(c) For $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$ and a fixed (non-random) p-dimensional vector $\boldsymbol{a}$ with $\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} \neq 0$ we obtain

$$\frac{\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}} \sim \chi^2_n \quad .$$

(d) For $p = 1$, the elements of $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ are identically distributed according to $N_1(0, \sigma^2)$ and $\boldsymbol{x}^\top \boldsymbol{x} \sim W_1(\sigma^2, n)$. This means that the distribution $W_1(\sigma^2, n)$ is the same distribution as $\sigma^2 \chi^2_n$.

(e) For the diagonal elements of $\boldsymbol{W}$ we have $w_{ii} \sim \sigma_i^2 \chi^2_n$.

(f) If $\boldsymbol{W_1} \sim W_p(\boldsymbol{\Sigma}, n_1)$ and $\boldsymbol{W_2} \sim W_p(\boldsymbol{\Sigma}, n_2)$, and $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are independent, then we have $\boldsymbol{W_1} + \boldsymbol{W_2} \sim W_p(\boldsymbol{\Sigma}, n_1 + n_2)$.

Multivariate normality is of particular importance in multivariate statistics, because many methods and tests assume that the data follow a multivariate normal distribution. An approximation of the multivariate normal distribution is obtained by the following theorem:

Theorem 1.6.1 (Central Limit Theorem)
For n independent observations originating from a population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ (finite, non-singular), we have for large $n - p$ approximatively:

$$\sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{1.20}$$

and

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \chi^2_p \ . \tag{1.21}$$

How can you now check if a data set originates from a multivariate normal distribution? Similar as in the univariate case, there are tests and graphical tools available. The following **tests for multivariate normality** are widely used: chi-squared test (Conover, 1980), Kolmogoroff-Smirnoff test (Smirnov, 1948; Afifi und Azen, 1979), Anderson-Darling test (Anderson and Darling, 1952), and Shapiro-Wilks test (Shapiro und Wilk, 1965). The latter two test are in general more reliable. We do not go into detail here, but just refer to the cited literature.

Here we want to show a graphical procedure to check for multivariate normality. For the univariate Q-Q plot (Hazen, 1914), the quantiles of the empirical distribution are drawn against the quantiles of a standard normal destribution. The extension to the multivariate case is called $\boldsymbol{\chi}^2$**plot** or Gamma plot (Easton und McCulloch, 1990), and it is constructed as follows:

- For every observation $\boldsymbol{x}_{i.}$, $i = 1 \ldots, n$, one computes the squared Mahalanobis distances
$$\mathrm{MD}_i^2 = (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}) \tag{1.22}$$
and these values are sorted, $\mathrm{MD}_{(1)}^2 \leq \ldots \leq \mathrm{MD}_{(n)}^2$. Here, $\bar{\boldsymbol{x}}$ is the arithmetic mean vector and $\boldsymbol{S}$ is the sample covariance matrix.

- Compute $\chi^2_{p,i/n}$, i.e. the quantiles i/n of the χ^2_p distribution, and draw the pairs $(\chi^2_{p,i/n}, \mathrm{MD}_{(i)}^2)$ into the plot.

If the points are approximately following a linear trend, one can assume that the data follow a multivariate normal distribution. This test is only recommended if both n and $n - p$ are bigger than 30. Although the distances MD_i^2 are not independent of each other and not exactly χ^2 distributed, this graphical display is quite informative.

Example 1.6.1 *Let us consider again the 4-dimensional data set from Table 1.2. The corresponding χ^2 plot is shown in Figure 1.6. One can see that the outliers (observations 9 and 16) with the biggest distances are not well fitting the linear trend of the remaining points.*

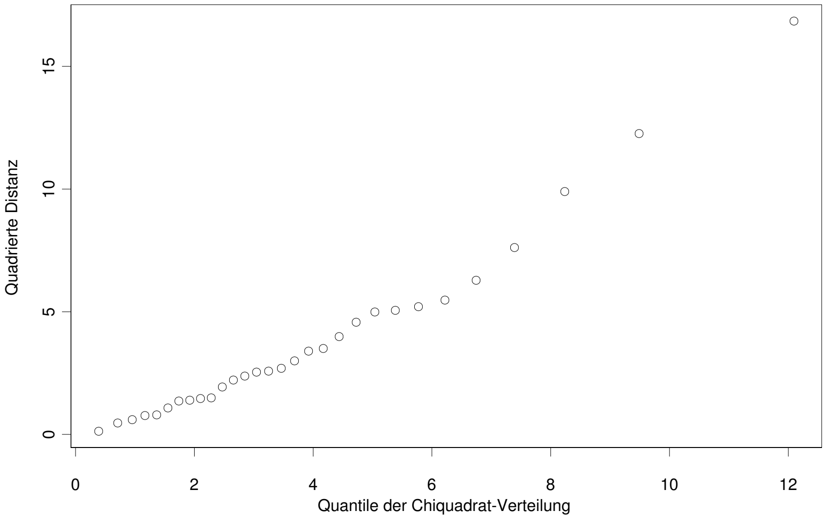


Figure 1.6: χ^2 plot for the data from Table 1.2.

1.7 Transformation to normality

If we realize that the underlying data set is not following a normal distribution, but we are still employing methods that require this assumption, the results and corresponding conclusions can be biased. However, there is still the possibility to transform the data in order to come closer to normality. If this is successful, one can then use those methods requiring normality.

A data transformation can be seen as a change of the data scale. For example, if the data are skewed to the right in one variable, a logarithmic transformation can lead to a distribution which is symmetric around the mean, and it can be close to normality.

In general, transformations are applied to every single variable of the multivariate data set. Based on theoretical considerations, the following transformations can be recommended for the listed types of data:

Original scaling	*Transformation*
Count data y	$\sqrt{y}$
Ratios, proportions $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$
Correlations r	Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$

In many cases it can only be decided based on the data set, which of the transformations is best suitable. A whole class of transformations is the *power transformation*: Let x be an observation (with strictly positive values) and λ be a parameter (can also be negative). Then the power transformation is defined as x^λ. For $\lambda = 0$ it is defined as $x^0 = \ln x$. For finding the appropriate power one can look at the histogram of the transformed data, or at the Q-Q plot. Once there is symmetry around the mean, and if the distribution comes close to normality, the optimal power has been identified. However, there is no guarantee that such a power exists to approach normality.

A slightly adapted version has been introduced by Box and Cox (1964):

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \tag{1.23}$$

This transformation is defined for real-valued λ and $x > 0$. Given observations $x_1, \ldots, x_n$, then the Box-Cox solution for the appropriate power λ is that one, where the expression (likelihood function)

$$l(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{i=1}^{n} (x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \tag{1.24}$$

is maximized. Here, $x_i^{(\lambda)}$ is defined in Equation (1.23), and

$$\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(\lambda)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i^\lambda - 1}{\lambda} \right) \tag{1.25}$$

is the arithmetic mean of the transformed observations.

$l(\lambda)$ can also be computed for many values of λ, and in a graphical display one can easily select the maximum. Usually one takes a value close to the optimum, which is also interpretable, such as $\lambda = 0$ (logarithm), $\lambda = \frac{1}{2}$ (square root), etc.

As mentioned above, in the multivariate case one can use the transformation for each single variable of the data set. Note that even the "best" transformation may not necessarily lead to normality, but probably just a good approximation.

1.8 Tests, confidence regions

In analogy to the univariate case, statistical tests can also be carried out in the multivariate case. Well known test such as tests of the mean can easily be transfered to the multivariate case. The difference now is that the p correlated variables need to be analyzed *jointly.*

If in the univariate case one wants to test whether the mean μ of a population is equal to a value μ_0, against the alternative hypothesis $H_1 : \mu \neq \mu_0$, then the corresponding test statistic (for unknown variance) is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \ , \tag{1.26}$$

which follows a t distribution with $n-1$ degrees of freedom, assuming i.i.d. (independent and identically distributed) samples from normal distribution.

We consider now in the multivariate case independent samples from a multivariate normal distribution, $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. A test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ can be carried out using the test statistic

$$T^2 = (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \left(\frac{1}{n}\boldsymbol{S}\right)^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \ , \tag{1.27}$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i.} \ , \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \text{ and } \boldsymbol{\mu}_0 = (\mu_{10}, \ldots, \mu_{p0})^\top \ .$$

This test statistic is known under *Hotelling's* T^2 (after the famous statistician Harold Hotelling), and it is distributed according to

$$T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,n-p} \ , \tag{1.28}$$

i.e. an F distribution with p and $n-p$ degrees of freedom. Thus, for a significance level of α we have:

$$\alpha = P\left[T^2 > \frac{(n-1)p}{(n-p)} F_{p,n-p;1-\alpha}\right] \tag{1.29}$$

In the univariate case, it is possible to construct a confidence interval as an equivalent to the hypothesis test. Similarly, this can be done in the multivariate case, but here we obtain a *confidence region.* Based on the above test we can thus easily construct a $100(1-\alpha)\%$ confidence region for the mean of a p-dimensional normally distributed variable. This is an ellipsoid determined by all $\boldsymbol{\mu}$, for which

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{p(n-1)}{(n-p)} F_{p,n-p;1-\alpha} \tag{1.30}$$

is fulfilled.

References

A.A. Afifi and S.P. Azen. *Statistical Analysis. A Computer Oriented Approach.* Acad. Press, New York, 1979.

T.W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34:122–148, 1963.

T.W. Anderson and D.A. Darling. Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23(2), 193–212, 1952.

D. Asimov. The grand tour. A tool for viweing multidimensional data. *SIAM Journal on Science and Statistical Computing*, 6, 128-143, 1985.

R.A. Becker, W.S. Cleveland, and A.R. Wilks. Dynamic graphics for data analysis. *Statistical Science*, 2(4):355–395, 1987.

G.E.P. Box and D.R. Cox. An analysis of transformations (with discussion). *J. Roy. Statist. Soc. B*, 26(2):211–252, 1964.

W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, New York, 1984.

G.S. Easton and R.E. McCulloch. A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association*, 85(410):376–386, 1990.

B. Everitt. *Graphical Techniques for Multivariate Data.* North-Holland, New York, 1978.

D.M. Hawkins, editor. *Topics in Applied Multivariate Analysis.* Cambridge University Press, Cambridge, 1982.

A. Hazen. Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, 77:1529–1669, 1914.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 6th edition, 2007.

A.M. Kshirsagar. *Multivariate Analysis.* M. Dekker, New York, 1972.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

P.J. Rousseeuw, I. Ruts, and J.W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

G.A.F. Seber. *Multivariate observations.* John Wiley & Sons, New York, 1984.

S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965.

N.V. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

J.W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass., 1977.

Chapter 2

Cluster analysis

2.1 Introduction

Cluster analysis aims at finding groups in a data set. One could be interested in identifying groups of observations, or groups of variables, or even both at the same time. Usually, it is not known beforehand if there is a grouping structure in the data, neither is it known how many groups there could be. Thus, cluster analysis is called an *unsupervised method*, which automatically (according to some defined algorithm) identifies the clusters.

Here we focus on identifying clusters of the n observations (objects). Obviously, the maximum number of clusters is n, and the minimum 1. Generally, observations which are assigned to one cluster should be similar to each other, while observations from different clusters are supposed to be dissimilar. Similarity is measured by a measure for homogeneity, while dissimilarity is characterized by heterogeneity (between the clusters). Usually, both of these measures are based on a distance definition.

In the following we assume p-variate observations $\boldsymbol{x}_{i.} = (x_{i1}, \ldots, x_{ip})^\top$, for $i = 1, \ldots, n$, sometimes simply called objects.

A widely used *distance or dissimilarity measure* between the i-th and the j-th object is the **Euclidean distance**

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\| \ , \tag{2.1}$$

also known under the term L_2 norm distance. An alternative, being less sensitive to outliers, is the **Manhattan distance** (or "city-block" distance), defined as

$$d(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| = \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|_1 \ , \tag{2.2}$$

also called L_1 norm distance.

The distance can be computed for every observation pair, which results in an $n \times n$ distance matrix $\boldsymbol{D} = [(d_{ij})]$. Clearly, this matrix is symmetric, and the main diagonal elements are all zero.

The distance matrix can be computed in R with the command `dist()`. However, the output only shows a lower diagonal matrix, for reasons of memory storage.

2.2 Types of classifications

It depends a lot on the application, which type of classification is most useful for the data at hand. One can distinguish between:

- **exhaustive classification:** Every observation is assigned to a cluster.
- **non-exhaustive classification:** Some of the observations might not be assigned to a cluster.

Non-exhaustive clustering would typically be applied if one expects outliers in the data, and these outliers should not be assigned to a group. In the following we assume an exhaustive classification.

In either case, there are still more types of possible and useful classifications:

- **Partition:** An observation is assigned only to one cluster. Thus, the clusters are non-overlapping. Usually, the number of clusters is fixed beforehand. The best known method creating partitions is *K-means clustering.*
- **Hierarchy:** A hierarchically organized clustering structure is obtained, in which one level of the hierarchy consists of a partition. The hierarchy usually starts with a partition into n clusters (consisting of single observations), and in higher levels of the hierarchy, different clusters are merged to obtain a partition with fewer clusters. This can be done as long until all the clusters are combined in a single cluster, consisting of n observations. This procedure is also called *agglomerative clustering.* The contrary would be a `divisive clustering`, where one cluster consisting of all n observations is decomposed step-by-step in the hierarchy, until all objects form a single cluster.

There are also more specific clustering methods which allow for overlapping clusters. These will not be treated here. However, we will consider a procedure called **fuzzy clustering**, where an observation is proportionally distributed among all the clusters.

Hierarchical clustering methods are appealing because the hierarchy can be visualized, and this provides an idea of a potentially underlying number of clusters. Such a visualization is called **dendrogram**.

2.3 Hierarchical clustering methods

As mentioned above, hierarchies consist of a sequence of partitions. We will only describe procedures for agglomerative clustering, where at the beginning, each object forms an own class, leading to n different clusters. At each step of the algorithm, the number of clusters is reduced by one, where the most similar classes are combined.

The "similarity" of the combined pair can be measured, and a "height" is associated with this newly formed class. At the end of the process there is only one single cluster left.

This calls for a new definition of similarity, expressing the distances between a group of observations in one cluster with indexes in the set C_k, and another group in the second cluster with indexes in C_l. The number of observations in each group is n_k and n_l, respectively. Now we can define different distance measures between clusters, which are also reflecting the name of the corresponding clustering algorithms:

- Complete Linkage:
$$\max_{i \in C_k, j \in C_l} d(i,j)$$
The similarity between two clusters is thus defined by that pair of objects from the different clusters which has the biggest distance.

- Single Linkage:
$$\min_{i \in C_k, j \in C_l} d(i,j)$$
Here, the similarity is given by the closest observations from two clusters. Single linkage tends to be unbalanced in the sense that big clusters are quickly combined. This procedure tends to produce many small groups and few large groups. Single linkage is also suitable to detect outliers.

- Average Linkage:
$$\frac{1}{n_k n_l} \sum_{i \in C_k} \sum_{j \in C_l} d(i,j)$$
The similarity is defined as the average of all pairwise distances.

- Centroid method:

  Here, one first needs to compute the arithmetic means (vectors) of the observations of each cluster, say $\bar{\boldsymbol{x}}(C_k)$ and $\bar{\boldsymbol{x}}(C_l)$. Then the similarity between the two clusters is given by the Euclidean distance of the cluster centers (centroids),

- Ward's method:

  The similarity between two clusters is defined as the increase of the variance when merging the two clusters,
$$\frac{\|\bar{\boldsymbol{x}}(C_k) - \bar{\boldsymbol{x}}(C_l)\|^2}{1/n_k + 1/n_l} .$$

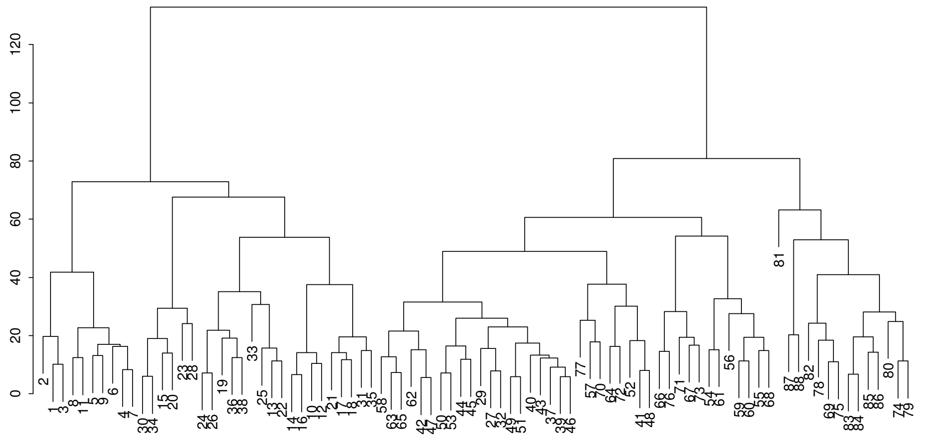
Those two clusters will be merged where the increase is the smallest possible.

For all of the above methods, the similarity needs to be computed for every pair of clusters, and then the pair with the smallest values of the measure will be merged. The resulting hierarchy can be presented in a **dendrogram**, where the horizontal axis presents the observations and the vertical axis the "height", which is the similarity measure at which clusters are merged. Thus, at the bottom we can find all the n clusters with single observations, and they are step-by-step merged, indicated by horizontal lines, untill all observations are joined in a single cluster (top). The observations are arranged in the plot in such a way to avoid overplotting of the lines.

Example 2.3.1 *The R package* `bootstrap` *contains the data set* `scor` *with the results of 88 students on exams in the five subjects Meachanics, Geometry, Algebra, Analysis, and Statistics. In each subject, a maximum of 100 points could be collected. It is not clear whether there are groups in the data, neither how many groups there could be.*

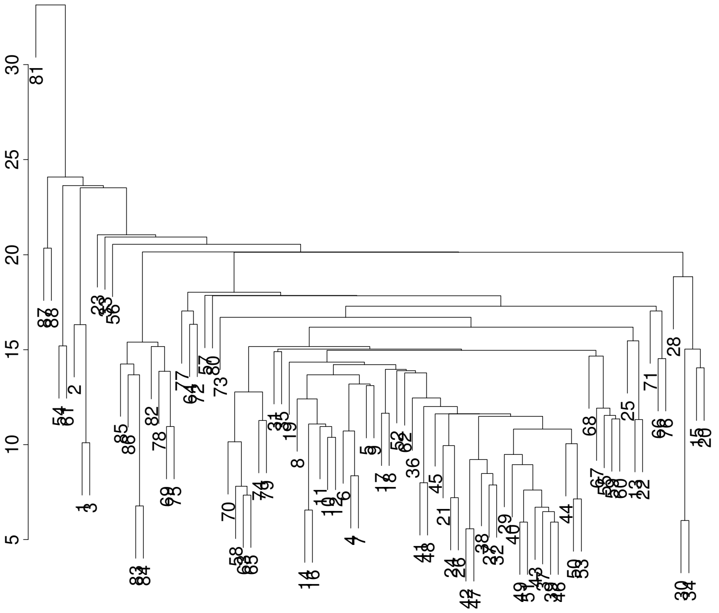
We are using hierarchical clustering based on complete linkage as well as single linkage. The results are shown by dendrograms in Figures 2.1 and 2.2. Hierarchical cluster analysis can be computed in R with the function `hclust()`*, where the input matrix needs to be a distance matrix, as computed with* `dist()`*, and the argument* `method` *indicates the similarity measure.* `plot()` *of the result object shows the dendrogram.*

Figure 2.1: Cluster dendrogram of the4 student exam data using *Complete Linkage.*



The complete linkage result (Figure 2.1) reveals that there seems to be a quite stable grouping structure with 2 clusters. Observation 81 seems to be an outlier. Single linkage in Figure 2.2 seems to focus more on identifying outliers and atypical observations. In either case, the dendrograms can be helpful to decide which final number of clusters should be chosen. The group assignments can then be obtained with the function `cutree(resobj,k)`*, where* `resobj` *is the cluster result object, and* `k` *the desired number of clusters.*

Figure 2.2: Cluster dendrogram of the student exam data using *Single Linkage*.



2.4 Partitioning methods

The best known algorithm for creating partitions is the **K-means algorithm**. Consider an index set C_k containing the indexes of the observations of the k-th cluster. One can define the so-called *total point scatter* as

$$T = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2(i,j) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\left(\sum_{j\in C_k} d^2(i,j) + \sum_{j\notin C_k} d^2(i,j)\right) .$$

It is then possible to decompose T as $T = W(C) + B(C)$ into a *within-cluster* point scatter

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j\in C_k} d^2(i,j)$$

and a *between-cluster* point scatter

$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j\notin C_k} d^2(i,j)$$

for a given a cluster partition C. Then, $B(C)$ tends to be large when observations assigned to different clusters are far apart. Thus, it is desirable to find a partition (for fixed K) which maximizes $B(C)$, which is equivalent to minimizing $W(C)$.

For K-means clustering one takes the Euclidean distance $d(i,j) = \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|$ as the distance measure. Then one can rewrite $W(C)$ from above as

$$W(C) = \sum_{k=1}^{K} n_k \sum_{i \in C_k} \|\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}_k\|^2 \ , \tag{2.3}$$

where n_k is the number of observations in the k-th cluster, and $\bar{\boldsymbol{x}}_k$ is the arithmetic mean vector of those observations. Minimizing this criterion means that the n observations are assigned to the K clusters in a way that the (average) distances of the observations from a cluster to their cluster center are minimized. This can be solved by an iterative algorithm:

Step 1. Cluster centers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$ are initialized. Usually, this is done by randomly selecting K observations as initial cluster centers.

Step 2. Minimize the objective function (2.3) by assigning each observation to the closest cluster center.

Step 3. Minimize (2.3) by replacing the cluster centers from Step 2. by the arithmetic means of the observations per cluster.

Step 4. Repeat Steps 2. and 3. until the assignments do not change.

Each of the Steps 2. and 3. reduce the value of $W(C)$ and thus convergence of this procedure is assured. However, one could end up in a local optimum, and it is thus recommended to re-start the procedure with different random initializations, and take that solution which gives the smallest value $W(C)$.

2.5 Model-based clustering

Model-based clustering can also be used to obtain a cluster partition, but the result would even give a “probability” for the assignment of an observation to a cluster. As the name indicates, model-based clustering makes use of a statistical model for the shape of the clusters. The standard “model” is multivariate normal distribution, i.e., it is assumed that the cluster has the density of a multivariate normal distribution, with a certain location and covariance. This is a big advantage over K-means clustering, since the cluster shapes can be more flexible. The result of K-means is typically spherically shaped clusters (due to the use of Euclidean distances to the center).

A detailed description of model-based clustering can be found in Fraley and Raftery (2002), and in many other sources of these authors. Assume that the data consist of K clusters, generated by multivariate normal densities with expectation $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, for $k = 1, \ldots, K$. Further, the class probabilities are given by so-called mixing coefficients $\pi_1, \ldots, \pi_K$, where $\pi_1 + \ldots + \pi_K = 1$. All these parameters are unknown, and they are estimated using the EM (expectation maximization) algorithm. Note that the covariance matrices are $p \times p$ matrices, and for larger p

there are many parameters to estimate from the available data, which can lead to instability. For this reason, the cluster “models” can be simplified, by imposing restrictions on the cluster covariance structures.

The simplest possibility for such restrictions is $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$, for $k = 1, \ldots, K$, where $\boldsymbol{I}$ is the identity matrix and σ^2 is a parameter for the variance. This would imply that all clusters are spherical, with the same radius. The estimation of the covariances thus reduces to estimating only one parameter, the variance σ^2. A less restricted covariance structure is $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$, for $k = 1, \ldots, K$. In this case, the clusters are still spherical, but their size can be different according to their variance σ_k^2, which needs to be estimated. Figure 2.3 illustrates different covariance structures.

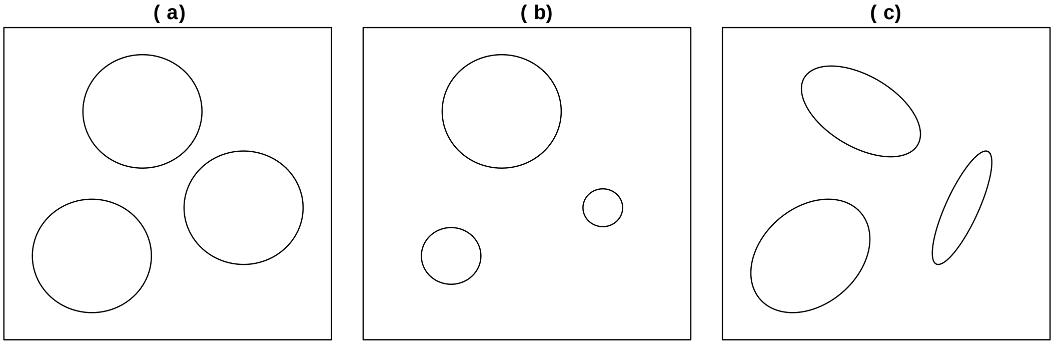


Figure 2.3: Different covariances for three clusters: (a) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \sigma^2 \boldsymbol{I}$; (b) $\boldsymbol{\Sigma}_j = \sigma_j^2 \boldsymbol{I}$, for $j = 1, 2, 3$; (c) all $\boldsymbol{\Sigma}_j$ different and of no special structure.

The R package `mclust` implements model-based clustering. The function `Mclust()` can also take an argument to compute all cluster solutions for a varying number of clusters. The optimal model can be selected according to a BIC criterion (Schwarz, 1978).

Example 2.5.1 *We consider the toy data set* `Nclus` *from the R package* `flexmix`*, consisting of 4 clusters in two dimensions. Figure 2.4 shows the outcome of the function. The left plot shows the BIC values (vertical axis) for different numbers of clusters (horizontal axis). Different cluster models are used, corresponding to the structure of the covariance matrices. The maximum BIC value points at the optimal model, which is a model with four clusters, and covariance structure “VVV”, meaning that all covariances are different from each other. The right plot presents the assignments of the observations to the different clusters, and also the shapes of the covariance structures together with the estimated group centers. Obviously, the clusters have been perfectly identified. The mixing coefficients correspond to the proportion of observations in the different clusters.*

2.6 Fuzzy clustering

Partitioning methods are sometimes called *hard* clustering methods, since they assign an observation to a cluster or not (1 or 0). In contrast, fuzzy clustering meth-

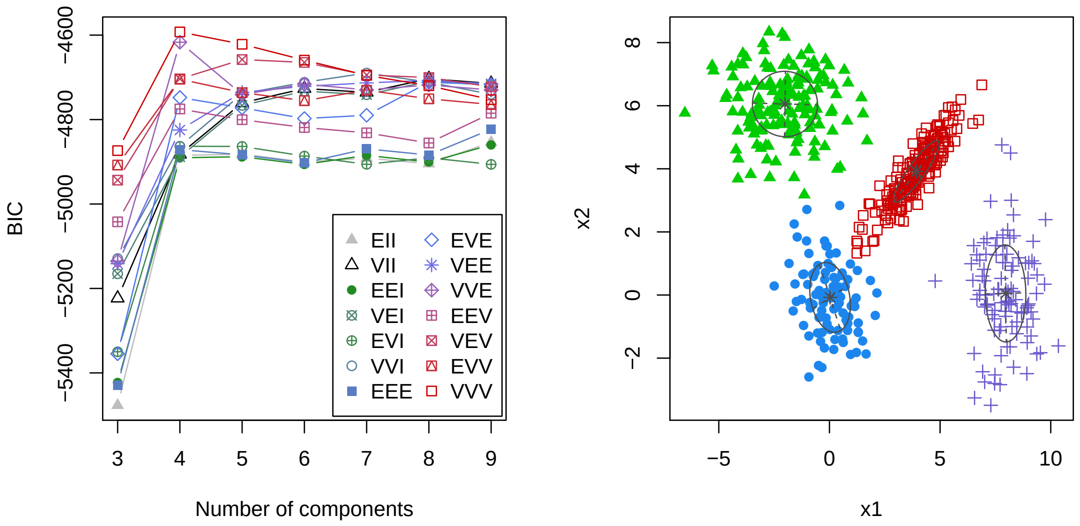


Figure 2.4: Result of model-based clustering: Left: the choice of the optimal cluster model based on the BIC criterion; right: the resulting cluster assignments.

ods allow for a proportional assignment of an observation to all clusters, where the sum of the proportions is 1. The coefficients for the proportional assignments are called *membership coefficients* $u_{ik} \in [0, 1]$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, with $\sum_{k=1}^{K} u_{ik} = 1$ for all i.

The best-known fuzzy clustering algorithm is called **fuzzy K-means** algorithm (Bezdek, 1974; Dunn, 1974), and it works very similar to the K-means procedure. First of all, K has to be given. The objective function (2.3) of K-means clustering is replaced by

$$\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{m}_k\|^2 \ , \tag{2.4}$$

which has to be minimized. Here, $\boldsymbol{m}_k = (m_{k1}, \ldots, m_{kp})^\top$ is the weighted cluster center of cluster k, defined as

$$m_{kj} = \frac{\sum_{i=1}^{n} u_{ik}^2 x_{ij}}{\sum_{i=1}^{n} u_{ik}^2}$$

for $j = 1, \ldots, p$.

One can show that the following equality holds:

$$\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{m}_k\|^2 = \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ik}^2 u_{jk}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|^2}{2 \sum_{j=1}^{n} u_{jk}^2} \tag{2.5}$$

This shows that the squared Euclidean distances between the observations enters the objective function, and thus it is also possible to use the distance matrix rather than the data matrix as an input of the procedure.

Of course, there are many more proposals of objective functions, and they determine the cluster solution (membership coefficients).

Fuzzy clustering is implemented in the R package `e1071` as function `cmeans`. We use the same data as before with model-based clustering, and set $K = 4$. Figure 2.5 shows the resulting matrix of membership coefficients as gray-scale symbols – the darker, the closer the value is to 1. Every plot represents the memberships for one particular cluster. Note that similar as for K-means clustering, also here the procedure ends up with clusters which tend to be spherically shaped.

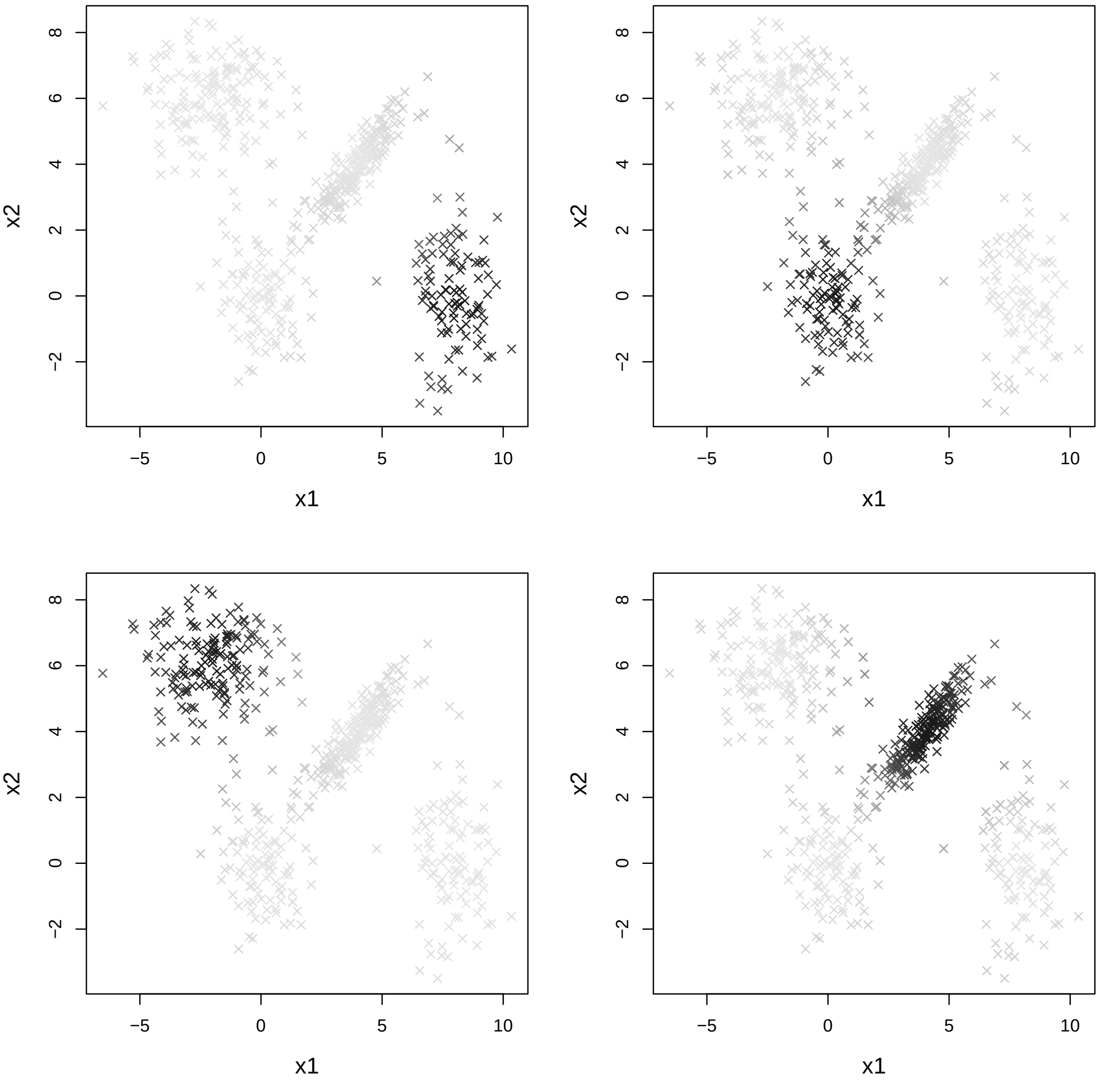


Figure 2.5: Result of fuzzy clustering: The four plots show the resulting membership coefficients for each of the four clusters in grey scale; dark means high value, light means low value.

2.7 Evaluation of the classification

The difficulty with cluster analysis is not only that there are various different procedures how to perform the clustering (hierarchical, partitioning, fuzzy clustering, etc.), but that for each procedure there exist several different algorithms. Moreover, several cluster algorithms require input parameters, like the number of clusters, and depending on this choice, the results can differ quite a lot. Consequently, there is a need for comparing the outcomes, and this is done by using so-called *cluster validity measures.*

The main goal of cluster analysis is to achieve highly homogeneous clusters, i.e. the observations within a cluster should be very similar to each other. On the other hand, different clusters should be dissimilar, because otherwise they should have been merged into one cluster. In other words, heterogeneity between different clusters should be achieved. Heterogeneity can be measured by

$$B_K = \sum_{k=1}^{K} \|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}\|^2, \tag{2.6}$$

where $\bar{\boldsymbol{x}}_k$ is the k-th cluster center ($k = 1, \ldots, K$), and

$$\bar{\boldsymbol{x}} = \frac{1}{K} \sum_{k=1}^{K} \bar{\boldsymbol{x}}_k$$

is the overall mean of the cluster centers. This term is also called the *between cluster sum of squares.* Homogeneity within the clusters can be defined by

$$W_K = \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}_k\|^2. \tag{2.7}$$

This term is called the *within cluster sum of squares*, since it considers squared Euclidean distances from the observations to their own cluster center.

While B_K should be large, W_K should be small. However, both measures depend on the number K of clusters, and thus this needs to be considered in a validity measure. Two prominent measures are the **Calinski-Harabasz index**

$$\mathrm{CH}_K = \frac{B_K/(K-1)}{W_K/(n-K)}$$

and the **Hartigan index**

$$\mathrm{H}_K = \ln \frac{B_K}{W_K}.$$

Practically, one considers a range of values for the possible number of clusters and computes the validity measure(s) for each cluster solution. The largest value of the index determines the optimal number of clusters.

Another prominent validity measure is the **average silhouette width** (Kaufman and Rousseeuw, 1990). Before computing this value, some definitions have to be provided first. The average dissimilarity of an observation $\boldsymbol{x}_{i.}$ belonging to cluster C_k to all other observations of the same cluster is given by

$$d_{i,C_k} = \frac{1}{n_k - 1} \sum_{i,j \in C_k, i \neq j} d^2(i,j),$$

where n_k is the number of observations in cluster C_k. The average dissimilarity of $\boldsymbol{x}_{i.}$ to observations from another cluster C_l is given by

$$d_{i,C_l} = \frac{1}{n_l} \sum_{j \in C_l} d^2(i,j).$$

The smallest of these values is

$$d_{i,C} = \min_l d_{i,C_l},$$

and it corresponds to the smallest dissimilarity of the i-th observation to its "closest" cluster. The *silhouette value* is defined as

$$s_i = \frac{d_{i,C} - d_{i,C_k}}{\max(d_{i,C_k}, d_{i,C})}.$$

The values of s_i are within the interval $[-1, 1]$. If the value of s_i is close to 1, the observation is well classified, a value of zero means that the observation is in between two clusters, and a value of -1 refers to a poor classification. Observations with negative silhouette values are probably assigned to a wrong cluster. The average silhouette width is

$$\frac{1}{n} \sum_{i=1}^{n} s_i,$$

and the higher this value, the better the classification.

The **silhouette plot** is implemented in the R package `cluster` as function `silhouette()`.

Example 2.7.1 *We use the result from fuzzy clustering of the* `Nclus` *data with hard assignments of the observations to the 4 clusters. Figure 2.6 shows the silhouette plot, as a result of the R command* `plot(silhouette(res$cluster,dist(X)))`, *where* `res` *is the fuzzy cluster object, and* `X` *the data frame. It can be seen that only one observation would have been better placed in a different cluster. The average silhouette width is 0.65. In contrast, Figure 2.7 shows the silhouette plot for fuzzy clustering with only 3 clusters. Obviously, the results are much worse, and also the average silhouette width went down to 0.52.*

A final useful cluster validity measure is the **Gap statistic** (Tibshirani et al., 2001), implemented in the R package `cluster` as function `clusGap()`, see also

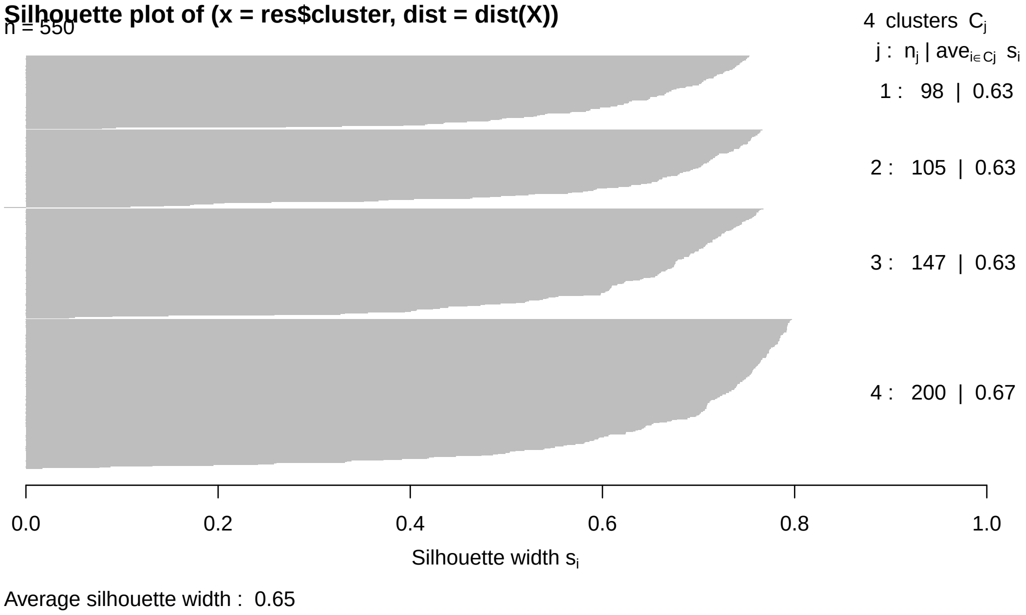


Figure 2.6: Resulting silhouette plot for fuzzy clustering (with hard cluster assignments) using 4 clusters.

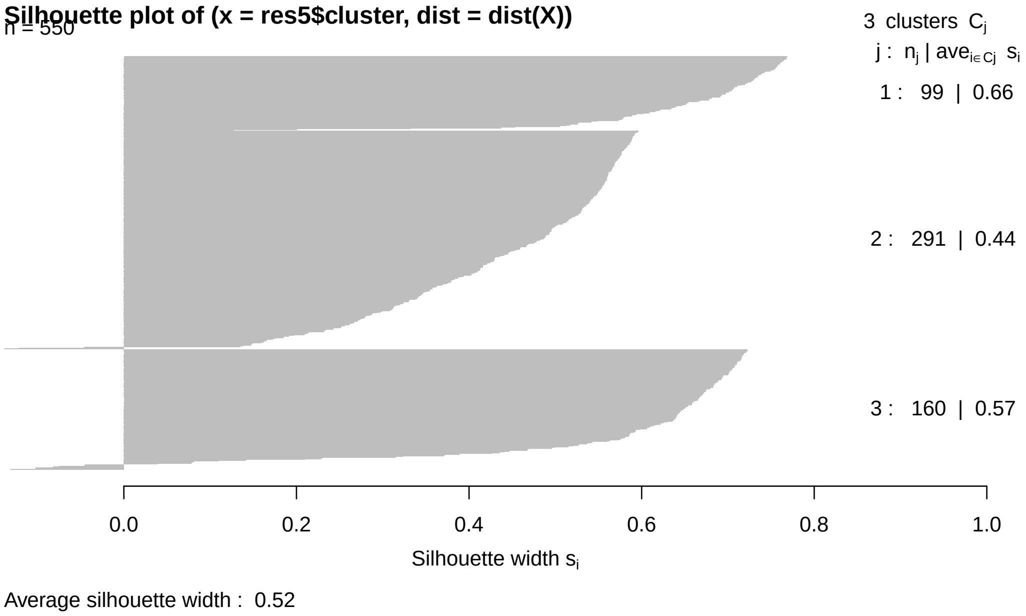


Figure 2.7: Resulting silhouette plot for fuzzy clustering (with hard cluster assignments) using 3 clusters.

`https://statweb.stanford.edu/ gwalther/gap`. The results can be computed e.g. by `clusGap(x, FUN=kmeans, K.max=10)`, where K-means clustering is used here for a maximum number of clusters $K = 10$. The results can be shown by `plot()`, and the function `maxSE()` returns the optimal number of clusters.

The idea is to consider

$$\tilde{W}_K = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,j \in C_k} d^2(i,j),$$

which is the pooled within-cluster sum of squares around the cluster means. The smaller the value $\tilde{W}_K$, the more compact are the points in the clusters – but this also depends on K: if K is very big, the clusters are naturally compact. The task is thus to find a possibly small value of K which still yields a small value of $\tilde{W}_K$.

The Gap statistic is defined as

$$\mathrm{Gap}_n(K) = E_n^*\{\log(\tilde{W}_K)\} - \log(\tilde{W}_K)\ , \tag{2.8}$$

where E_n^* denotes the expectation under a sample of size n from an appropriate reference distribution. For the reference distribution, a unimodal distribution is simulated, generated from a uniform distribution on the hypercube determined by the ranges of the input data. The expectaton is estimated by an average of B (e.g. $B = 100$) copies $\log(\tilde{W}_K^*)$, each of which is computed from a bootstrap sample $\boldsymbol{x}_{1.}^*, \ldots, \boldsymbol{x}_{n.}^*$ of this reference distribution.

Practically, we search for the smallest K such that the Gap statistic yields a local maximum. Since bootstrapping is used, one cannot just compute the average, but also a standard deviation and thus a standard error. Therefore, it is more advisable to look for the smallest K such that the value of the Gap statistic is not more than one standard error away from the first local maximum.

Example 2.7.2 *We apply the gap statistic to the* `Nclus` *data set, using fuzzy clustering, Forbesup to 10 clusters. Figure 2.8 shows the resulting plot. The first local maximum is for 4 clusters, and this is also the global maximum. The error bars reflect the standard errors.*

References

J.C. Bezdek. Cluster validity with fuzzy sets. *Cybernetics*, 3:58–72, 1974. Scripta Publ. Comp., Washington, D.C.

H.H. Bock. Automatische Klassifikation. In K.P. Grotemeyer, D. Morgenstern, and H. Tietz, editors, *Studia Mathematica/Mathematische Lehrbücher*, volume XXIV. Vandenhoeck & Ruprecht, Göttingen, 1974.

J.C. Dunn. A fuzzy relative of the isodata progress and its use in detecting compact well-separated clusters. *Cybernetics*, 3:32–57, 1974. Scripta Publ. Comp., Washington, D.C.

C. Fraley, and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the Americal Statistical Association*, 97(458), 611–631, 2002.

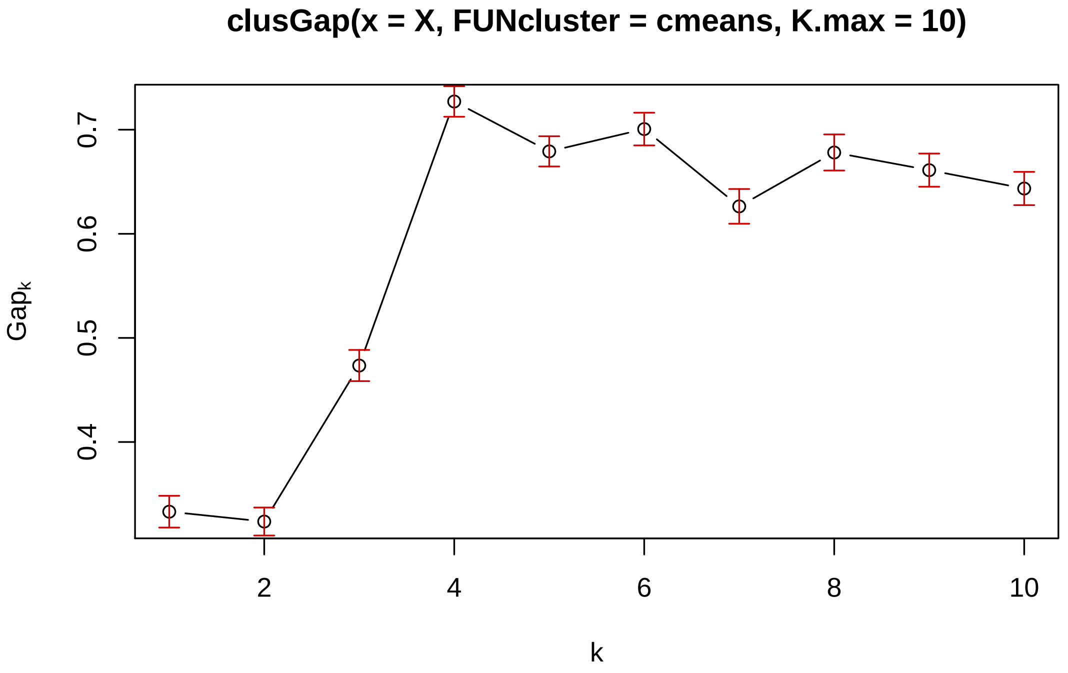


Figure 2.8: Resulting gap statistic plot for fuzzy clustering with up to 10 clusters.

J.A. Hartigan. *Clustering algorithms.* Wiley & Sons, New York, 1975.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data.* Wiley & Sons, New York, 1990.

M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems*, 1:239–253, 1978.

G.E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464, 1978.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statististic. *Journal of the Royal Statistical Society B*, 63, Part 2, , 411–423, 2001.

Chapter 3

Multivariate linear regression

3.1 Introduction

While in multiple linear regression we are interested in predicting the values of a response variable by using several explanatory variables, in multivariate regression analysis we want to predict several responses based on several explanatory variables. In both cases, we are investigating linear relationships.

We will start with a reminder of the multiple linear regression model, and then extend to the multivariate case.

3.2 Multiple linear regression

Consider the quantities $x_1, \ldots, x_p$, which are used to predict a response variable y. For example, the response could be the price of a car, and we are using car characteristics such as horsepower, miles per gallon, weight, height, etc., to predict the price. In our model, the price y will be treated as a random variable, while the characteristics are fixed (non-random) quantities.

The multiple linear regression model is thus of the form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \ . \tag{3.1}$$

The unknown regression parameters $\beta_0, \beta_1, \ldots, \beta_p$ determine the functional relationship which in this model is linear. Since the explanatory variables $x_1, \ldots, x_p$ will in general not perfectly describe the response y, there is also an additive error term ε involved.

Assume we have now n independent observations of y as well as of $x_1, \ldots, x_p$. Then our regression model can be written in terms of these observations as

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\
\vdots &\quad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n
\end{aligned}$$

or in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ . \tag{3.2}$$

$\boldsymbol{y}$ and the error terms $\boldsymbol{\varepsilon}$ are n-dimensional random variables. The first column of the *design matrix* $\boldsymbol{X}$ is a column of 1's, since these are the multipliers of the intercept term β_0. Thus, $\boldsymbol{X}$ has dimension $(n \times (p+1))$.

We will use the following assumptions for the error terms:

1. $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and

2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \boldsymbol{I}_n$.

σ^2 is the variance of the error terms, and it is the same for all components. Moreover, the different error terms are uncorrelated.

Later on we will also need distributional assumptions (normal distribution) in order to test hypotheses and to construct confidence intervals.

3.3 The least-squares estimator

Based on the observations y_i and $x_{i1}, \ldots, x_{ip}$, for $i = 1, \ldots, n$, we want to estimate the regression parameters and the error variance σ^2 by focusing on the differences $y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}$, which inform about the deviation between the observed response and the model fit. These differences are called **residuals**.

The *least-squares method* minimizes the sum of the squared residuals,

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \ . \end{aligned} \tag{3.3}$$

The coefficient minimizing this criterion is denoted as $\widehat{\boldsymbol{\beta}}$, and it is the least-squares estimator of the regression parameter $\boldsymbol{\beta}$.

We can multiply the terms in Equation (3.3):

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.4}$$

The resulting terms are scalars (not vectors or matrices), and thus we have

$$\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y})^\top = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.5}$$

Thus, Equation (3.3) is equal to

$$S(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.6}$$

The partial derivative with respect to the vector $\boldsymbol{\beta}$ yields

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 - 2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.7}$$

Setting this equation to zero gives the least-squares estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad . \tag{3.8}$$

Remark: If $\boldsymbol{X}$ does not have full rank $p + 1 \leq n$, we cannot compute $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and a way out would be to use a generalized inverse (e.g. Moore-Penrose) $(\boldsymbol{X}^\top \boldsymbol{X})^{-}$.

With $\widehat{\boldsymbol{\beta}}$ one can now compute the *fitted values* $\widehat{\boldsymbol{y}}$ of $\boldsymbol{y}$ by

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \ , \tag{3.9}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ is called "hat"-matrix. The *estimated residuals* are

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \ . \tag{3.10}$$

We can find that $\boldsymbol{X}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{0}$ and $\widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \widehat{\boldsymbol{\varepsilon}} = 0$. This means that the estimated residuals are orthogonal to the columns of $\boldsymbol{X}$ and to the fitted values $\widehat{\boldsymbol{y}}$.

Theorem 3.3.1 *For the linear regression model (3.2) we have:*
The least-squares estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ *is unbiased, i.e.* $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, *and* $Cov(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.
The estimated residuals $\widehat{\boldsymbol{\varepsilon}}$ *have the following properties:*
$E(\widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{0}$ *and* $Cov(\widehat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$. $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\varepsilon}}$ *are uncorrelated.*

Proof: see, e.g. Johnson und Wichern (2007).

This implies that the estimator has the following optimal properties:

Theorem 3.3.2 (Gauss-Markov)
Assume in the regression model (3.2) that the error terms are uncorrelated, and that $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ *(homoscedasticity). Then we have:*

(a) $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ *is a uniquely determined efficient linear estimator for* $\boldsymbol{\beta}$,

(b) $s^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{n-p-1} = \frac{1}{n-p-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ *is an unbiased estimator for the residual variance* σ^2.

Proof: see, e.g., Mardia et al. (1979), p. 172.

An efficient estimator has a covariance matric which is smaller than that of any other linear unbiased estimator. The least-squares estimator is thus also called a *best* linear unbiased estimator (BLUE).

Theorem 3.3.3 *If in addition to the assumptions of the Gauss-Markov theorem (Theorem 3.3.2) we have that*

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \quad , \tag{3.11}$$

we can construct confidence intervals for the parameters $\beta_j \quad (j = 1, \ldots, p)$ *and* σ^2. *We hereby use the notation* $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = [(g_{ij})]$ *for* $i, j = 1, \ldots, p$:

(a) $\left[\widehat{\beta}_j - t_{n-p-1;1-\frac{\alpha}{2}}\sqrt{s^2 g_{jj}},\ \widehat{\beta}_j + t_{n-p-1;1-\frac{\alpha}{2}}\sqrt{s^2 g_{jj}}\ \right]$
is a confidence interval for $\beta_j \quad (j = 1, \dots, p)$ *with coverage* $1-\alpha$.

(b) $\left[\frac{(n-p-1)s^2}{\chi^2_{n-p-1;1-\frac{\alpha}{2}}},\ \frac{(n-p-1)s^2}{\chi^2_{n-p1;\frac{\alpha}{2}}}\right]$
is a confidence interval for σ^2 *with coverage* $1-\alpha$.

3.4 Multivarate linear regression

While in the multiple linear regression model we considered one response variable, we have m response variables $y_1, \dots, y_m$ in the multivariate regression case. Thus, it is possible to construct for each single response variable a regression model based on the explanatory variables $x_1, \dots, x_p$:

$$\begin{aligned}
y_1 &= \beta_{01} + \beta_{11}x_1 + \ldots + \beta_{p1}x_p + \varepsilon_1 \\
\vdots \quad & \quad \vdots \\
y_j &= \beta_{0j} + \beta_{1j}x_1 + \ldots + \beta_{pj}x_p + \varepsilon_j \\
\vdots \quad & \quad \vdots \\
y_m &= \beta_{0m} + \beta_{1m}x_1 + \ldots + \beta_{pm}x_p + \varepsilon_m
\end{aligned} \tag{3.12}$$

For the error term $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_m)^\top$ we now have the assumptions $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. This means that the error terms for the different responses can be correlated with each other.

Consider now a sample of size n, then we can write down the regression model separately for every response variable $\boldsymbol{y}_j = (y_{1j}, \dots, y_{nj})^\top$ $(j = 1 \dots, m)$ in the same way as in the previous section:

$$\begin{aligned}
y_{1j} &= \beta_{0j} + \beta_{1j}x_{11} + \ldots + \beta_{pj}x_{1p} + \varepsilon_{1j} \\
y_{2j} &= \beta_{0j} + \beta_{1j}x_{21} + \ldots + \beta_{pj}x_{2p} + \varepsilon_{2j} \\
\vdots \ & \ \vdots \\
y_{nj} &= \beta_{0j} + \beta_{1j}x_{n1} + \ldots + \beta_{pj}x_{np} + \varepsilon_{nj}
\end{aligned}$$

All these equations can be collected in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{3.13}$$

with the matrices

- $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_m)$ of dimension $n \times m$,
- $\boldsymbol{X}$, still of dimension $n \times (p+1)$,
- $\boldsymbol{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)$ of dimension $(p+1) \times m$,
- $\boldsymbol{E} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_m)$ of dimension $n \times m$.

According to the assumptions in the multiple linear regression model, we have

$$E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$$

$$\mathrm{Cov}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k) = \sigma_{jk}\boldsymbol{I}_n \tag{3.14}$$

for $j, k = 1, \ldots, m$. This means that the m characteristics of the i-th trial ($i = 1, \ldots, n$) have covariance $\boldsymbol{\Sigma} = [(\sigma_{jk})]$, but observations from different trials are uncorrelated.

Also in multiple linear regression we are interested here in deriving the least-squares estimator. Similar as in Equation (3.3), we minimize the sum of squared residuals:

$$S(\boldsymbol{B}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \beta_{0j} - \beta_{1j}x_{i1} - \ldots - \beta_{pj}x_{ip})^2$$

One can show that this is the same as

$$S(\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{Y}^\top\boldsymbol{Y}) - 2\ \mathrm{tr}(\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{B}) + \mathrm{tr}(\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B})\ ,$$

where "tr()" denotes the trace of the matrix, i.e. the sum of the diagonal elements. The partial derivative with respect to $\boldsymbol{B}$ yields

$$\frac{\partial S(\boldsymbol{B})}{\partial \boldsymbol{B}} = -2\boldsymbol{X}^\top\boldsymbol{Y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\ ,$$

and this leads to the least-squares estimator

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\ . \tag{3.15}$$

Interestingly, this is just the same solution as if we would use the least-squares estimator for each single response variable: Denote $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_m)$, then we have

$$\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}_j\ , \tag{3.16}$$

as the least-squares estimator of $\boldsymbol{\beta}_j$ for the response $\boldsymbol{y}_j$, for $j = 1, \ldots, m$.

The fitted values of the responses are obtained by

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}\ . \tag{3.17}$$

Since the structure of the estimator is still the same as in the multiple linear regression case, still all the results are valid, in particular those of the Gauss-Markov theorem.

Example 3.4.1 *As an example we consider the data set* `mtcars` *from the package* `MASS`*, with 11 characteristics of 32 cars. We are interested in jointly predicting the variables* `mpg` *(miles per gallon),* `disp` *(displacement),* `hp` *(gross horsepower) and* `wt` *(weight) with the explanatory variables* `cyl` *(number of cylinders),* `am` *(transmission, with 0 = automatic, 1 = manual), and* `carb` *(number of carburetors). The essential R code is:*

```
> mtcars$cyl <- factor(mtcars$cyl) # define as factor variable
> Y <- as.matrix(mtcars[,c("mpg","disp","hp","wt")])
> mvmod <- lm(Y ~ cyl + am + carb, data=mtcars)
```

This results in the estimated coefficient matrix $\widehat{\boldsymbol{B}}$

```
> coef(mvmod)
                  mpg      disp         hp         wt
(Intercept) 25.320303 134.32487 46.5201421  2.7612069
cyl6        -3.549419  61.84324  0.9116288  0.1957229
cyl8        -6.904637 218.99063 87.5910956  0.7723077
am           4.226774 -43.80256  4.4472569 -1.0254749
carb        -1.119855   1.72629 21.2764930  0.1749132
```

Note that cars with 4 cylinders are taken as baseline (included in the intercept). Thus, having more cylinders has a negative effect on `mpg`*, but a positive effect on* `hp`*, as an example. Still, it is not clear if those effects are significant. Moreover, for significance considerations it would not be useful to consider single multiple regression models, since the error terms here are correlated.*

Unfortunately, the result of `summary()` *on this model indeed only provides inference for the models on the single responses:*

```
> mvmod.sum <- summary(mvmod)
> str(mvmod.sum)

List of 4
~~~~~~~~ snip ~~~~~~~~

> mvmod.sum[1]

Response mpg :

Call:
lm(formula = mpg ~ cyl + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9074 -1.1723  0.2538  1.4851  5.4728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.3203     1.2238  20.690  < 2e-16
cyl6         -3.5494     1.7296  -2.052 0.049959
cyl8         -6.9046     1.8078  -3.819 0.000712
am            4.2268     1.3499   3.131 0.004156
carb         -1.1199     0.4354  -2.572 0.015923
```

```
Residual standard error: 2.805 on 27 degrees of freedom
Multiple R-squared:  0.8113,Adjusted R-squared:  0.7834
F-statistic: 29.03 on 4 and 27 DF,  p-value: 1.991e-09
```

Indeed, when only considering the response `mpg`*, then* `cyl` *has a significant effect. Going back to the multivariate case, this would be an inappropriate test for testing significance of the parameters in the first column of $\widehat{\boldsymbol{B}}$. Appropriate tests would go even further: one could test for significance in different columns, or rows, or also of single elements of $\widehat{\boldsymbol{B}}$.*

Significance tests of the parameters in the multiple linear regression model involve the diagonal elements of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and s^2, the estimated residual variance. In the multivariate model we need to involve an estimator of the covariance $\boldsymbol{\Sigma}$, see Equation (3.14).

An unbiased estimator of $\boldsymbol{\Sigma}$ is obtained through the residual matrix $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ as

$$\boldsymbol{S}_R = \frac{1}{n-p-1} \boldsymbol{R}^\top \boldsymbol{R} \,. \tag{3.18}$$

Instead of the full model using all the p explanatory variables, we could now consider a reduced model with only $q < p$ variables – but still all the responses! Thus, instead of

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{i1} + \ldots + \beta_{pj} x_{ip} + \varepsilon_{ij}$$

we have the reduced model

$$y_{ij} = \beta^*_{0j} + \beta^*_{1j} x_{i1} + \ldots + \beta^*_{qj} x_{iq} + \varepsilon^*_{ij}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where here for simplicity, the reduced model omits the last $p - q$ variables from the full model.

Denote the least-squares estimator from the reduced model as $\widehat{\boldsymbol{B}}^*$, and fitted values as

$$\widehat{\boldsymbol{Y}}^* = \boldsymbol{X} \widehat{\boldsymbol{B}}^* \,,$$

and the residual matrix as $\boldsymbol{R}^* = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^*$. The estimator for the error covariance matrix is then

$$\boldsymbol{S}_{R^*} = \frac{1}{n-q-1} \boldsymbol{R}^{*\top} \boldsymbol{R}^* \,. \tag{3.19}$$

For testing the hypothesis that the reduced model is true (here this means that the last $p - q$ rows of $\boldsymbol{B}$ are zero), one can use the likelihood-ratio test statistic

$$\Lambda = \left(\frac{|\boldsymbol{S}_R|}{|\boldsymbol{S}_{R^*}|} \right)^{n/2} \tag{3.20}$$

which follows a *Wilks lambda* distribution. One can approximate the distribution for large n as follows:

$$-\nu \log \Lambda \sim \chi^2_{m(p-q)}$$

with $\nu = n - p - 1 - \frac{1}{2}(m - p + q + 1)$.

There are also some alternative tests, such as the tests of Pillai, Hotelling-Lawley, or Roy, which make use of similar "ingredients".

Example 3.4.2 *We continue our above example here, and we are interested in the question, whether the reduced model*

```
> mvmod0 <- lm(Y ~ am + carb, data=mtcars)
```

(without `cyl`) is preferable over the full model. This can be tested by

```
> anova(mvmod, mvmod0, test="Wilks")

Model 1: Y ~ cyl + am + carb
Model 2: Y ~ am + carb
  Res.Df Df Gen.var.   Wilks approx F num Df den Df    Pr(>F)
1     27      29.862
2     29  2   43.692 0.16395   8.8181      8     48 2.525e-07
```

or by

```
> anova(mvmod, mvmod0, test="Pillai")

Model 1: Y ~ cyl + am + carb
Model 2: Y ~ am + carb
  Res.Df Df Gen.var. Pillai approx F num Df den Df    Pr(>F)
1     27      29.862
2     29  2   43.692 1.0323   6.6672      8     50 6.593e-06
```

and these tests agree that the full model is preferable.

One can also test if the matrix $\boldsymbol{B}$ *is zero, agains the alternative that any of the numbers is different from zero. This can be done in R as follows:*

```
> res <- manova(Y ~ cyl + am + carb, data=mtcars)
> summary(res, test="Wilks")

          Df   Wilks approx F num Df den Df    Pr(>F)
cyl        2 0.06911  16.8240      8     48 1.451e-11
am         1 0.46547   6.8901      4     24 0.0007671
carb       1 0.29896  14.0696      4     24 4.798e-06
```

Here, a sequential model comparison is made, starting from a pure intercept model (thus, $\boldsymbol{B}$ *is zero), and adding step-by-step one of the explanatory variables, in the sequence defined by the formula. We can see that each additional explanatory variable has a significant contribution to the multivariate response. Alternative tests*

```
> summary(res, test="Pillai")
> summary(res, test="Hotelling-Lawley")
> summary(res, test="Roy")
```

lead to the same conclusions.

Finally, we are interested in checking the model assumption – which seems to be even more important in the multivariate case. Unfortunately, in R we get

```
> plot(mvmod)
Error: 'plot.mlm' is not implemented yet
```

. . . and this since many years already. A (suboptimal) way out would be to build multiple linear regression models for each response separately, and then to look at all the diagnostics plot. This is of course not ideal. One could, however, do some things "by hand". At least

```
> Yhat <-  predict(mvmod)
```

works, and this allows to compute the residual matrix:

```
> R <- Y - Yhat
```

A diagnostic for the multivariate residuals can be based on computing the Mahalanobis distances (to zero), and looking at the chi-square plot (see Chapter 1), which is shown in Figure 3.1.

```
> n <- nrow(R)
> m <- ncol(R)
> p <- 3        # number of explanatory variables
> sig <- 1/(n-p-1)*t(R)%*%R
> MD2 <- mahalanobis(R,rep(0,m),sig)
> plot(qchisq((1:n-0.5)/n,m),sort(MD2))
```

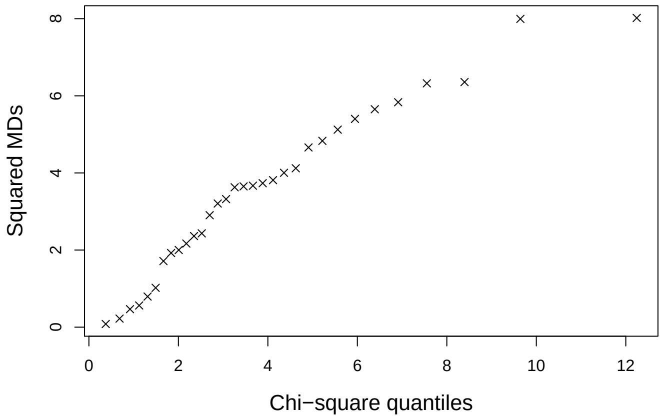


Figure 3.1: Chi-square plot for the multivariate residuals for the cars data set.

Figure 3.1 reveals some deviation from a straight line, which means that the normal distribution assumptions might be violated. This could have affected the inference tests.

References

N.R. Draper and H. Smith. *Applied Regression Analysis.* Wiley & Sons, New York, 1981.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 6th edition, 2007.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

G.A.F. Seber. *Multivariate Observations.* John Wiley & Sons, New York, 1984.

Chapter 4

Robust statistics – some concepts and tools

4.1 Introduction

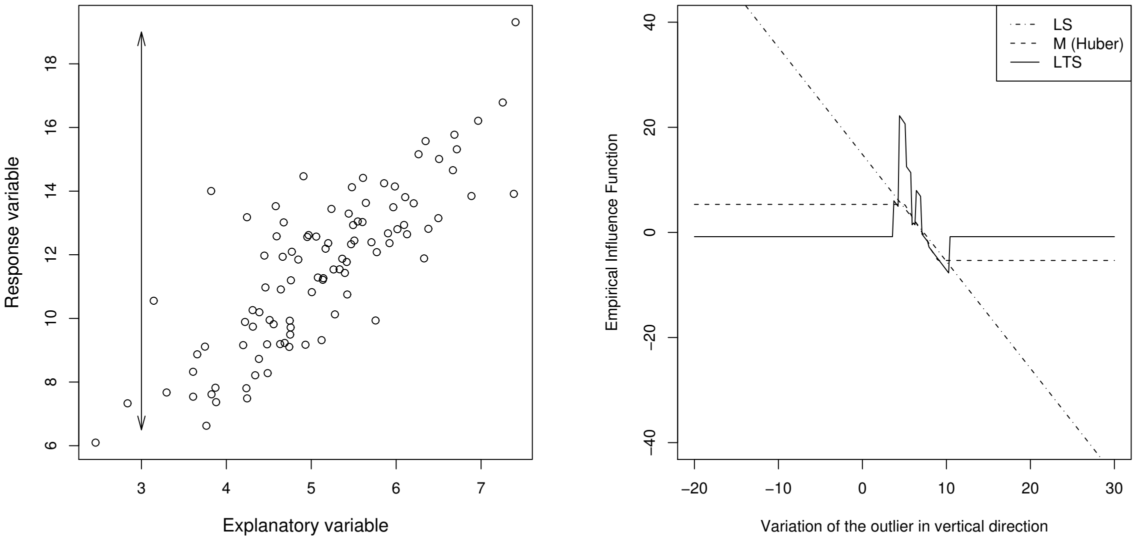
So far we have seen that multivariate statistical methods rely on quite strict model assumptions. If they are not met, the results could be misleading. The difficulty is that it is often hard to check if the assumptions are fulfilled. For example, tests on the distribution again have requirements such as independence of the (multivariate) observations, which might not be easy to check formally. Even worse, although there are diagnostic tools, such as the hat matrix in regression to identify leverage points (x-outliers), these tools are themselves sensitive to outliers.

The goal of *robust statistics* is to have less dependency on strict model assumptions. One still works with models and assumptions, but certain deviations and violations are tolerated. In a more human language one could say that robust methods focus on fitting only the majority of the data, where the requirements need to be fulfilled, but allow for deviations from the minority.

Here our focus is on robust regression and on robust covariance estimation. Understanding the robustness concepts in regression analysis allows to get an idea of how to robustify also many other methodologies. Further, covariance estimation is so crucial in multivariate statistics, so that various methods can be robustified just by plugging in a robust covariance estimate.

It is needless to say that real data almost always contain outliers. These are not necessarily wrong observations, but they can originate from different processes, they can refer to special phenomena, and thus they are often the most interesting observations. Good robust estimators downweight outliers during the estimation, but then provide (robust) diagnostics in order to reveal those observations.

The material presented in this chapter can be found in much more detail in the book Maronna et al. (2006).



(a) (b)

Figure 4.1: Empirical influence functions for regression estimators. Subplots show (a) the data with varying positions of the outlier and (b) the empirical influence functions of the slope parameter estimated by least squares (LS), M-regression and least trimmed squares (LTS).

4.2 Basic concepts of robustness

Which properties should a robust estimator have? It should be resistant to a sizeable proportion of outliers or deviation from assumptions. It should also still yield reasonable results if these ideal assumptions are valid. In general, we are interested in the influence function, the maxbias curve and breakdown point, and in the statistical efficiency. These issues are explained in the following.

Influence function

One of the basic ideas of robustness is that a limited amount of contamination should only have a small effect on the estimator. This can be simply empirically checked by varying single data points and looking at the effect on the estimator. As an example, Figure 4.1 shows a simple linear regression problem, where the observations were generated from a normal distribution. We are interested in the effect of one observation when it is varied along the indicated vertical line. For each position of the data point, we are interested in the change of the slope parameter of the following regression estimators: least-squares regression and two robust regression estimators: Huber M-regression and LTS regression (see later in this chapter). The result is known as the empirical influence function (EIF), but in order to make the results of the different estimators comparable, we compute the difference of the slopes for the contaminated and uncontaminated data, and divide by the amount $1/n$ of contamination. The right subplot of Figure 4.1 shows the results. It can be seen that the least-squares estimator has an unbounded EIF. This means that a single gross outlier can have an arbitrarily large effect on the estimator. Both robust estimators possess a bounded EIF. However, not only the bound but also the shape of the EIF is of importance in order to understand how a robust estimator deals with contamination. Ideally the EIF should be smooth: it should not show local spikes or should not be a step function. In practice, the effect of placing a data point at one location and then shifting it to a very close position should be very small. One observes that indeed the M-estimator has a smooth EIF and is thus virtually insensitive to local shifts in the data. On the contrary, the response of the least trimmed squares estimator to small data perturbations is far from smooth.

The concept of the EIF can be formalized by the so-called influence function (IF). The IF measures the influence an infinitesimal amount of contamination has on an estimator with respect to its position in space. More precisely, the influence function of an estimator T at a given distribution G is defined as:

$$\mathrm{IF}(\boldsymbol{x}, T, G) = \lim_{\varepsilon \downarrow 0} \frac{T\left[(1-\varepsilon)G + \varepsilon\delta_{\boldsymbol{x}}\right] - T(G)}{\varepsilon}, \tag{4.1}$$

where ε is the fraction of contamination and $\delta_{\boldsymbol{x}}$ is a probability measure which puts all the mass at $\boldsymbol{x}$. For robust estimators, the effect of small contamination will be limited.

Maxbias curve

So far we have considered small amounts of contamination. What happens if instead one changes more data points? What one expects is that a robust estimator can withstand a certain fraction of contamination. The mathematical tool to examine to which extent an estimator is distorted with respect to the fraction of contamination in the data is the maxbias curve. The maxbias curve measures the bias an estimator has with respect to the percentage of the worst possible type of contamination. Let $\boldsymbol{X}$ be the original data set and $\check{\boldsymbol{X}}$ be a data set in which m out of n observations have been replaced with arbitrary values, and let $\| \cdot \|$ denote the Euclidean norm. Then the maxbias curve for an estimator T is defined as:

$$\mathrm{maxbias}(m, T, \boldsymbol{X}) = \sup_{\check{\boldsymbol{X}}} \| \, T(\check{\boldsymbol{X}}) - T(\boldsymbol{X}) \, \| \, . \tag{4.2}$$

It is known that for some estimates of regression the worst possible type of outliers is found at points where y, x and the fraction y/x increase to infinity. Non-robust estimators, such as the least-squares regression estimator, turn out to reach a maxbias of infinity already at small amounts of contamination. This brings us to a further important concept.

Breakdown point

For every estimators, there exists a point where the maxbias tends to infinity. This point is referred to as the *breakdown point.* Loosely, the breakdown point indicates which percentage of the data may be replaced with outliers before the estimator yields aberrant results. Based on the maxbias curve, the breakdown point of an estimator T at a given sample $\boldsymbol{X}$ is given by:

$$\varepsilon_n^*(T, \boldsymbol{X}) = \min\left\{\frac{m}{n}; \mathrm{maxbias}(m, T, \boldsymbol{X}) = \infty\right\}. \tag{4.3}$$

For $n \to \infty$ one obtains the *asymptotic breakdown point*, denoted ε^*. For least-squares regression it holds that $\varepsilon^* = 0$. The maximal possible value of the asymptotic breakdown point of regression estimators equals 0.5 (if one asks for certain equivariance properties).

One of the goals in designing robust estimators is obtaining a high breakdown point. However, bounded influence and high breakdown should not result in a drastic decrease in efficiency.

Statistical efficiency

An important property to any statistical estimator is the variance. For instance, the least-squares estimator is also the minimum variance unbiased estimator for the linear model (the Gauss-Markov theorem). This implies that predictions made by any other regression estimator for data which follow the linear model with normally distributed error terms, will have a higher uncertainty than the least-squares predictions. Also robust estimators for regression are prone to an increase in variance compared to least-squares if the model assumptions are met. They are said to be less *efficient* than parametric estimators. The goal is thus to design robust estimators which are also efficient under normality, but at the same time achieve high (or at least positive) breakdown point and reasonable behavior of the influence function. These are goals that seem to be hard to achieve. However, one has to keep in mind that even if a robust estimator loses some efficiency, it will outperform the least-squares estimator in case of model violations, because the latter is only optimal at the exact normal model.

4.3 Robust regression

We already know that outliers will have an impact on traditional least-squares regression. However, which type of outliers has the biggest impact? How can one protect against such outliers?

These are some questions which will be answered in this section. Let us first have a look at the multiple linear regression model (here we don't treat multivariate linear regression).

We have observations of the explanatory variables, collected in the $n \times (p+1)$ matrix $\boldsymbol{X}$, with elements x_{ij}, and the n-vector $\boldsymbol{y}$ for the response, with elements y_i. The first column of ones takes care of the intercept term. Call $\boldsymbol{x}_{i.} = (1, x_{i1}, \ldots, x_{ip})^\top$ the column vector containing the elements of the i-th observation, i.e. the i-th row of $\boldsymbol{X}$. The linear regression model is then given by

$$y_i = \boldsymbol{x}_{i.}^\top \boldsymbol{\beta} + \varepsilon_i \ , \ i = 1, ..., n, \tag{4.4}$$

with the unknown regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, and the error terms ε_i, which are assumed to be i.i.d. random variables.

For a given estimator $\hat{\boldsymbol{\beta}}$, call $r_i = r_i\left(\hat{\boldsymbol{\beta}}\right) = y_i - \boldsymbol{x}_{i.}^\top \hat{\boldsymbol{\beta}}$ the i-th residual. The classical least-squares (LS) estimator is defined as

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i\left(\boldsymbol{\beta}\right)^2 \ . \tag{4.5}$$

An ancient alternative to LS is the L_1 estimator defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left|r_i\left(\boldsymbol{\beta}\right)\right| . \tag{4.6}$$

4.3.1 M-estimators

One can rewrite Equations (4.5) and (4.6) as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i\left(\boldsymbol{\beta}\right)), \tag{4.7}$$

where $\rho(r) = r^2$ for LS and $\rho\left(r\right) = |r|$ for L_1. By taking other ρ functions, different estimators are obtained. In fact, Equation (4.7) is the definition of a whole class of estimators commonly referred to as *M-estimators.*

In order to not depend on the scale of the response, a more suitable definition of M-estimators of regression is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i\left(\boldsymbol{\beta}\right)}{\hat{\sigma}}\right) \ , \tag{4.8}$$

where $\hat{\sigma}$ is a robust scale estimator of the residuals, that can be estimated either previously or simultaneously with the regression parameters.

Differentiating (4.8) with respect to $\boldsymbol{\beta}$ we get that the estimate fulfils the system of *M-estimating equations*

$$\sum_{i=1}^{n} \psi\left(\frac{r_i\left(\boldsymbol{\beta}\right)}{\hat{\sigma}}\right) \boldsymbol{x}_i = 0 \tag{4.9}$$

where $\psi = \rho'$.

For LS, $\psi\left(r\right) = r$, and (4.9) are the well-known normal equations. We may then in general interpret (4.9) as a robustified version of the normal equations, where the residuals are curbed.

Put $W\left(r\right) = \psi\left(r\right)/r$ and $w_i = W\left(r_i\left(\boldsymbol{\beta}\right)/\hat{\sigma}\right)$. Then (4.9) may be rewritten as

$$\sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \boldsymbol{x}_i = 0 \ . \tag{4.10}$$

Then (4.10) is a weighted version of the normal equations, and hence the estimator can be seen as weighted LS, with the weights depending on the data. For LS, W is constant. For an estimator to be robust, observations with large residuals should receive a small weight, which implies that $W\left(r\right)$ has to decrease to zero fast enough for large r.

Figure 4.2 shows some examples of ρ-functions, the resulting ψ-functions, and the corresponding weights. The first row in this figure are the corresponding functions for the LS estimator. It is clear from Equation (4.9) that a single observation can have an arbitrarily big effect on the estimator. In fact, it can be shown that **the influence function of the estimator is proportional to the ψ function**.

The second row of Figure 4.2 refers to the *Huber* family, with ρ' given by

$$\psi(r) = \begin{cases} r & \text{for} \quad |r| \leq b \\ b \text{ sign}\,(r) & \text{otherwise} \end{cases} \tag{4.11}$$

The extreme cases $b \to \infty$ and $b \to 0$ correspond to LS and L_1, respectively.

The last row in Figure 4.2 refers to the Tukey *bisquare* family, with

$$\rho\,(r) = \begin{cases} \left(\frac{r}{k}\right)^2 \left(3 - 3\left(\frac{r}{k}\right)^2 + \left(\frac{r}{k}\right)^4\right) & \text{for} \;\; |r| \leq k \\ 1 & \text{else} \end{cases} . \tag{4.12}$$

When $k \to \infty$, the corresponding estimate tends to LS and hence becomes more efficient and at the same time less robust. Thus, k is a tuning parameter the choice of which is a compromise between efficiency and robustness. The usual practice is to choose k to attain a given efficiency, such as 0.90.

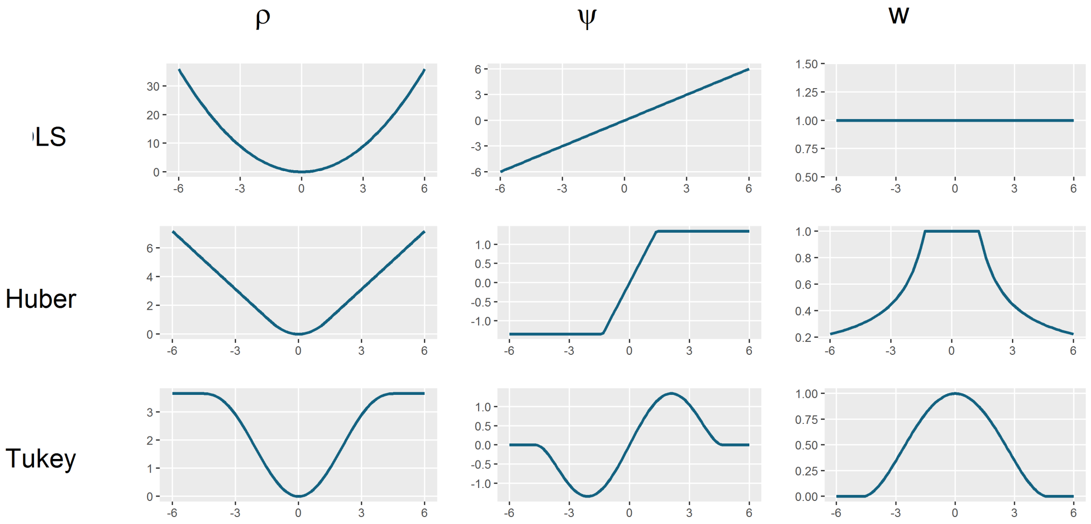


Figure 4.2: Different options for the function ρ, ψ and the weights w.

Note that the Tukey bisquare function has a bounded ρ-function. This is desirable because otherwise if some $\boldsymbol{x}_{i.}$ is large, then the i-th term will dominate the sum in (4.10), which would be unfortunate if $(\boldsymbol{x}_{i.}, y_i)$ is atypical (a so-called bad leverage point). For this reason it is better to use M-estimators given by (4.8) with a *bounded* ρ.

Computing M-estimators

Assume we have an initial value $\hat{\boldsymbol{\beta}}_0$, and $\hat{\boldsymbol{\beta}}_m$ is the approximation at iteration m. Then given $\hat{\boldsymbol{\beta}}_m$, compute the residuals $r_i = r_i\left(\hat{\boldsymbol{\beta}}_m\right)$ and then the weights $w_i = W\left(r_i/\hat{\sigma}\right)$, and solve (4.10) to obtain $\hat{\boldsymbol{\beta}}_{m+1}$. The procedure is called *iterative reweighted least squares* (IRWLS).

It is crucial that the initial estimator $\hat{\boldsymbol{\beta}}_0$ is robust – otherwise the algorithm might converge to a non-robust solution.

A good initial estimator is also necessary to obtain a robust estimate of the residual scale $\hat{\sigma}$. Some options are mentioned below. If there are no leverage points one could use L_1 as an initial $\hat{\boldsymbol{\beta}}_0$, and compute $\hat{\sigma}$ as a robust scale of the residuals r_i (e.g. the MAD).

4.3.2 Regression estimators based on a robust residual scale

Given $\boldsymbol{\beta}$**,** let $\boldsymbol{r}\left(\boldsymbol{\beta}\right) = \left(r_1\left(\boldsymbol{\beta}\right), ..., r_n\left(\boldsymbol{\beta}\right)\right)$. We consider an estimator of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}\left(\boldsymbol{r}\left(\boldsymbol{\beta}\right)\right) \tag{4.13}$$

where $\hat{\sigma}$ is a robust scale estimator.

Least Median of Squares (LMS) estimator:

The simplest robust scale estimator is the median of the absolute residuals:

$$\hat{\sigma}\left(\boldsymbol{r}\right) = \operatorname*{med}_i \left|r_i\right| \tag{4.14}$$

The corresponding regression estimator achieves highest breakdown point 50%, but a low efficiency. A tradeoff between breakdown and efficiency can be obtained by considering another quantile instead of the median.

Least Trimmed Squares (LTS) estimator:

Call $\left|r\right|_{(i)}$ the ordered absolute values of the residuals r_i, i.e. $\left|r\right|_{(1)} \leq ... \leq \left|r\right|_{(n)}$. A smoother alternative is to consider a scale more similar to the standard deviation, namely the *trimmed squares scale*

$$\hat{\sigma}\left(\boldsymbol{r}\right) = \left(\frac{1}{h}\sum_{i=1}^{h}\left|r\right|_{(i)}^2\right)^{1/2}, \tag{4.15}$$

with $h \in [n/2, n]$. Smaller values of h lead to higher breakdown point, but to lower efficiency.

M-estimator of scale

This estimator is defined as the solution σ of an equation of the form

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\hat{\sigma}}\right) = b \tag{4.16}$$

where ρ is a bounded ρ-function and b is a constant. It can be shown that the breakdown point of $\hat{\sigma}$ is $\min(b, 1-b)$. Equation (4.16) is nonlinear, but it is easy to solve iteratively. Put

$$W(z) = \frac{\rho(z)}{z^2}. \tag{4.17}$$

Then (4.16) can be rewritten as

$$\hat{\sigma}^2 = \frac{1}{nb} \sum_{i=1}^{n} w_i r_i^2$$

with $w_i = W(r_i/\hat{\sigma})$, which displays $\hat{\sigma}$ as a weighted RMSE. Given some starting value $\hat{\sigma}_0$, an iterative procedure can be implemented as was done for regression M-estimators.

The choice $\rho(z) = z^2$ and $b = 1$ yields the RMSE. The choice $\rho(z) = \mathrm{I}(|z| > 1)$ and $b = 0.5$ yields $\hat{\sigma} = \mathrm{med}(|r|)$.

Regression estimators with $\hat{\sigma}$ given by (4.16) are called **S-estimators**. Although S-estimators achieve the maximum breakdown point, they have low efficiency. However, they can be used as initial estimator for the M-estimator (4.8). The resulting estimator is called **MM-estimator**; it inherits the breakdown point of the S-estimator, but has controllable efficiency.

Example 4.3.1 *The package `robustbase` implements several robust procedures, also MM-regression (function `lmrob()`). Here we consider a simple example, the data set `Animals2` included in this package. It contains measurements of body and brain weight of different animals. There are also 3 dinosaurs, see Figure 4.3 (left), and they act as bad leverage points. The regression fits are shown in the plot, and obviously LS regression was affected by those outliers. MM-regression automatically downweights these observations, and the weights can be seen in the right plot of Figure 4.3.*

```
library(robustbase)
data("Animals2")
dat <- data.frame(body.log=log(Animals2$body),brain.log=log(Animals2$brain))
plot(dat)
res.ls <- lm(brain.log~body.log,data=dat)
abline(res.ls,col=2,lty=2)
res.mm <- lmrob(brain.log~body.log,data=dat)
abline(res.mm,col=3,lty=1)

plot(res.mm$residuals,res.mm$rweights)
```

4.4 Robust location and covariance

Multivariate location and covariance play a central role in multivariate statistics because many multivariate methods directly build on these estimates. For example,

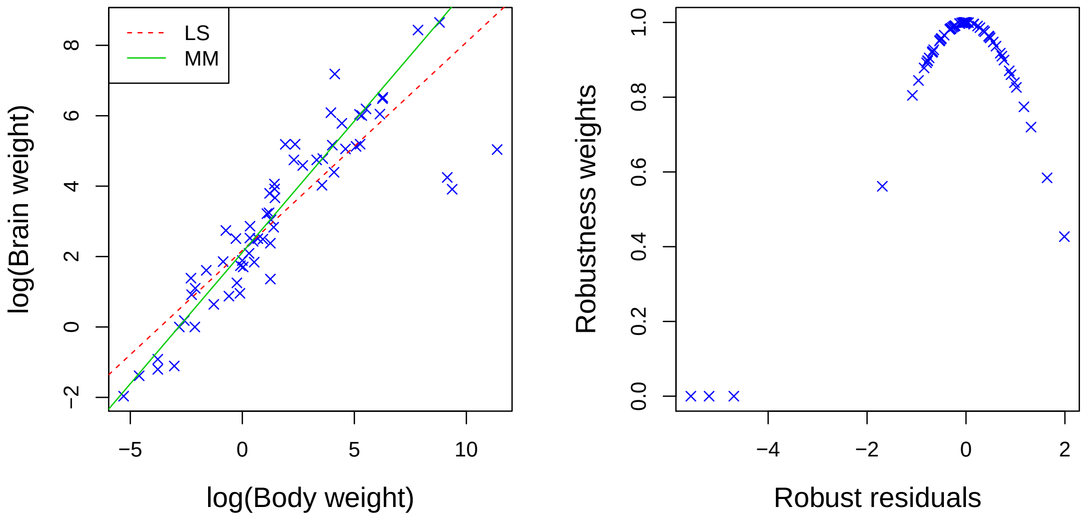


Figure 4.3: Body and brain weight of different animals. Left: linear regression using Least-squares and MM-regression; right: Residuals versus weights of the observations from MM-regression.

principal component analysis is carried out on the centered data, and the standard method uses a decomposition of the covariance matrix to find the principal components. Outliers or deviations from a model distribution can lead to very different results, and thus it is necessary to robustly estimate multivariate location and covariance. We will first think about desired properties of robust location and covariance estimators. Aside from robustness issues, a central property is affine equivariance which will be discussed below.

4.4.1 Affine equivariance

It is desirable that location and covariance estimates respond in a mathematically convenient form to certain transformations of the data. One can define a transformation that is using a nonsingular $p \times p$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$ of length p to transform the p-dimensional observations $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}$ by $\boldsymbol{A}\boldsymbol{x}_{j.} + \boldsymbol{b}$. This transformation performs any desired nonsingular linear transformation of the original data. Thus, if $\boldsymbol{t}$ denotes a location estimator, it is requested that

$$\boldsymbol{t}(\boldsymbol{A}\boldsymbol{x}_{1.} + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_{n.} + \boldsymbol{b}) = \boldsymbol{A} \cdot \boldsymbol{t}(\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}) + \boldsymbol{b}, \tag{4.18}$$

and for a covariance estimator $\boldsymbol{C}$ we require

$$\boldsymbol{C}(\boldsymbol{A}\boldsymbol{x}_{1.} + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_{n.} + \boldsymbol{b}) = \boldsymbol{A} \cdot \boldsymbol{C}(\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}) \cdot \boldsymbol{A}^{\top}. \tag{4.19}$$

Location and covariance estimators that fulfil (4.18) and (4.19) are called *affine equivariant* estimators. These estimators transform properly under changes of the origin, the scale, or under rotations.

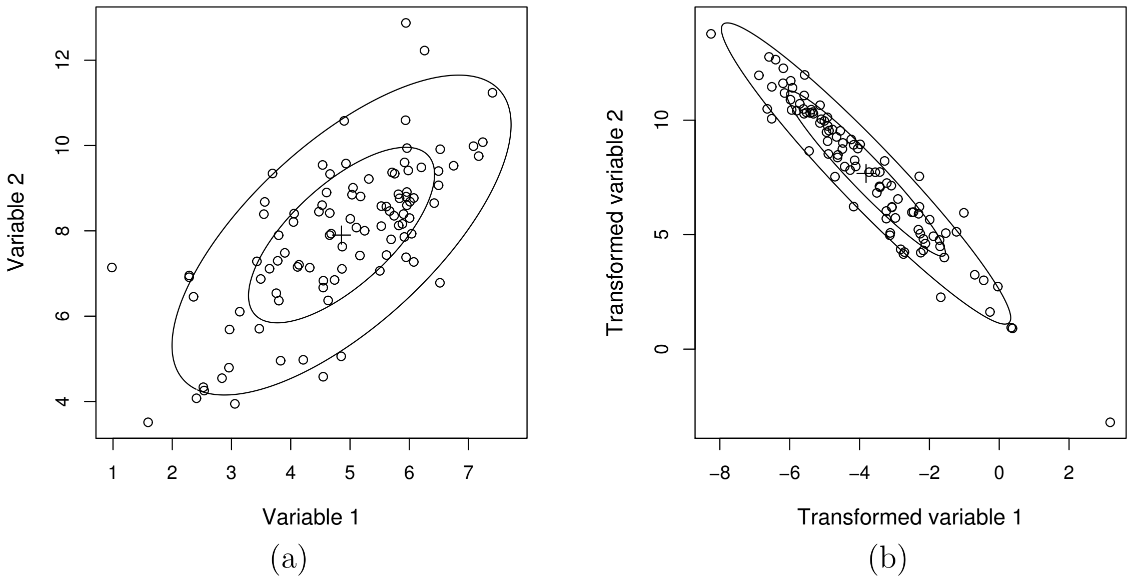


Figure 4.4: Bivariate data with estimated location (+) and covariance matrix (visualised by tolerance ellipses); plotted are (a) the original data and (b) the transformed data, together with the transformed estimates.

Figure 4.4a shows a bivariate data set where the location (+) was estimated by the arithmetic mean and the covariance by the sample covariance matrix. The latter is visualized by so-called *tolerance ellipses*: In case of normally distributed data the tolerance ellipses would contain a certain percentage of data points around the center and according to the covariance structure. Here we show the 50% and 90% tolerance ellipses. Figure 4.4b pictures the data after applying the transformation

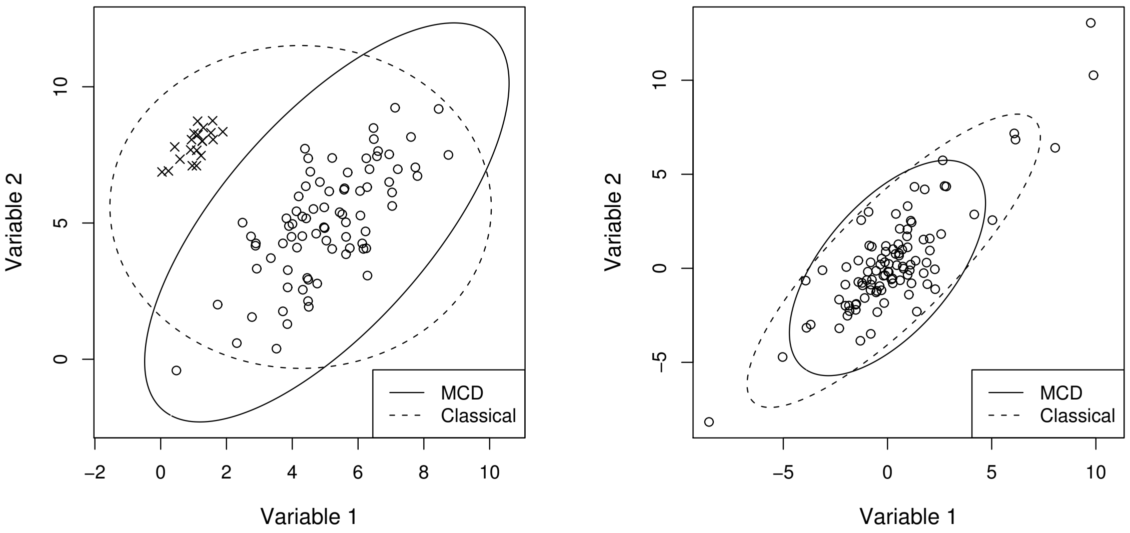
$$\boldsymbol{A}\boldsymbol{x}_{j.} + \boldsymbol{b} = \begin{pmatrix} -2 & 3 \\ 1 & -1 \end{pmatrix} \boldsymbol{x}_{j.} + \boldsymbol{b}$$

to each data point. The location and covariance estimates were not recomputed for the transformed data but were transformed according to the Equations (4.18) and (4.19). It is obvious from the figure that the transformed estimates are the same as if they would have been derived directly from the transformed data. Note that the transformation matrix $\boldsymbol{A}$ is close to singularity because the spread of the data becomes very small in one direction. The property of affine equivariance is only valid for nonsingular transformation matrices.

Note that the coordinate-wise median as a robust location estimator is not affine equivariant. Also a robust covariance estimate based on pairwise robust covariances would not be affine equivariant.

4.4.2 The MCD estimator

An estimator of multivariate location and covariance which is affine equivariant and has high breakdown point is the *Minimum Covariance Determinant* (MCD)



(a) (b)

Figure 4.5: Tolerance ellipses (97.5%) based on the MCD estimator and on the classical sample mean and sample covariance matrix for (a) bivariate normally distributed data with outliers and (b) bivariate T_2 (heavy tailed) distributed data.

estimator. The idea behind this estimator is in fact related to the LTS estimator. Here, one is searching for those h data points for which the determinant of the empirical covariance matrix is minimal. The location estimator $\boldsymbol{t}$ is the mean of these h observations, and the covariance estimator $\boldsymbol{C}$ is given by the covariance matrix with the smallest determinant, but multiplied by a constant to obtain consistency for normal distribution. The parameter h determines the robustness but also the efficiency of the resulting estimator. The highest possible breakdown point can be achieved if $h \approx n/2$ is taken, but this choice leads to a low efficiency. On the other hand, for higher values of h the efficiency increases but the breakdown point decreases. Therefore, a compromise between efficiency and robustness is considered in practice.

Figure 4.5 shows a comparison between the MCD estimator and classical location and covariance estimation for two simulated data sets. The covariance estimates are visualised by 97.5% tolerance ellipses, the location estimates are the centres of the ellipses. In Figure 4.5a a bivariate normally distributed data set is used, where 20% of the data points are generated with a different mean and covariance. Note that these deviating data points cannot be identified as outliers by inspecting the projections on the coordinates. Not only the location estimate is influenced by the deviating points, but especially the covariance structure. The data coming from the outlier distribution are inflating the tolerance ellipse based on the classical estimators while that based on the MCD is much more compact and reflects the structure of the majority of data.

The second example shown in Figure 4.5b is simulated from a bivariate T distribution with 2 degrees of freedom with a certain covariance structure. Also here the inflation of the classical ellipse due to some very distant points is visible.

We can also compute the correlation coefficient using the classical and robust covariance estimates. In the first example the classical correlation is 0.00 while the MCD gives a correlation of 0.70, which is also the result of the classical correlation for the data without outliers. For the second example we obtain a value of 0.84 for the classical correlation and 0.60 for the robust correlation.

4.4.3 Other robust covariance estimators

Multivariate S-estimators

Similar as in the regression context, see Equation (4.16), it is possible to define S-estimators in the context of robust location and covariance estimation. The idea is to make the Mahalanobis distances small. The squared Mahalanobis distances are defined as

$$\mathrm{MD}^2(\boldsymbol{x}_{i.}, \boldsymbol{t}, \boldsymbol{C}) = (\boldsymbol{x}_{i.} - \boldsymbol{t})^\top \boldsymbol{C}^{-1} (\boldsymbol{x}_{i.} - \boldsymbol{t}) \qquad \text{for } i = 1, \ldots, n$$

for a location estimator $\boldsymbol{t}$ and a covariance estimator $\boldsymbol{C}$.

Small Mahalanobis distances can be achieved by using an M-estimator of scale $\hat{\sigma}$, and minimizing

$$\hat{\sigma}(\mathrm{MD}^2(\boldsymbol{x}_{1.}, \boldsymbol{t}, \boldsymbol{C}), \ldots, \mathrm{MD}^2(\boldsymbol{x}_{n.}, \boldsymbol{t}, \boldsymbol{C}))$$

under the restriction that the determinant $|\boldsymbol{C}| = 1$. This restriction avoids a degenerated solution for $\boldsymbol{C}$.

Multivariate MM-estimators

Like in robust regression, a drawback of S-estimators is that their asymptotic efficiency might be rather low. MM estimators for multivariate location and covariance combine both high breakdown point and high efficiency. The resulting estimators are affine equivariant and have bounded influence function. As for the previous estimators, the solution for the estimators can be found by an iterative algorithm.

Example 4.4.1 *The R package* `robustbase` *contains the data set* `phosphor` *with measurements of inorganic and organic Phosphorus in the soil. Figure 4.6 shows the data, together with a 97.5% tolerance ellipsoid. To compute a tolerance ellispoid, we assume multivariate normal distribution, and then the squared Mahalanobis distances follow a χ^2_p distribution, here $p = 2$. Fixing the quantile $\chi^2_{2;0.975}$ results in constant values of the squared Mahalanobis distance, and this results in the presented ellipse. In the left plot, the Mahalanobis distances have been computed with the MCD estimates of location and covariance, while in the right plot, the classical estimates arithmetic mean and sample covariance were used.*

Data points which are outside of the tolerance ellipse are potential multivariate outliers. Clearly, this is only reliable when using robust estimators.

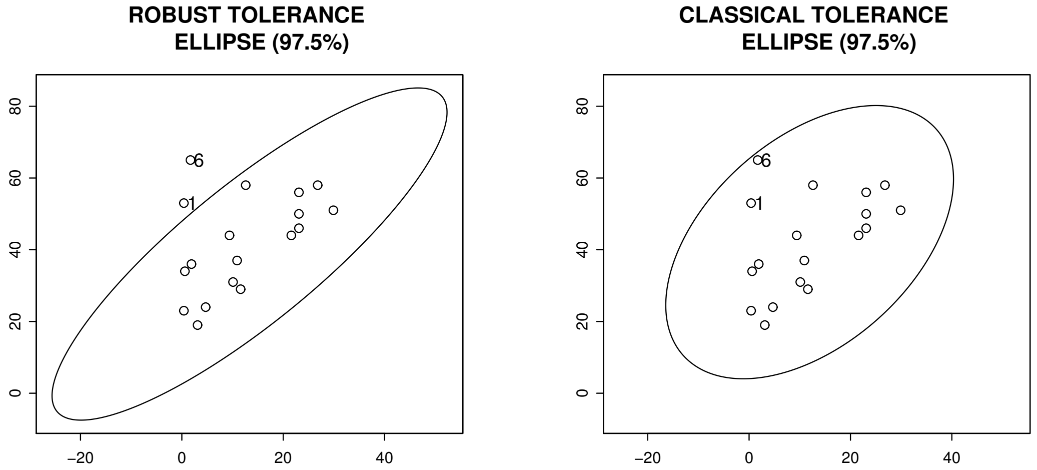


Figure 4.6: Phosphorus data set: 97.5% tolerance ellipses based on the MCD (left) and on classical estimators (right).

Note that for p-dimensional data with $p > 2$ it is hardly possible to show tolerance ellipsoids, but one can still perform multivariate outlier detection, because Mahalanobis distances are always univariate. Multivariate outliers would be identified as follows:

- With the observations $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}$, compute the MCD estimator of location $\boldsymbol{t}$ and covariance $\boldsymbol{C}$.
- Compute the squared Mahalanobis distances $\mathrm{MD}^2(\boldsymbol{x}_{i.}, \boldsymbol{t}, \boldsymbol{C})$ for every observation.
- An observations is a multivariate outlier if $\mathrm{MD}^2(\boldsymbol{x}_{i.}, \boldsymbol{t}, \boldsymbol{C}) > \chi^2_{p;0.975}$.

4.5 Back to regression: robust regression diagnostics

LS regression is sensitive to outliers. These could be outliers in the response, so-called “vertical” outliers, or outliers in the explanatory variables, so-called “leverage points”. The latter are particularly influential if they have large residual (bad leverage points).

In LS regression, the hat matrix $\boldsymbol{H}$ is used to identify leverage points. The hat matrix is defined as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, and the LS fit is obtained as $\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{H}\boldsymbol{y}$.

This means that

$$\widehat{y}_{i_{LS}} = (h_{i1} \cdots \underline{h_{ii}} \cdots h_{in}) \begin{pmatrix} y_1 \\ \vdots \\ \underline{y_i} \\ \vdots \\ y_n \end{pmatrix} ,$$

and thus h_{ij} can be interpreted as the effect of the j-th observation y_j on $\widehat{y}_{i_{LS}}$. Particularly, h_{ii} indicates the influence of the i-th observation on its own fit $\widehat{y}_{i_{LS}}$.

Properties of h_{ii}:

- $\boldsymbol{H}^\top = \boldsymbol{H}$, $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$
  $\mathrm{tr}(\boldsymbol{H}) = \mathrm{tr}[\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}] = \mathrm{tr}(\boldsymbol{I}_{p+1}) = p+1$
  It follows that the average of the h_{ii} elements is equal to $\frac{p+1}{n}$.

- $h_{ii} = (\boldsymbol{H}\boldsymbol{H})_{ii} = \sum_{j=1}^n h_{ij}h_{ji} = \sum_{j=1}^n h_{ij}^2 \quad \Longleftrightarrow \quad h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$
  It follows that $0 \leq h_{ii} \leq 1$.
  If $h_{ii} = 0$, then $h_{ij} = 0$ for all j, and then we have no other contribution of $\boldsymbol{y}$ on $\widehat{y}_{i_{LS}}$.
  If $h_{ii} = 1$, then $h_{ij} = 0$ for all $j \neq i$, and one obtains an exact fit $\widehat{y}_{i_{LS}} = y_i$.

If there is an exact fit, one can assume that this was caused by a leverage point, because an extremely strong influence was caused to the own estimation. In general, one needs to be careful if there are large values of h_{ii}. As a rule of thumb one could define the threshold $h_{ii} > 2 \cdot \frac{p+1}{n}$ for the identification of leverage points.

Example 4.5.1 *The R package* **`robustbase`** *contains the data set* **`hbk`** *(for Hawkins, Bradu, Kass), which is a simulated data set to test regression estimators for their robustness. The data set consists of 75 observations in four dimensions. The first 14 observations are outliers in the x-space (leverage points), created in two groups: 1-10 are "bad" leverage points with large residuals, and 11-14 are "good" leverage points with small residuals.*

Figure 4.7 shows the diagonal elements of the hat matrix against the index of the observations. The dashed line is at twice the average, i.e. $2 \cdot \frac{p+1}{n} = 2 \cdot \frac{4}{75} = 0.107$. *Only observations 12, 13 and 14 exceed this threshold, and they would thus be identified as leverage points. However, all the bad leverage points have regular values* h_{ii}. *This phenomenon that the "true" outliers are not revealed is called masking effect.*

What is the reason for the masking effect? In fact, one can show that the following relationship holds in case of regression with intercept:

$$h_{ii} = \frac{1}{n-1}\mathrm{MD}_i^2 + \frac{1}{n} \tag{4.20}$$

Here,

$$\mathrm{MD}_i^2 = (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})$$

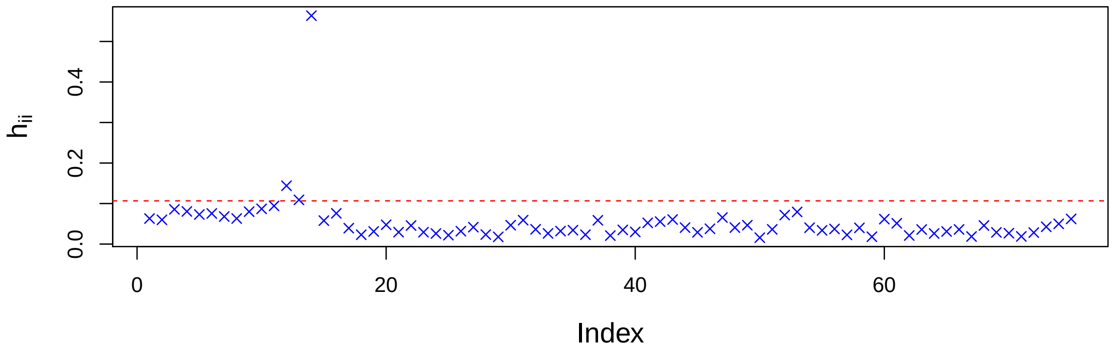


Figure 4.7: Diagonal elements of the hat matrix of the Hawkins-Bradu-Kass data, with the threshold indicating "extreme" values.

for $i = 1, \dots, n$, and $\boldsymbol{x}_{i.} = (x_{i1}, \dots, x_{ip})^\top$ (without "1" for the intercept). Thus, the diagonal elements of the hat matrix are proportional to the "classical" (non-robust) squared MDs. Clearly, non-robust diagnostics can be spoiled by outliers. The outliers may also lead to an inflation of the empirical covariance matrix estimation, and this hides (masks) the outliers. In fact, this inflation effect has been observed already in Figure 4.6.

A robust alternative to the hat matrix elements h_{ii} for leverage diagnostics can be inspired by Equation (4.20): one can simply use Mahalanobis distances based on robust estimates of location and covariance. For instance, the robustified (squared) distances

$$\text{MD}_i^2 = (\boldsymbol{x}_{i.} - \boldsymbol{t})^\top \boldsymbol{C}^{-1} (\boldsymbol{x}_{i.} - \boldsymbol{t})$$

with the estimators of location $\boldsymbol{t}$ and covariance $\boldsymbol{C}$ of the MCD estimator are suitable. A threshold for outlyingness is $\chi^2_{p;0.975}$.

Figure 4.8 shows the robustified MDs together with the outlier threshold. Now it is immediate which observations are leverage points (those with index 1-14).

Remark: Note that the values h_{ii} only consider the x-values (explanatory variables), but not the response y. Thus, it is not possible to distinguish between good and bad leverage points.

Regression diagnostics

The different types of outliers we would like to distinguish are illustrated in the case of simple linear regression schematically in Figure 4.9. These are:

- Regular observations: $\boldsymbol{x}_{i.}$ is in the usual data range, and y_i fits to the model.
- Vertical outliers: $\boldsymbol{x}_{i.}$ is in the usual data range, but y_i does not fit to the model.
- Good leverage points: $\boldsymbol{x}_i$ is an outlier, thus unusual in the x-space, but y_i fits to the model.

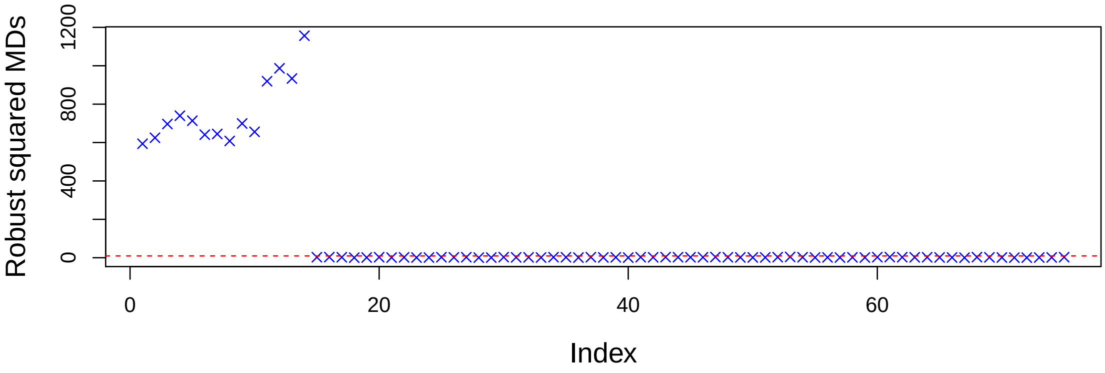


Figure 4.8: Robust squared Mahalanobis distances for the Hawkins-Bradu-Kass data, with the threshold indicating “extreme” values.

- Bad leverage points: $\boldsymbol{x}_i$ is an outlier, thus unusual in the x-space, and y_i does not fit to the model.

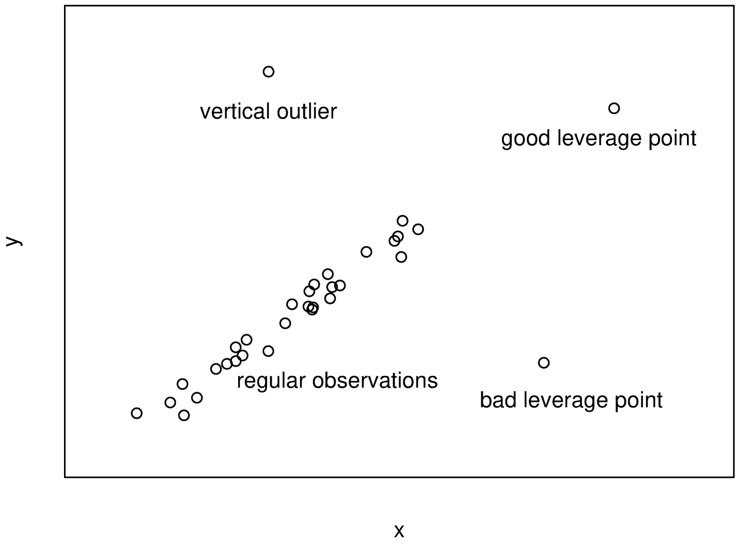


Figure 4.9: Different types of outliers, illustrated in simple linear regression.

In general, good leverage points have the advantage that they are along the regression line (hyperplane) and thus they even allow for a more accurate estimation of the regression parameters. Bad leverage points can have a strong impact on the (LS) estimation, and they can even lead to a leverage of the regression line (hyperplane).

The regression diagnosic plot allows to distinguish these 4 types of observations. Outlyingness in the x-space can be recognized by robust Mahalanobis distances. Large (absolute) residuals can be recognized as scaled deviations in the y-space from the robust regression fit. Thus, the robust residuals need to be scaled by a robust estimate of the residual scale. Then one can argue according to the normal theory that scaled residuals outside $[-2, 2]$ or $[-2.5, 2.5]$ are extreme and thus very unusual.

Figure 4.10 shows the resulting regression diagnostic plot, where the robust Mahalanobis distances are plotted against the robust scaled residuals. The horizontal lines are at the thresholds ±2.5 for the scaled residuals, and the vertical line is at the threshold based on the chi-square quantile.

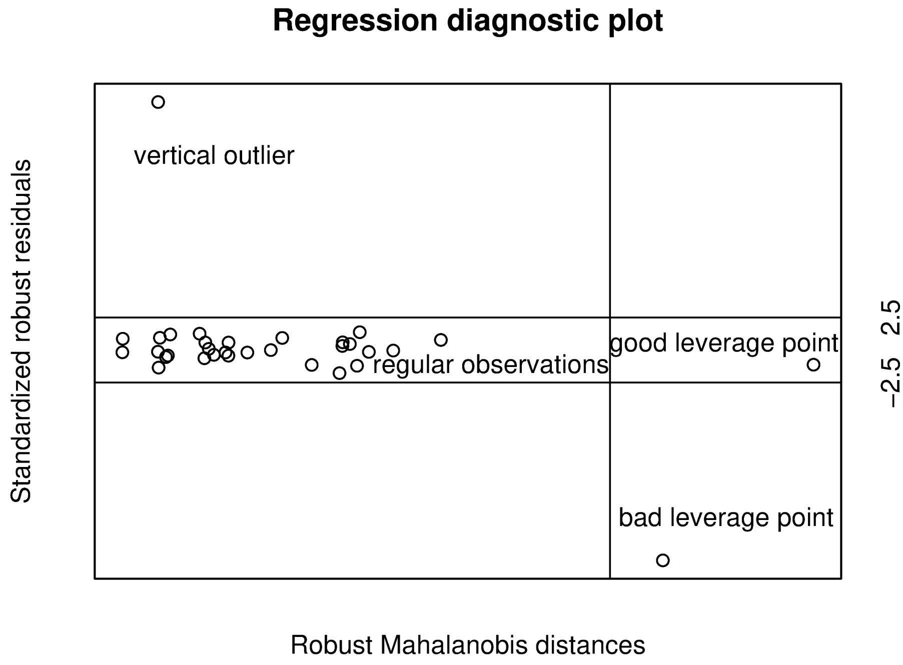


Figure 4.10: Identication of the four different categories of observations in the regression diagnostic plot.

Example 4.5.2 *We go back to the Hawkins-Bradu-Kass data set and compute the MM-estimator of regression to create the regression diagnostic plot:*

```
library(robustbase)
data(hbk)
res.rob <- lmrob(Y~., data=hbk)
plot(res.rob,which=1)
```

The plot is shown in Figure 4.11. It is not so clearly seen from the indexes, but the first 10 observations are the bad leverage points with large robust distance and large residual, and the observations with index 11-14 are the good leverage points. This corresponds to how the data have been generated.

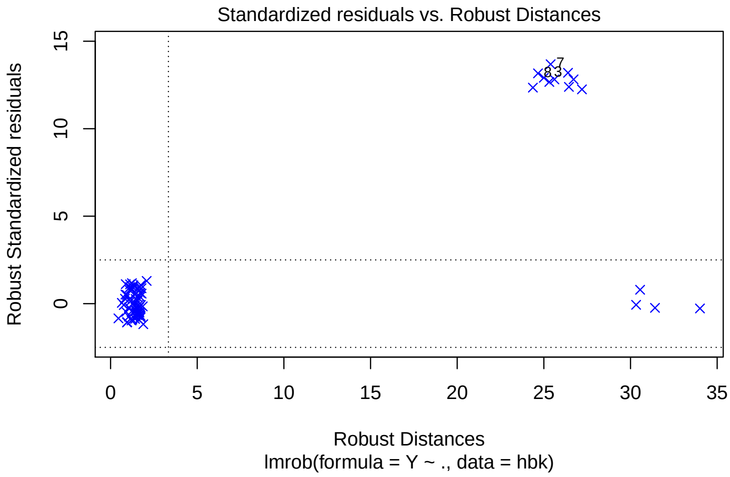


Figure 4.11: Identification of the four different categories of observations in the regression diagnostic plot.

4.6 Robust multivariate regression

Consider again the multivariate regression problem with the observations $\boldsymbol{y}_{i.} = (y_{i1}, \ldots, y_{im})^\top$, $\boldsymbol{x}_{i.} = (1, x_{i1}, \ldots, x_{ip})^\top$, and the error terms $\boldsymbol{e}_{i.} = (e_{i1}, \ldots, e_{im})^\top$, and the model

$$\boldsymbol{y}_{i.} = \boldsymbol{B}^\top \boldsymbol{x}_{i.} + \boldsymbol{e}_{i.} \ ,$$

for $i = 1, \ldots, n$, see also Equation (3.13). We are interested in robustly estimating the $(p+1) \times m$ matrix $\boldsymbol{B}$ of regression coefficients, as well as the covariance matrix $\boldsymbol{\Sigma}$ of the error terms. For this purpose, one can use multivariate S-estimators, see Section 4.4.3, and Van Aelst and Willems (2005). Here, the Mahalanobis distances are based on the residuals $\boldsymbol{r}_{i.} = \boldsymbol{y}_{i.} - \boldsymbol{B}^\top \boldsymbol{x}_{i.}$ and on an estimator $\boldsymbol{C}$ of $\boldsymbol{\Sigma}$. Thus, one has to minimize

$$\hat{\sigma}(\boldsymbol{r}_{1.}^\top \boldsymbol{C}^{-1} \boldsymbol{r}_{1.}, \ldots, \boldsymbol{r}_{n.}^\top \boldsymbol{C}^{-1} \boldsymbol{r}_{n.})$$

using an M-estimator of scale $\hat{\sigma}$, under the contraint $|\boldsymbol{C}| = 1$.

Example 4.6.1 *The R package* ***FRB*** *contains the function* ***FRBmultiregS()*** *to compute the estimators described above. "FRB" is short for "fast and robust bootstrap", and this technique allows to do robust statistical inference. The package also contains the data set* ***schooldata*** *which we use for illustration. There are 70 observations with scores of 3 different tests (reading, mathematics, selfesteem), and they should be predicted by using 5 explanatory variables, see help pages for details.*

```
library(FRB)
data(schooldata)
Sres <- FRBmultiregS(cbind(reading,mathematics,selfesteem)~.,
                     data=schooldata, bdp = 0.25)
```

The parameter **`bdp`** *defines the breakdown point by setting the appropriate constant b for the M-estimator of scale. The estimated regression parameters are:*

```
> Sres

Coefficients:
            reading mathematics selfesteem
(Intercept)  1.9630      2.5348     0.0775
education    0.1223      0.0117    -0.0471
occupation   4.9368      6.1086     2.1510
visit        0.0721      0.0201     0.2368
counseling  -0.7878     -0.8265    -0.0833
teacher     -0.1920     -0.2937     0.0219
```

Statistical inference based on using the robust bootstrap is obtained for each response as follows:

```
> summary(Sres)

Response reading:

Coefficients:
            Estimate   Std.Error   p-value
(Intercept)   1.9630      1.0458   0.06471   .
education     0.1223      0.0744   0.06466   .
occupation    4.9368      1.2811   0.00000 ***
visit         0.0721      0.3284   0.80437
counseling   -0.7878      0.2252   0.00392  **
teacher      -0.1920      0.1758   0.17079

Response mathematics:

Coefficients:
            Estimate   Std.Error   p-value
(Intercept)   2.5348      1.2084    0.0273   *
education     0.0117      0.0857    0.9094
occupation    6.1086      1.3151    0.0000 ***
visit         0.0201      0.3150    0.9041
counseling   -0.8265      0.3334    0.0132   *
teacher      -0.2937      0.2194    0.0931   .

Response selfesteem:

Coefficients:
            Estimate   Std.Error   p-value
(Intercept)   0.0775      0.3859     0.981
```

```
education    -0.0471      0.0406     0.255
occupation    2.1510      0.5430     0.000 ***
visit         0.2368      0.0796     0.000 ***
counseling   -0.0833      0.1260     0.458
teacher       0.0219      0.0437     0.662

Robust residual scale: 2.2

Error covariance matrix estimate:
            reading mathematics selfesteem
reading       14.49        14.4       2.55
mathematics   14.44        21.5       2.60
selfesteem     2.55         2.6       1.55
```

Interestingly, the value for "self esteem" is significantly affected by the number of parental visits (the more, the higher).

Finally, we show a diagnostic plot in Figure 4.12 which is obtained by:

```
> diagplot(Sres)
```

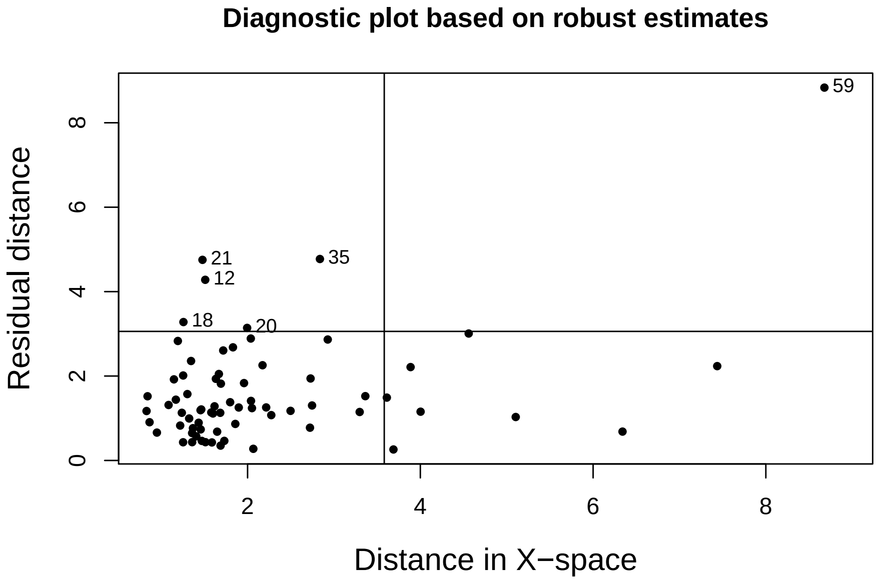


Figure 4.12: Diagnostic plot for robust multivariate regression with the school data.

Figure 4.12 shows Mahalanobis distances of the x-variables versus those from the residuals. Observation 59 is revealed as a bad leverage point. When looking closer at this observation one can see that it has exceptionally high values in all the variables.

References

N.R. Draper and H. Smith. *Applied Regression Analysis.* Wiley & Sons, New York, 1981.

D.M. Hawkins, D. Bradu, and G.V. Kass. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26, 197-208, 1984.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods.* John Wiley & Sons, Chichester, 2006.

P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871-880, 1984.

P.J. Rousseeuw. Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (eds.), *Mathematical Statistics and Applications*, Volume B. Akadémiai Kiadó, Budapest, pp. 283-297, 1985.

P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

G.A.F. Seber. *Multivariate Observations.* John Wiley & Sons, New York, 1984.

S. Van Aelst and G. Willems. Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica*, 15, 981-1001, 2005.

Chapter 5

Principal component analysis

5.1 Introduction

Principal component analysis, or, for short, PCA, goes back to the famous statistician Karl Pearson (1901), and it was developed by Harold Hotelling (1933). Its aim is to describe complex relationships in given data in a simpler form. The data are represented by linear combinations of specific components in a way to preserve as much information as possible. Thus, the dimensionality is reduced to the number of those components. PCA is mainly known for its ability to reduce the dimension of the data space by keeping the information loss low.

PCA is one of the most important methods in multivariate statistics – one could say it is the mother (father, grand-m/f, etc.) of all multivariate methods. Its concept is relatively simple, it allows for a powerful insight into the data structure, and it is the basis for many other multivariate methods.

5.2 Definition of principal components

Let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ be a p-dimensional random vector with expectation $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance matrix

$$\boldsymbol{\Sigma} = E\Big[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\Big] \ . \tag{5.1}$$

Further, $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ is a $(p \times p)$ matrix with fixed values (non-random), with the constraint that its column vectors $\boldsymbol{\gamma}_i$ are unitary vectors, i.e. $\boldsymbol{\gamma}_i^\top \boldsymbol{\gamma}_i = 1$ for $i = 1, \ldots, p$. Moreover, different columns of $\boldsymbol{\Gamma}$ are orthogonal, i.e. $\boldsymbol{\gamma}_i^\top \boldsymbol{\gamma}_j = 0$ for $i \neq j$. This implies that $\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^{-1}$.

Consider the linear transformation

$$\boldsymbol{z} = \boldsymbol{\Gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.2}$$

or, expressed in components,

$$z_i = \boldsymbol{\gamma}_i^\top (\boldsymbol{x} - \boldsymbol{\mu}) \qquad \text{for} \qquad i = 1, \ldots, p \ . \tag{5.3}$$

The result of the above transformation is a new random variable $\boldsymbol{z}$ of dimension p. The variance of z_i $(i = 1, \ldots, p)$ is

$$\mathrm{Var}(z_i) = E\Big[\boldsymbol{\gamma}_i^\top(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\boldsymbol{\gamma}_i\Big] = \boldsymbol{\gamma}_i^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_i \ . \tag{5.4}$$

So far it is not clear which matrix $\boldsymbol{\Gamma}$ we should choose – in spite of the constraints, there are still infinitely many possibilities. Since we are interested in preserving information content, we should look at the variance of the transformed variables. In other words, we would like to obtain such a transformation which maximizes the variances of the components of $\boldsymbol{z}$.

Considering also the contraints on $\boldsymbol{\Gamma}$, we can mathematically formulate a maximization problem in terms of Lagrange optimization.

First principal component:

For $i = 1$ we would like to obtain a component z_1 such that $\mathrm{Var}(z_1)$ is maximized, and $\boldsymbol{\gamma}_1^\top\boldsymbol{\gamma}_1 = 1$. The corresponding Lagrangian problem can be stated as:

$$\phi_1 = \boldsymbol{\gamma}_1^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 - a_1(\boldsymbol{\gamma}_1^\top\boldsymbol{\gamma}_1 - 1) \tag{5.5}$$

The partial derivatives with respect to the unknowns $\boldsymbol{\gamma}_1$ are set equal to zero, and we obtain

$$\frac{\partial\phi_1}{\partial\boldsymbol{\gamma}_1} = 2\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 - 2a_1\boldsymbol{\gamma}_1 = \boldsymbol{0} \tag{5.6}$$

or

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 = a_1\boldsymbol{\gamma}_1 \ . \tag{5.7}$$

This is an eigenvector/eigenvalue problem, and the solution is that the unknown coefficient vector $\boldsymbol{\gamma}_1$ is an eigenvector of the covariance matrix $\boldsymbol{\Sigma}$ to the eigenvalue a_1.

Our problem, however, is that $\boldsymbol{\Sigma}$ has p eigenvectors in total. Which one should we take?

With Equation (5.7) we can see that

$$\mathrm{Var}(z_1) = \boldsymbol{\gamma}_1^\top(\boldsymbol{\Sigma}\boldsymbol{\gamma}_1) = \boldsymbol{\gamma}_1^\top(a_1\boldsymbol{\gamma}_1) = a_1\boldsymbol{\gamma}_1^\top\boldsymbol{\gamma}_1 = a_1,$$

and since we want to maximize variance, we take that eigenvector $\boldsymbol{\gamma}_1$ corresponding to the largest eigenvalue a_1. Component z_1 now is denoted as *first principal component (PC)*, and $\boldsymbol{\gamma}_1$ is the direction of this component.

Second principal component:

In the next step, for $i = 2$, we again want to maximize variance, more clearly, $\mathrm{Var}(z_2)$ should be maximized under the constraint $\boldsymbol{\gamma}_2^\top\boldsymbol{\gamma}_2 = 1$. In addition we also want that z_1 and z_2 are uncorrelated (in order to uncover new information in z_2). The latter condition means:

$$Cov(z_1, z_2) = Cov(\boldsymbol{\gamma}_1^\top(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{\gamma}_2^\top(\boldsymbol{x} - \boldsymbol{\mu})) = Cov(\boldsymbol{\gamma}_1^\top\boldsymbol{x}, \boldsymbol{\gamma}_2^\top\boldsymbol{x}) = \boldsymbol{\gamma}_1^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 =$$

$$= \boldsymbol{\gamma}_2^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2^\top a_1\boldsymbol{\gamma}_1 = a_1\boldsymbol{\gamma}_2^\top\boldsymbol{\gamma}_1 = 0$$

Since $a_1 \neq 0$, the condition of uncorrelated components is equivalent to orthogonality of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. That's a remarkable fact which you don't easily get for other methods.

Now we can again formulate the Lagrangian problem,

$$\phi_2 = \boldsymbol{\gamma}_2^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_2 - a_2(\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_2 - 1) - b\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_1 \tag{5.8}$$

with the Lagrange coefficients a_2 and b. The partial derivatives with respect to the unknowns $\boldsymbol{\gamma}_2$ are set equal to zero, which yields:

$$\frac{\partial \phi_2}{\partial \boldsymbol{\gamma}_2} = 2\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 - 2a_2\boldsymbol{\gamma}_2 - b\boldsymbol{\gamma}_1 = \mathbf{0} \tag{5.9}$$

Multiplication from the left-hand side with $\boldsymbol{\gamma}_1^\top$ result in

$$2\boldsymbol{\gamma}_1^\top \boldsymbol{\Sigma}\boldsymbol{\gamma}_2 - 2a_2\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_2 - b\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 = 0 - 0 - b \cdot 1 = 0 \ ,$$

and thus $b = 0$. Therefore we can reduce (5.9) to

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 = a_2\boldsymbol{\gamma}_2 \ ,$$

and thus $\boldsymbol{\gamma}_2$ is eigenvector to $\boldsymbol{\Sigma}$ to the next largest eigenvalue a_2, and z_2 is called *second PC*.

Further principal components:

The k-th PC $(2 < k \leq p)$ is defined in an analogous way: maximize $\mathrm{Var}(z_k)$ under the constraints $\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_k = 1$ and uncorrelatedness to all previous PCs, i.e. $Cov(z_k, z_j) = 0$ for $k > j$, which is equivalent to orthogonality $\boldsymbol{\gamma}_k^\top \boldsymbol{\gamma}_j = 0$.

As expected, the solution is that $\boldsymbol{\gamma}_k$ is eigenvector of $\boldsymbol{\Sigma}$ to the k-th largest eigenvalue a_k.

Let us now collect all the eigenvectors as columns in the matrix $\boldsymbol{\Gamma}$, thus $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$, and the corresponding eigenvalues $a_1, \ldots, a_p$ sorted in ascending order $(a_1 \geq a_2 \geq \ldots \geq a_p \geq 0)$ in the diagonal of the matrix $\boldsymbol{A}$, thus $\boldsymbol{A} = \mathrm{Diag}(a_1, \ldots, a_p)$. Then we can express the PC solution is matrix form as

$$\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{A} \tag{5.10}$$

or also as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^\top$, see Theorem 1.4.3 (*spectral decomposition*).

Let us go back again to our initial transformation (5.2). With our definition of $\boldsymbol{\Gamma}$, this linear transformation

$$\boldsymbol{z} = \boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.11}$$

is know as *principal component transformation*, and the i-th element z_i of the vector $\boldsymbol{z}$ is called *i-th principal component*.

Remark: We can see that the PC transformation is not scale invariant, which means that the resulting PCs depend on the scale (units) of the variables. If we would first scale (and probably also center) the variables to mean 0 and variance 1,

$$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} \quad \text{for} \quad i = 1, \ldots, p, \tag{5.12}$$

we would in general obtain different PCs, since then we would perform the eigenvector/eigenvalue decomposition on the correlation matrix, and not on the covariance matrix. Therefore, if scale invariance is desired (and in most applications it is), the variables need to be standardized first.

Since our random variable $\boldsymbol{x}$ has been centered, we obtain that the expectation of the PCs is zero:

$$E(\boldsymbol{z}) = \boldsymbol{\Gamma}^\top \Big[E(\boldsymbol{x} - \boldsymbol{\mu})\Big] = \boldsymbol{0} \tag{5.13}$$

The covariance matrix of the PCs is

$$\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{\Gamma}^\top \mathrm{Cov}(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\top \boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{A} \ . \tag{5.14}$$

Indeed, this is how we constructed the PCs: The variance of the i-th PC is a_i, the i-the eigenvalue of $\boldsymbol{\Sigma}$, and different PCs are uncorrelated, since $\boldsymbol{A}$ is a diagonal matrix. The total variance of all PCs is the sum of all eigenvalues. This is of course equal to the total variance of $\boldsymbol{x}$.

The matrix $\boldsymbol{\Gamma}$ obviously relates $\boldsymbol{x}$ and $\boldsymbol{z}$; it is also called *loadings matrix.* Its elements γ_{ij} reflect the influence of x_i on z_j.

A further interesting measure for the relationship between $\boldsymbol{x}$ and $\boldsymbol{z}$ is the correlation. In particular, the squared correlation represents a measure of determination, the variation of z_j explained by x_i. The covariance between the variables and the PCs is

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{z}) &= E\Big[(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{z}^\top\Big] \\ &= E\Big[\boldsymbol{\Gamma}\boldsymbol{z}\boldsymbol{z}^\top\Big] \\ &= \boldsymbol{\Gamma}\boldsymbol{A} \end{aligned} \tag{5.15}$$

or

$$\mathrm{Cov}(x_i, z_j) = \gamma_{ij} a_j \qquad \text{for} \qquad i, j = 1, \ldots, p \ . \tag{5.16}$$

Thus, the correlation between variables and PCs is

$$\mathrm{Cor}(x_i, z_j) = \lambda_{ij} = \frac{\mathrm{Cov}(x_i, z_j)}{\sqrt{\mathrm{Var}(x_i)}\sqrt{\mathrm{Var}(z_j)}} = \frac{\gamma_{ij} a_j}{\sigma_{ii}^{\frac{1}{2}} a_j^{\frac{1}{2}}} = \gamma_{ij}\sqrt{\frac{a_j}{\sigma_{ii}}} \tag{5.17}$$

or

$$\mathrm{Cor}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\Lambda} = \Big(\mathrm{Diag}(\boldsymbol{\Sigma})\Big)^{-\frac{1}{2}} \boldsymbol{\Gamma}\boldsymbol{A}^{\frac{1}{2}} \ . \tag{5.18}$$

5.3 Geometric interpretation of PCs

In Chapter 1 we have already defined the Mahalanobis distance (MD). Setting its square equal to a constant, see Equation (1.17), i.e.

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) := c^2 \quad , \tag{5.19}$$

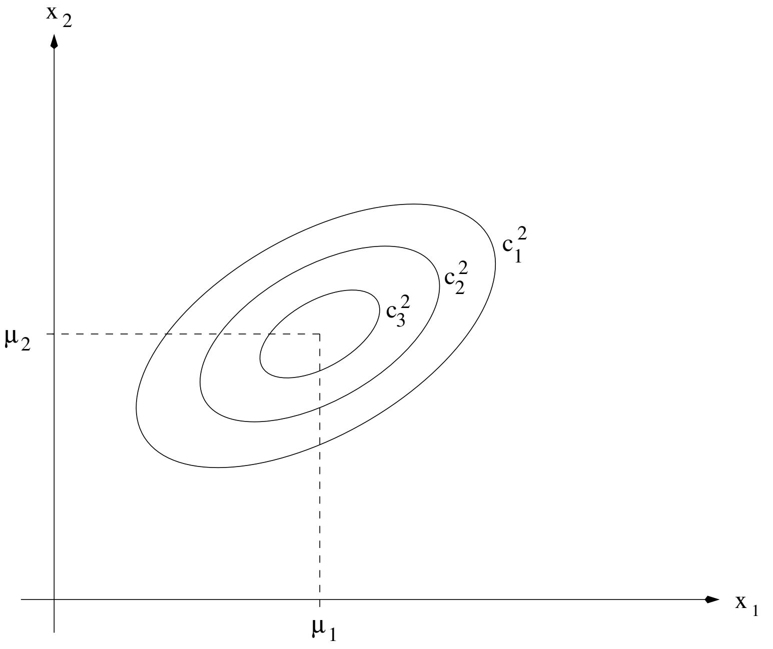


Figure 5.1: Ellipses with constant densities for the 2-dimensional normal distribution.

describes an ellipsoid in the p-dimensional space with center $\boldsymbol{\mu}$. Therefore, the density of the p-variate normal distribution, which contains the squared MD in its exponent, is constant along these ellipsoids. Figure 5.1 visualizes this in the 2-dimensional case, for different constants.

Using the PC transformation $\boldsymbol{z} = \boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})$, we can re-express the squared MD:

$$\begin{aligned} c^2 &= (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Gamma} \boldsymbol{A}^{-1} \boldsymbol{\Gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \\ &= \left[\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})\right]^\top \boldsymbol{A}^{-1}\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{z}^\top \boldsymbol{A}^{-1} \boldsymbol{z} \\ &= \sum_{i=1}^{p} \frac{z_i^2}{a_i} = \frac{z_1^2}{a_1} + \frac{z_2^2}{a_2} + \ldots + \frac{z_p^2}{a_p} \quad . \end{aligned}$$

The components $z_1, \ldots, z_p$ of $\boldsymbol{z}$ thus represent the main axes of the ellipsoids, as shown in Figure 5.2 for the 2-dimensional case. The first PC z_1 is along the largest expansion of the ellipsoid, the second PC z_2 is along the largest expansion which is orthogonal to the direction of the first PC, etc.

Re-expressing the MD by means of PCs also has another interesting aspect: MDs of observations to the center in the variable space turn to Euclidean distances in the PC space, because the covariance matrix of the PCs is a diagonal matrix. The PC transformation thus allows for a "distorted" data representation where we can think in terms of the usual Euclidean distances, which our "Euclidean eyes" clearly prefer.

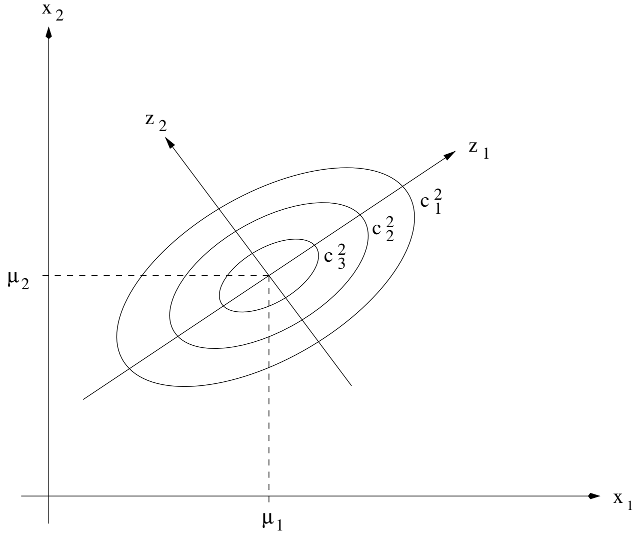


Figure 5.2: Geometrical view of the PCs in 2 dimensions.

5.4 PCs based on data

Consider an $n \times p$ data matrix $\boldsymbol{X}$ with n observations and p variables. For the PCA transformation we need to estimate the expectation vector and the covariance matrix. This can be done by the empirical estimates, the sample mean and the sample covariance matrix,

$$\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i.} \ , \tag{5.20}$$

and

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \ , \tag{5.21}$$

where $\boldsymbol{x}_{i.}$ is the i-th row of $\boldsymbol{X}$.

In analogy to Equation (5.11), the sample principal components are computed by

$$\boldsymbol{Z} = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\Gamma}} \tag{5.22}$$

or

$$\boldsymbol{z}_j = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\gamma}}_j \qquad \text{for} \qquad j = 1, \ldots, p \ . \tag{5.23}$$

Here, $\mathbf{1}$ is a column vector with n entries of 1, and $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ is the matrix of eigenvectors of $\boldsymbol{S}$. Similar to the population definition of PCA, we obtain a decomposition

$$\hat{\boldsymbol{\Gamma}}^\top \boldsymbol{S}\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{A}} \ , \tag{5.24}$$

where $\hat{\boldsymbol{A}} = \mathrm{Diag}(\hat{a}_1, \ldots, \hat{a}_p)$ is a diagonal matrix with the eigenvalues to the corresponding eigenvectors of $\boldsymbol{S}$, arranged in descending order.

The matrix $\boldsymbol{Z}$ of PCs has the same dimension as the data matrix $\boldsymbol{X}$. In fact, it just represents the data information in an orthgonally rotated coordinate system.

In this representation, the data are centered, and thus the original data center could not be reconstructed. The elements z_{ij} of the matrix $\boldsymbol{Z}$ are called (PC) *scores.*

In analogy to Equation (5.17) we can compute the correlations between the variables and the PCs as

$$\hat{\lambda}_{ij} = \frac{\hat{\gamma}_{ij}\hat{a}_j^{\frac{1}{2}}}{s_{ii}^{\frac{1}{2}}} \qquad \text{for} \qquad i,j = 1,\ldots,p \; , \tag{5.25}$$

where s_{ii} is the i-th diagonal element of $\boldsymbol{S}$. In matrix notation, this correlation is

$$\hat{\mathbf{\Lambda}} = \Big(\mathrm{Diag}(\boldsymbol{S})\Big)^{-\frac{1}{2}} \hat{\mathbf{\Gamma}} \hat{\boldsymbol{A}}^{\frac{1}{2}} \; . \tag{5.26}$$

Example 5.4.1 *Consider the exam data set* `scor` *from the R package* `bootstrap`*, with the points of 88 students in 5 different subjects. We are computing the PCs now:*

The arithmetic mean vector is

$$\bar{\boldsymbol{x}} = (39.0 \quad 50.6 \quad 50.6 \quad 46.7 \quad 42.3)^{\top} \; ,$$

and the empirical covariance matrix is

$$\boldsymbol{S} = \begin{pmatrix} 305.8 & 127.2 & 101.6 & 106.3 & 117.4 \\ & 172.8 & 85.2 & 94.7 & 99.0 \\ & & 112.9 & 112.1 & 121.9 \\ & & & 220.4 & 155.5 \\ & & & & 297.8 \end{pmatrix}$$

(lower half is just symmetric). With the spectral decomposition we obtain the matrix of eigenvectors

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.51 & -0.75 & -0.30 & 0.30 & 0.08 \\ 0.37 & -0.21 & 0.42 & -0.78 & 0.19 \\ 0.35 & 0.08 & 0.15 & 0.00 & -0.92 \\ 0.45 & 0.30 & 0.60 & 0.52 & 0.29 \\ 0.53 & 0.55 & -0.60 & -0.18 & 0.15 \end{pmatrix},$$

and the diagonal matrix with the eigenvalues,

$$\hat{\boldsymbol{A}} = \mathit{Diag}(687.0, 202.1, 103.7, 84.6, 32.2).$$

The PCs (columns of the matrix $\boldsymbol{Z}$) are computed as

$$\boldsymbol{z}_j = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^{\top})\hat{\boldsymbol{\gamma}}_j = (\boldsymbol{x}_1\hat{\gamma}_{1j} + \ldots + \boldsymbol{x}_p\hat{\gamma}_{pj}) - \mathbf{1}\bar{\boldsymbol{x}}^{\top}\hat{\boldsymbol{\gamma}}_j,$$

which is

$$\begin{aligned}
\boldsymbol{z}_1 &= 0.51\boldsymbol{x}_1 + 0.37\boldsymbol{x}_2 + 0.35\boldsymbol{x}_3 + 0.45\boldsymbol{x}_4 + 0.53\boldsymbol{x}_5 + 99.5 \cdot \mathbf{1} \\
\boldsymbol{z}_2 &= -0.75\boldsymbol{x}_1 - 0.21\boldsymbol{x}_2 + 0.08\boldsymbol{x}_3 + 0.30\boldsymbol{x}_4 + 0.55\boldsymbol{x}_5 + 1.4 \cdot \mathbf{1} \\
\boldsymbol{z}_3 &= -0.30\boldsymbol{x}_1 + 0.42\boldsymbol{x}_2 + 0.15\boldsymbol{x}_3 + 0.60\boldsymbol{x}_4 - 0.60\boldsymbol{x}_5 + 19.2 \cdot \mathbf{1} \\
\boldsymbol{z}_4 &= 0.30\boldsymbol{x}_1 - 0.78\boldsymbol{x}_2 + 0.00\boldsymbol{x}_3 + 0.52\boldsymbol{x}_4 - 0.18\boldsymbol{x}_5 - 11.5 \cdot \mathbf{1} \\
\boldsymbol{z}_5 &= 0.08\boldsymbol{x}_1 + 0.19\boldsymbol{x}_2 - 0.92\boldsymbol{x}_3 + 0.29\boldsymbol{x}_4 + 0.15\boldsymbol{x}_5 - 14.4 \cdot \mathbf{1}
\end{aligned}$$

The contribution of the variables to the PCs are the loadings. One can see that the first PC has positive contributions from all variables, in about the same order of magnitude, and thus this PC could be interpreted as an average result of the students on all subjects. PC 2 is more difficult to interpret.

The empirical variances of the PCs are equal to the eigenvalues,

$$\begin{aligned}
\widehat{a}_1 &= \widehat{Var}(\boldsymbol{z}_1) = 687.0 \\
\widehat{a}_2 &= \widehat{Var}(\boldsymbol{z}_2) = 202.1 \\
\widehat{a}_3 &= \widehat{Var}(\boldsymbol{z}_3) = 103.7 \\
\widehat{a}_4 &= \widehat{Var}(\boldsymbol{z}_4) = \;\; 84.6 \\
\widehat{a}_5 &= \widehat{Var}(\boldsymbol{z}_5) = \;\; 32.2 \quad .
\end{aligned}$$

We see that PC 1 has a variance which is about 3 times bigger than that of PC 2, and thus PC 1 is by far more important in terms of recovering the information.

Figure 5.3 shows a plot of the first two PC scores. Since PC 1 has a meaning in terms of an average result, it is interesting to see an ordering of the numbers (IDs of students) in the plot. Obviously, the original data matrix has already been ordered according to some average result on all subjects.

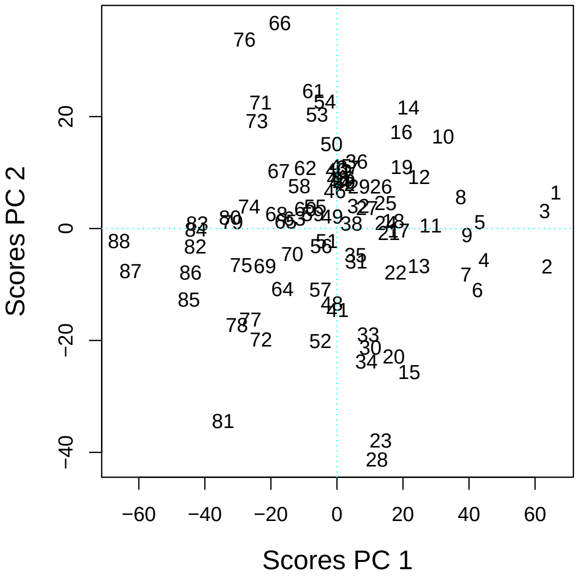


Figure 5.3: First and second PC scores of the exam data.

5.5 Number of relevant PCs

Since a major goal of PCA is dimension reduction, we are usually not interested in the last few PCs which are characterized by the smallest variances. So, how many PCs are “relevant”? The answer might depend on the purpose, what the user wants to do with the PCs. If it is for visual inspection of the data, it might be sufficient to look at those PCs which are covering the most important data information. One could also argue, that part of the information just consists of noise, and this should be contained in the last few PCs, which are not interesting for the inspection.

PCA tries to explain as much of the total variance as possible with fewer dimensions. What is the total variance? It is:

$$\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

This is the same as,

$$\frac{1}{n-1} \text{trace} \left((\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)^\top (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top) \right) = \text{trace}(\boldsymbol{S}) = \text{trace}(\hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{A}} \hat{\boldsymbol{\Gamma}}^\top) = \text{trace}(\hat{\boldsymbol{A}}) \ ,$$

where "trace" is the sum of the diagonal elements of the matrix, here the sum of the eigenvalues.

5.5.1 Statistical tests

The task is thus to determine the number k of relevant PCs. A possible test for this purpose is a test on equality of the last (i.e. smallest) $p-k$ eigenvalues, thus $a_p = a_{p-1} = \ldots = a_{k+1}$. This means that the last $p-k$ PCs cover the same amount of information, presumably originating just from irrelevant noise.

Assuming that the data are originating from a distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, one practically applies a series of tests, starting with $k = 0$, and increasing k as long as the null hypothesis can no longer be rejected. We work with the following test statistic:

$$\left(n - \frac{2p+11}{6} \right) (p-k) \ln \left(\frac{m_a}{m_g} \right) \sim \chi^2_{(p-k+2)(p-k-1)/2} \tag{5.27}$$

with

$$m_a = \frac{\hat{a}_{k+1} + \ldots + \hat{a}_p}{p-k} \ , \qquad m_g = \sqrt[p-k]{\hat{a}_{k+1} \ldots \hat{a}_p} \ . \tag{5.28}$$

When working with standardized variables, the eigenvalues are determined from the correlation matrix, which can be estimated with the sample correlation matrix $\boldsymbol{R}$. Then the above test needs to be modified, and one obtains the test statistic

$$(n-1)(p-k) \ln \left(\frac{m_a^*}{m_g^*} \right) \sim \chi^2_{(p-k+2)(p-k-1)/2} \quad , \tag{5.29}$$

where m_a^* und m_g^* are arithmetic and geometric mean of the smallest $(p-k)$ eigenvalues of $\boldsymbol{R}$, respectively.

There exist several other tests, such as a test of the hypothesis that the explained variance of the first k PCs exceeds a certain threshold. Details can be found e.g. in Anderson (2003).

Example 5.5.1 *Consider again the exam data set. We have already computed the eigenvalues from the sample covariance matrix. We start the test with $k = 0$, thus*

$$H_0 : \qquad a_1 = a_2 = a_3 = a_4 = a_5$$

versus the alternative, that this is not true. The value of the test statistic is 221.4, and for $\alpha = 0.05$, the corresponding quantile of the χ^2 distribution is $\chi^2_{14;0.95} = 23.7$. Thus, H_0 is clearly rejected.

We proceed with $k = 1$ to test

$$H_0 : \qquad a_2 = a_3 = a_4 = a_5 \ .$$

The value of the test statistic is 26.1, and $\chi^2_{9;0.95} = 16.1$, thus still rejection.

Now set $k = 2$ to test

$$H_0 : \qquad a_3 = a_4 = a_5 \quad .$$

The value of the test statistic is 7.30*, and the critical value is $\chi^2_{5;0.95} = 11.07$. Therefore, we can no longer reject the null hypothesis, and conclude that only the first 2 PCs are relevant.*

5.5.2 Rules of thumb

In the literature, one can find different "rules of thumb" to determine the number of relevant PCs, and they are not based on statistical reasoning. A major point is the argumentation: Do we assume (as above) that the non-relevant PCs are generated from noise? Rules of thumb might not be appropriate to answer this question. Another thought is related to the purpose: What are you doing with the "relevant" PCs? If it is "just" for visual exploration, the precise answer might not be that crucial.

One rule of thumb concerning the number of relevant PCs is the proportion of explained variance using k PCs on the total variance, which is defined as

$$\frac{\sum_{j=1}^{k} \hat{a}_j}{\sum_{i=1}^{p} \hat{a}_i} \geq \gamma \ , \tag{5.30}$$

where γ could be set to 80% or to 90%.

A further frequently applied criterion is to exclude those PCs which have a variance (eigenvalue) lower than the average. If the data are standardized, the sum of the eigenvalues is p, and thus the average is 1.

Example 5.5.2 *Consider the exam data again. Then the proportion of explained variance of the first j PCs is*

$\frac{\hat{a}_j}{\sum_i \hat{a}_i}$ in %	61.9	18.2	9.3	7.6	2.9
total %	61.9	80.2	89.5	97.1	100

Considering $\gamma = 80\%$*, one would take the first 2 PCs. Based on the average, one would only take the first PC.*

A further possibility to select the number of relevant PCs is the *scree graph*, which shows the proportion of explained variance of each PC versus the number of the PC. We exclude those (smallest) PCs where the proportion follows approximately a linear trend.

Example 5.5.3 *The scree graph for the exam data is shown in Figure 5.4. One can see that the contributions of the last 3 PCs follow a linear trend, and thus these PCs can be excluded.*

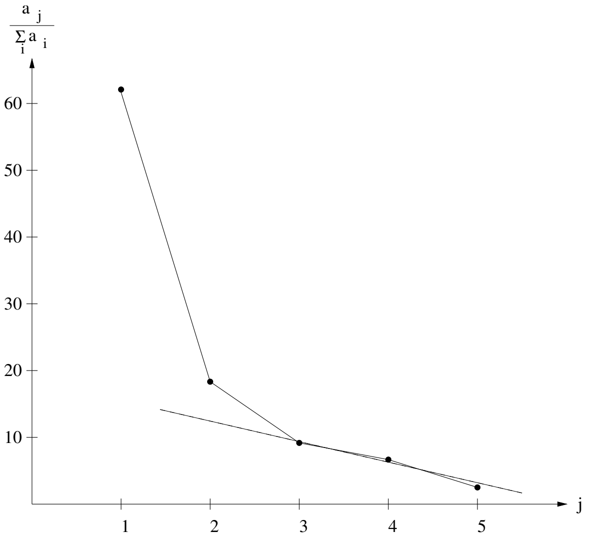


Figure 5.4: Scree graph for the exam data example.

A next possibility is to inspect the values of the squared correlation coefficients between variables and components, see Equation (5.25). If those squared correlations are low for the last PCs, they can be excluded without much loss of information.

Example 5.5.4 *The squared correlations for the exam data set are given in Table 5.1. One can see that the first PC explains* 57.4% *of* $\boldsymbol{x}_1$*,* 53.9% *of* $\boldsymbol{x}_2$*, etc. Accordingly, PC 1 is quite important for all the variables. This is different for PC 2, where we can only find important contributions to variable 1 and 5. If we would exclude the last PC (and we will), we will lose* 24.3% *of the information of* $\boldsymbol{x}_3$*. Excluding also PC 4 means to lose substantial information of the second variable.*

In practice one needs to find a compromise between the different rules.

There are also so-called *resampling* procedures to determine the number of relevant PCs, such as *Jackknife* (Efron, 1982), *Bootstrap* (Efron, 1979; 1981) and *Cross Validation* (Stone, 1974; Eastment and Krzanowski, 1982). The idea is to consider only random subsets of the available data, to derive some "statistic", and to repeat this procedure several times with new random subsets. This allows to estimate the uncertainty (e.g. of the explained variance by k PCs) and thus to give a more accurate answer, usually without using strict distributional assumptions.

Table 5.1: Squared correlations between variables $\boldsymbol{x}_j$ and PCs $\boldsymbol{z}_k$ for the exam data set.

$\hat{\lambda}^2_{jk}$	$\boldsymbol{z}_1$	$\boldsymbol{z}_2$	$\boldsymbol{z}_3$	$\boldsymbol{z}_4$	$\boldsymbol{z}_5$
$\boldsymbol{x}_1$	0.574	0.371	0.030	0.024	0.001
$\boldsymbol{x}_2$	0.539	0.050	0.104	0.300	0.007
$\boldsymbol{x}_3$	0.727	0.018	0.019	0.000	0.243
$\boldsymbol{x}_4$	0.634	0.083	0.168	0.103	0.012
$\boldsymbol{x}_5$	0.660	0.204	0.125	0.009	0.002

5.6 Singular value decomposition

Singular value decomposition (SVD) can be viewed as an alternative algorithm to determine the PCs. It is not based on a decomposition of the covariance matrix, but it directly uses the data matrix for the decomposition. This is a particular advantage if $n < p$ (“flat” data matrices), which would always lead to a singular covariance matrix, where the last $n - p$ PCs would have variance 0.

Let us assume in the following that the columns of the real-valued data matrix $\boldsymbol{X}$ are centered to mean 0. Then there exists an orthogonal $n \times n$ matrix $\boldsymbol{U}$ and an orthogonal $p \times p$ matrix $\boldsymbol{V}$, such that we obtain the decomposition

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top , \tag{5.31}$$

where $\boldsymbol{D}$ is an $n \times p$ matrix with “diagonal” elements $d_{ii} \geq 0$ for $i = 1, \ldots, \min(n, p)$, and the remaining elements are 0. The positive values d_{ii} are called *singular values* of $\boldsymbol{X}$. The number of these positive values corresponds to the rank of $\boldsymbol{X}$.

If the rank of $\boldsymbol{X}$ is $k \leq min(n, p)$, then $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X} = \sum_{i=1}^{k} d_{ii} \boldsymbol{u}_i \boldsymbol{v}_i^\top , \tag{5.32}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the i-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Due to the orthogonality of $\boldsymbol{U}$ and $\boldsymbol{V}$ we obtain

$$\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{u}_i = d_{ii}^2 \boldsymbol{u}_i \tag{5.33}$$

and

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{v}_i = d_{ii}^2 \boldsymbol{v}_i \, . \tag{5.34}$$

This means that $\boldsymbol{u}_i$ is the i-th eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$ to the eigenvalue d_{ii}^2, and $\boldsymbol{v}_i$ is the i-th eigenvector of $\boldsymbol{X}^\top\boldsymbol{X}$ to the same eigenvalue d_{ii}^2. The eigenvalues for $i = 1, \ldots, k$ are strictly positive, and the remaining ones are zero.

As a summary we conclude that $\boldsymbol{U}$ has n orthogonal eigenvectors of $\boldsymbol{X}\boldsymbol{X}^\top$ in its columns, and $\boldsymbol{V}$ has p orthogonal eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$ in its columns.

It is now straightforward to see the connection to a covariance-based estimation of the PCs, as it was done in Section 5.4: Consider again mean-centered data $\boldsymbol{X}$.

Then the sample PCs were defined as $\boldsymbol{Z} = \boldsymbol{X}\hat{\boldsymbol{\Gamma}}$, and thus $\boldsymbol{X} = \boldsymbol{Z}\hat{\boldsymbol{\Gamma}}^\top$. Further, in this case the sample covariance matrix is

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X} = \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{A}}\hat{\boldsymbol{\Gamma}}^\top .$$

The matrix $\hat{\boldsymbol{\Gamma}}$ is the matrix with the normed eigenvectors of $\boldsymbol{S}$. In SVD, $\boldsymbol{V}$ is the matrix with the normed eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$, and therefore we can conclude that $\hat{\boldsymbol{\Gamma}} \equiv \boldsymbol{V}$. From Equation (5.34) we see that $d_{ii}^2 = (n-1)\hat{a}_i$. We can thus write

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \tag{5.35}$$

and therefore we have $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$.

Alternative definitions for PCs

With these considerations we can formulate the PCA problem based on a different objective function.

Let us first define the *Frobenius norm* of a matrix: Denote $\boldsymbol{x}_{i.}$ as the rows of $\boldsymbol{X}$, for $i = 1, \dots, n$. The Frobenius norm of $\boldsymbol{X}$ is defined as:

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^n \|\boldsymbol{x}_{i.}\|_2^2} = \sqrt{\sum_{i=1}^n\sum_{j=1}^p x_{ij}^2}$$

From above we have

$$\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D} = \boldsymbol{Z}.$$

Define $\boldsymbol{V}_m = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_m)$, and m is smaller than the rank of $\boldsymbol{X}$. Then the columns of $\boldsymbol{X}\boldsymbol{V}_m$ are the first m PCs. This is equivalent to a projection of $\boldsymbol{X}$ onto an m-dimensional subspace formed by $\boldsymbol{V}_m$. One can show that

$$\boldsymbol{V}_m = arg \max_{\boldsymbol{B}} \|\boldsymbol{X}\boldsymbol{B}\|_F^2$$

for any $p \times m$ matrix $\boldsymbol{B}$ with $rank(\boldsymbol{B}) \leq m$ and $\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}$.

An equivalent formulation is the following: We have

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{X}\boldsymbol{V}_m\boldsymbol{V}_m^\top + \boldsymbol{E}.$$

Then

$$\boldsymbol{V}_m = arg \min_{\boldsymbol{B}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\|_F^2$$

with the same definition of $\boldsymbol{B}$ as above. Note that $\hat{\boldsymbol{X}} := \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top$ has the same dimension as $\boldsymbol{X}$. Therefore we can view $\hat{\boldsymbol{X}}$ as a rank m approximation of $\boldsymbol{X}$, which is optimal in the above sense. For finding the PC directions we are in fact minimizing residual sum-of-squares, with the residual matrix $\boldsymbol{X} - \hat{\boldsymbol{X}}$.

5.7 Biplots: visualizing PC loadings and scores

Biplots have been introduced in Gabriel (1971) to display both variable and object information jointly in one plot – usually in 2 dimensions. The "bi" does not refer to the "2 dimensions" but to the joint presentation of variables and observations. There are even 3-dimensional versions of biplots (Gabriel, 1986), which at that time could be inspected on paper with red-green glasses – interesting but probably not very practical.

Here we will introduce biplots to represent loadings and scores from a PCA. Biplots have also been introduced for other methods, such as multidimensional scaling, correspondence analysis, see, e.g., Gower and Hand (1996).

We like 2-dimensional plots because they are easy to handle. A projection of the data into the plane should, however, make sure that the data have approximately rank 2. Then we are sure that indeed the projection represents the essential variability.

Let $\boldsymbol{X}$ be a mean-centered $n \times p$ data matrix of rank k (not much larger than 2). The biplot shows a representation of $\boldsymbol{X}$ by means of two groups of vectors with dimension n and p, respectively, which form a rank-2 approximation of $\boldsymbol{X}$. Denote this rank-2 approximation by $\boldsymbol{X}_{(2)}$.

We know that a least-squares based rank-2 approximation is given by the first 2 PCs, and we are aware of SVD to compute these first 2 PCs. In the following we will only need the first 2 columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, but we will still use the notation $\boldsymbol{U}$ and $\boldsymbol{V}$ in order to avoid new symbols. Thus:

$$\boldsymbol{X} \approx \boldsymbol{X}_{(2)} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \end{pmatrix} . \tag{5.36}$$

Since we want to represent the data by n score vectors and p loadings vectors (all 2-dimensional), we need a decomposition of $\boldsymbol{X}_{(2)}$ into an $n \times 2$ and an $p \times 2$ matrix, say

$$\boldsymbol{X}_{(2)} = \boldsymbol{G}\boldsymbol{H}^\top \tag{5.37}$$

with

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c} \tag{5.38}$$

and

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{5.39}$$

for $0 \le c \le 1$. Depending on the choice of c, the first 2 singular values are distributed among the matrices $\boldsymbol{G}$ and $\boldsymbol{H}$. The biplot consists of the rows of $\boldsymbol{G}$ and $\boldsymbol{H}$, i.e. of $n + p$ 2-dimensional vectors.

For the choice $c = 0.5$ we use the same scaling for the observation and variable vectors, which seems to be natural. However, we obtain nicer properties with the choice $c = 1$ (and with re-scaling):

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{g}_{1.}^\top \\ \vdots \\ \boldsymbol{g}_{n.}^\top \end{pmatrix} = \sqrt{n-1} \left(\boldsymbol{u}_1 \ \boldsymbol{u}_2 \right) \tag{5.40}$$

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{h}_{1.}^\top \\ \vdots \\ \boldsymbol{h}_{p.}^\top \end{pmatrix} = \frac{1}{\sqrt{n-1}} \left(\boldsymbol{v}_1 \ \boldsymbol{v}_2 \right) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \tag{5.41}$$

These properties are:

- The inner product between the rows of $\boldsymbol{G}$ and the rows of $\boldsymbol{H}$ approximate the values x_{ij} of the data matrix $\boldsymbol{X}$:

$$\boldsymbol{g}_{i.}^\top \boldsymbol{h}_{j.} = \sqrt{n-1}\ \boldsymbol{u}_{i.}^\top \frac{1}{\sqrt{n-1}} \left(\boldsymbol{v}_{j.}^\top \boldsymbol{D} \right)^\top = \boldsymbol{u}_{i.}^\top \boldsymbol{D} \boldsymbol{v}_{j.} \approx x_{ij} \ . \tag{5.42}$$

- The inner product between the rows of $\boldsymbol{H}$ approximates the covariance:

$$\begin{aligned} \boldsymbol{H}\boldsymbol{H}^\top &= \left(\frac{1}{\sqrt{n-1}} \boldsymbol{V}\boldsymbol{D} \right) \left(\frac{1}{\sqrt{n-1}} \boldsymbol{D}\boldsymbol{V}^\top \right) = \frac{1}{n-1} \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top \\ &= \frac{1}{n-1} \left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \right) \left(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \right) = \frac{1}{n-1} \boldsymbol{X}_{(2)}^\top \boldsymbol{X}_{(2)} \\ &\approx \frac{1}{n-1} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{S} \ . \end{aligned} \tag{5.43}$$

  It follows that the squared Euclidean norm of the rows of $\boldsymbol{H}$, i.e. $\|\boldsymbol{h}_{j.}\|^2 = \boldsymbol{h}_{j.}^\top \boldsymbol{h}_{j.}$, approximates the variance. Moreover, the cosine between $\boldsymbol{h}_{i.}$ and $\boldsymbol{h}_{j.}$ $(i, j = 1, \ldots, p)$ approximates the correlation between the variables,

$$\cos(\boldsymbol{h}_{i.}, \boldsymbol{h}_{j.}) = \frac{\boldsymbol{h}_{i.}^\top \boldsymbol{h}_{j.}}{\|\boldsymbol{h}_{i.}\| \|\boldsymbol{h}_{j.}\|} \approx r_{ij} \ . \tag{5.44}$$

- The Euclidean distance between the rows of $\boldsymbol{G}$ approximates the Mahalanobis distance between the observations,

$$\begin{aligned} \|\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\|^2 &= \left(\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.} \right)^\top \left(\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.} \right) = (n-1) \left(\boldsymbol{u}_{i.} - \boldsymbol{u}_{j.} \right)^\top \left(\boldsymbol{u}_{i.} - \boldsymbol{u}_{j.} \right) \\ &\approx \left(\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.} \right)^\top \boldsymbol{S}^{-1} \left(\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.} \right) \ , \end{aligned} \tag{5.45}$$

  because

$$\boldsymbol{x}_{i.}^\top \boldsymbol{S}^{-1} \boldsymbol{x}_{j.} \approx \left(\boldsymbol{u}_{i.}^\top \boldsymbol{D}\boldsymbol{V}^\top \right) (n-1) \left(\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top \right) \left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{u}_{j.} \right) = (n-1) \boldsymbol{u}_{i.}^\top \boldsymbol{u}_{j.} \tag{5.46}$$

  for $i, j = 1, \ldots, n$.

Example 5.7.1 *Consider once more the exam data, for which we already have computed loadings and scores. In order to obtain the properties of the biplot as described before, these loadings and scores need to be rescaled. This is done automatically in the default setting of the R function* `biplot()`*.*

The resulting biplot is shown in Figure 5.5. The numbers refer to the observations, and these are in fact the coordinates of the $n \times 2$ matrix $\boldsymbol{G}$. The scale of these coordinates is at the bottom and left side of the plot. The arrows link the origin with the coordinates of the $p \times 2$ matrix $\boldsymbol{H}$, with the scale on the right and top of the plot. As we can see, PC 1 is presenting an average result in the subjects, with good results for students on the right-hand side of the plot, and poor results for the very left. PC 2 seems to discriminate subjects requiring mathematical skills (high values, on the top) from geometrical skills (low values, on the bottom). Thus, the position of every student in the plot allows for a particular interpretation. We can also see that e.g. student 81 is somehow exceptional, with good mathematical and poor geometrical skills. Note that the geometrical subjects (`mec` *and* `vec`*) were with closed book, while the others were with open book.*

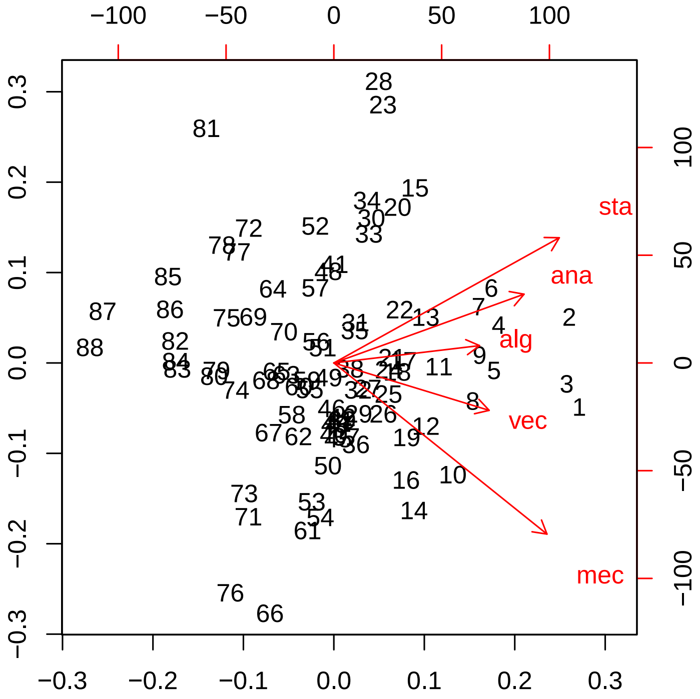


Figure 5.5: Biplot for the exam data set.

5.8 Diagnostics for PCA

Although PCA is not a typical method for multivariate outlier detection, one could still be interested if single observations are deviating from the data majority. For this purpose, two distance measures have been introduced: the *score distance* (SD) and the *orthogonal distance* (OD).

Figure 5.6 tries to explain the ideas behind these distances. We have observations in the 3-dimensional space, and the PCA space is built up with 2 PCs. The SD is a distance measure in the PCA space, and it is equal to the Mahalanobis distances of the observations projected into this space. The ellipses refer to these Mahalanobis distances. The OD is the distance orthogonal to the PCA space.

Here we can see three particular observations: Observation 1 has big OD but small SD; observation 2 has big OD and big SD; observation 3 has small OD but big SD. Similar to the diagnostics in regression one could classify these different types of observations as vertical outliers (1), and good (3) and bad (2) leverage points. In particular, bad leverage points can have a strong influence on the classical estimation of the PCs. Here, "classical" refers to the SVD estimation, or equivalently, the estimation based on the empirical covariance matrix. A robust PCA could easily be obtained through a robust estimation of the covariance matrix. In this case, the diagnostics would also make much more sense.

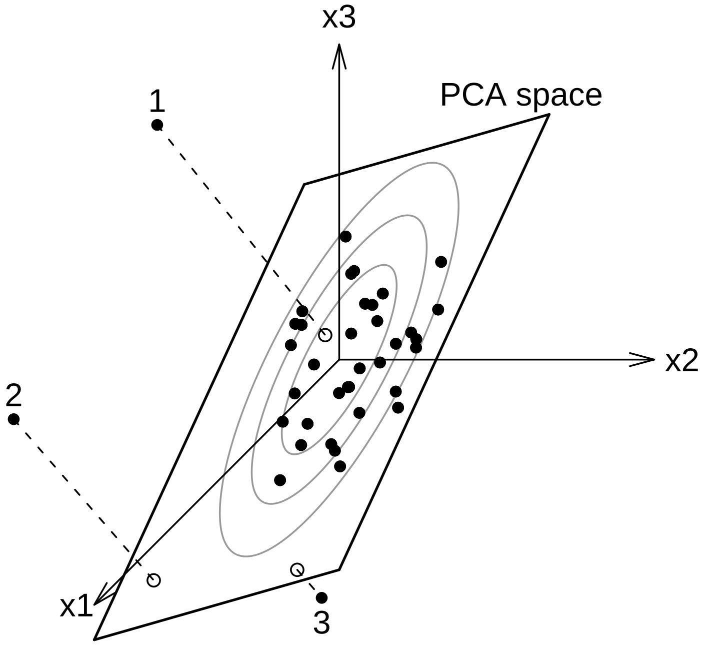


Figure 5.6: Diagnostic plot for PCA.

Formally, these distances are defined as follows. Let $\hat{\mathbf{\Gamma}}_k = (\hat{\gamma}_1, \ldots, \hat{\gamma}_k)$ be the matrix of the first k estimated PC loadings, $\boldsymbol{z}_{i.} = (z_{i1}, \ldots, z_{ik})^\top$ the i-th score vector, and $\hat{a}_1, \ldots, \hat{a}_k$ the corresponding variances of the PCs. The number k of PCs can be selected according to a criterion defined in Section 5.5, e.g. such that 80% of the variability is explained.

The SD for the i-th observation ($i = 1, \ldots, n$) is defined as

$$\mathrm{SD}_i = \left(\sum_{j=1}^{k} \frac{z_{ij}^2}{\hat{a}_j} \right)^{1/2} .$$

This is equal to the Mahalanobis distance of the score vector to the PCA center (zero) with respect to the covariance matrix ($\boldsymbol{A}$).

The OD for the i-th observation $\boldsymbol{x}_{i.}$ ($i = 1, \ldots, n$) is defined as

$$\mathrm{OD}_i = \|\boldsymbol{x}_{i.} - \hat{\boldsymbol{\Gamma}}_k \boldsymbol{z}_{i.}\|_2,$$

which is the Euclidean distance of the observation to its projection into the space of the first k PCs.

Similar to multivariate outlier detection, one can define cutoff values for both distance measures, which would refer to unusual values of SD and/or OD. Since SD is a Mahalanobis distance, a suitable cutoff value is $\sqrt{\chi^2_{k;0.975}}$. For the cutoff value for the OD it has been argued that $\mathrm{OD}^{2/3}$ is closer to normality, and thus a suitable cutoff value is

$$\left(\mathrm{median}_i(\mathrm{OD}_i^{2/3}) + \mathrm{MAD}_i(\mathrm{OD}_i^{2/3}) z_{0.975} \right)^{3/2} ,$$

where $z_{0.975}$ is the 0.975 quantile of the $N(0,1)$. MAD stands for the Median Absolute Deviation, which is defined for univariate values $y_1, \ldots, y_n$ as

$$\mathrm{MAD} = 1.483 \cdot \mathrm{median}_i(\mid y_i - \mathrm{median}_j(y_j) \mid).$$

Example 5.8.1 *For the exam data set we can get such diagnostic plots quite easily in R. In the following we are using a robust PCA method to estimate the PCs, implemented in the function* `PcaHubert()` *of the package* `rrcov` *– for details see help pages, as well as classical PCA using* `princomp()`*. The R code is as follows:*

```
library(bootstrap)
data(scor) # data
library(rrcov)
res1 <- PcaHubert(scor,scale=TRUE)  # robust
plot(res1)

X <- scale(scor)
X.pca <- princomp(X)                # not robust
library(chemometrics)
res <- pcaDiagplot(X,X.pca,a=3,plot=TRUE) # 3 PCs
```

The resulting diagnostic plots are shown in Figure 5.7. The robust analysis (left) shows some more unusual observations compared to classical PCA (right).

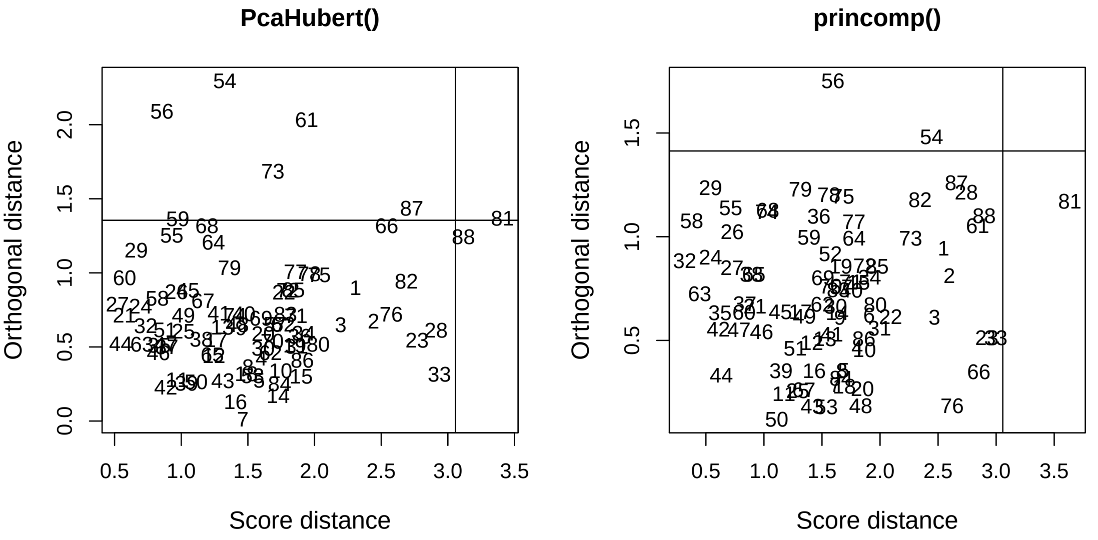


Figure 5.7: Diagnostic plots for the exam data set: left robust, right non-robust.

References

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*, Wiley, Chichester, 2003.

A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications.* Wiley & Sons, New York, 1994.

T.F. Cox and A.A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.

H.T. Eastment and W.J. Krzanowski. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, 24:73–77, 1982.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7:1–26, 1979.

B. Efron. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Methods. *Biometrika*, 68:589–599, 1981.

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* J.W. Arrowsmith Ltd., Bristol, 1982.

K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

K.R. Gabriel and C.L. Odoroff. Use of three-dimensional biplots for diagnosis of models. In W. Gaul and M. Schader, editors, *Classification as a Tool of Research*, pages 153–159, 1986.

J.C. Gower and D.J. Hand. *Biplots.* Chapman & Hall, London, 1996.

J.C. Gower and S.A. Harding. Nonlinear biplots. *Biometrika*, 75(3):445–455, 1988.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24:417–441, 498–520, 1933.

J.E. Jackson. *A User's Guide To Principal Components.* Wiley & Sons, New York, 1991.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag. (6)*, 2:559–572, 1901.

M. Stone. Cross-validatory choice and assessment of statistical prediction. *J. R. Stat. Soc. B*, 36:111–133, 1974.

Chapter 6

Factor analysis

6.1 Introduction

Consider n observations which have been measured at p characteristics, then these characteristics will usually correlate among each other. They could even influence each other, or there could be a "hidden" non-observed quantity which influences the measured characteristics.

In factor analysis we assume that what we observe is basically the result of underlying quantities which are not directly observable. These quantities are called **latent variables**, and they cannot be measured. The "factors" in factor analysis aim at isolating such latent variables, explaining the relationships in the data.

As an example, consider "intelligence" as a hidden factor. Intelligence cannot be measured directly, but only indirectly, e.g. in terms of attentiveness, knowledge, abilitied in jugdement, etc. Testing specific persons with respect to such observable characteristics may allow to isolate a factor "intelligence", underlying all these measurements.

Note that factor analysis is basically similar to PCA, since we also aim for a dimension reduction. However, we would like to have interpretable factors (the components in PCA are not necessarily interpretable), and we also have a statistical model (PCs were only defined by a linear transformation).

6.2 Factor analysis model

6.2.1 Definition

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ be a p-dimensional vector of random variables, $x_1, x_2, \ldots, x_p$, which will describe our characteristics or variables later on.

Since in PCA we had the problem of the dependency on the scale, we will right away center and scale the random variables to mean zero and variance one. Thus, $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^\top$ is the new random variable, obtained by

$$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}}, \qquad i = 1, \ldots, p \quad .$$

In the model we already assume that the information contained in $\boldsymbol{y}$ can be re-expressed by a smaller number $k < p$ of unknown random variables (factors) $\boldsymbol{f} = (f_1, \ldots, f_k)^\top$, up to an error term $\boldsymbol{e}$. Thus, dimension reduction is already intrinsically stated in the model, called k-factor model:

$$\boldsymbol{y} = \boldsymbol{\Lambda f} + \boldsymbol{e} \quad . \tag{6.1}$$

Here, $\boldsymbol{\Lambda} = [(\lambda_{ij})]$ is a $(p \times k)$ matrix with fixed values. It is called *loadings matrix*, and it describes the relationships between factors and variables (as in PCA). The error term $\boldsymbol{e} = (e_1, \ldots, e_p)^\top$ is often called *unique factor* (or uniqueness).

In this model we have the following assumptions:

$$\begin{aligned} E(\boldsymbol{f}) &= \boldsymbol{0},\ Cov(e_i, e_j) = 0 \qquad (i \neq j), \\ E(\boldsymbol{e}) &= \boldsymbol{0},\ \ Cov(\boldsymbol{f}, \boldsymbol{e}) = \boldsymbol{O}\ , \\ Var(f_i) &= 1\ . \end{aligned}$$

With these assumptions, we can see that the covariance matrix of the error term has a diagonal form:

$$Cov(\boldsymbol{e}) = \boldsymbol{\Psi} = Diag(\psi_{11}, \ldots, \psi_{pp})\ . \tag{6.2}$$

We can thus re-express the correlation matrix $\boldsymbol{\rho} = [(\rho_{ij})]$ of our random variables $\boldsymbol{x}$ by our model:

$$\begin{aligned} \boldsymbol{\rho} &= Cor(\boldsymbol{x}) = Cov(\boldsymbol{y}) = Cov(\boldsymbol{\Lambda f} + \boldsymbol{e}) \\ &= \boldsymbol{\Lambda} \underbrace{Cov(\boldsymbol{f})}_{\boldsymbol{\Phi}} \boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda} \underbrace{Cov(\boldsymbol{f}, \boldsymbol{e})}_{\boldsymbol{O}} + \underbrace{Cov(\boldsymbol{e}, \boldsymbol{f})}_{\boldsymbol{O}} \boldsymbol{\Lambda}^\top + \underbrace{Cov(\boldsymbol{e})}_{\boldsymbol{\Psi}} \\ &= \boldsymbol{\Lambda \Phi \Lambda}^\top + \boldsymbol{\Psi} \quad . \end{aligned}$$

Here, $\boldsymbol{\Phi}$ is the $(k \times k)$ matrix with the correlations between the factors. If we additionally assume that the factors are uncorrelated, i.e.

$$Cov(\boldsymbol{f}) = \boldsymbol{\Phi} = \boldsymbol{I} \quad , \tag{6.3}$$

we obtain

$$\boldsymbol{\rho} = \boldsymbol{\Lambda\Lambda}^\top + \boldsymbol{\Psi} \tag{6.4}$$

or

$$\boldsymbol{\rho}_{red} = \boldsymbol{\Lambda\Lambda}^\top = \begin{pmatrix} \kappa_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \kappa_2^2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \kappa_p^2 \end{pmatrix} = \boldsymbol{\rho} - \boldsymbol{\Psi} \quad , \tag{6.5}$$

with the *reduced correlation matrix* $\boldsymbol{\rho}_{red}$. The diagonal elements $\kappa_i^2 = 1 - \psi_{ii} = \sum_{j=1}^k \lambda_{ij}^2 \quad (i = 1, \ldots, p)$ are called *communalities.* They correspond to the row-sums of the squared factor loadings, and describe the proportion of variance of y_i, explained by the factors, because the total variance is trace($\boldsymbol{\rho}$), and thus trace($\boldsymbol{\rho}_{red}$) is the variance explained by the factor model.

6.2.2 Non-uniqueness of the factor loadings

Let $\boldsymbol{G}$ be an orthogonal matrix of dimension $(k \times k)$. Because of $\boldsymbol{G}^{-1} = \boldsymbol{G}^\top$ we have

$$\boldsymbol{y} = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{f}) + \boldsymbol{e} \ . \tag{6.6}$$

Since the new factors $\boldsymbol{G}^\top \boldsymbol{f}$ also fulfill our model assumptions,

$$\begin{array}{cc} E(\boldsymbol{G}^\top \boldsymbol{f}) = \boldsymbol{0}, & Cov(\boldsymbol{G}^\top \boldsymbol{f}) = \boldsymbol{I} \\ \text{and} & Cov(\boldsymbol{G}^\top \boldsymbol{f}, \boldsymbol{e}) = \boldsymbol{O} \ , \end{array}$$

this k-factor model is also valid with the new factors, i.e.

$$\boldsymbol{\rho} = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{\Lambda}^\top) + \boldsymbol{\Psi} \ . \tag{6.7}$$

This, however, means that the new factor loadings are not uniquely determined.

Basically, this is not a big problem, because later on we want to rotate the factors in any case in order to obtain a better interpretation. For now, if we just want to obtain a unique solution, we can impose further restrictions in order to achieve uniqueness. These are the constraints that either $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ or $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$ is diagonal.

6.2.3 Number of parameters

Practically, we can estimate the correlation matrix $\widehat{\boldsymbol{\rho}}$ based on real data. Then, $\widehat{\boldsymbol{\rho}}$ is used to estimate loadings $\widehat{\boldsymbol{\Lambda}}$ and error variances $\widehat{\boldsymbol{\Psi}}$, and they need to fulfill either

$$\widehat{\boldsymbol{\Lambda}}^\top \widehat{\boldsymbol{\Psi}}^{-1} \widehat{\boldsymbol{\Lambda}} = Diag \qquad \text{or} \qquad \widehat{\boldsymbol{\Lambda}}^\top \widehat{\boldsymbol{\Lambda}} = Diag \tag{6.8}$$

as well as

$$\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Lambda}}^\top + \widehat{\boldsymbol{\Psi}} \ . \tag{6.9}$$

Does the factor model lead to a simpler interpretation than the correlation matrix? In other words, is the number of parameters for the factor model lower compared to the correlation matrix? For the correlation matrix we need to estimate $\frac{1}{2}p(p+1)$ parameters. For a k-factor model we need to estimate $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, thus $pk + p$ parameters. For uniqueness of these estimates we ask for diagonal structure of either $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ or $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$, thus $\frac{1}{2}k(k-1)$ restrictions. Therefore, for the k-factor model we have $pk + p - \frac{1}{2}k(k-1)$ parameters to estimate. The difference to an unrestricted model is thus

$$\begin{aligned} s &= \frac{1}{2}p(p+1) - \left(pk + p - \frac{1}{2}k(k-1)\right) \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \quad . \end{aligned} \tag{6.10}$$

Only for the case $s > 0$, the k-factor model leads to a simpler interpretation than the correlation matrix, the other cases would not be valid (or useful) solutions. Thus, the requirement for $s > 0$ leads to an upper bound for the number k of factors.

6.2.4 Parameter estimation by Principal Factor Analysis (PFA)

The main job in factor analysis is to estimate the parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. There are different approaches, such as the *Maximum Likelihood Method (MLM)*, implemented in the R function `factanal()`, or *Principal Factor Analysis (PFA)*, implemented in the R package `StatDA` as function `pfa()`. For MLM we need to assume that the data are generated from a multivariate normal distribution, which is not explicitely required for PFA. In fact, MLM is quite technical, and thus we focus here on PFA only, since this is closely related to PCA.

We start from our model

$$\boldsymbol{\rho} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi} \tag{6.11}$$

and try to first estimate the communalities, which is equivalent to estimating the diagonal elements of $\mathbf{\Psi}$. Afterward we estimate the loadings matrix.

Estimation of the communalities

The communalities

$$\kappa_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = 1 - \psi_{ii} \qquad (i = 1, \ldots, p) \tag{6.12}$$

describe the porportion of variance explained by the k-factor model. They are in the interval $[0, 1]$. There are different options to estimate the communalities:

1. Highest correlation coefficient: $max_{i \neq j} \mid \hat{\rho}_{ij} \mid$
   For the estimation of the communalities we use from each column of $\boldsymbol{\rho}$ the largest (absolute) non-diagonal element.

2. Squared multiple correlation coefficient: $\hat{\rho}^2_{i,12\ldots)i(\ldots p}$
   This measure refers to an R^2 measure from a linear regression of the i-th variable on the remaining variables. Thus, it tells us about the variance porportion of the i-th variable explained from the other variables. We can compute this measure as:
   $$\hat{\rho}^2_{i,12\ldots)i(\ldots p} = 1 - \frac{1}{\hat{\rho}^{ii}} \quad , \tag{6.13}$$
   where $\hat{\rho}^{ii}$ is the i-th diagonal element of $\hat{\boldsymbol{\rho}}^{-1}$.

3. Iterative estimation:
   We first need to fix the number k of factors. This could be done by using a criterion from PCA on the number of components. Then we start to initialize the communalities by using one of the estimation methods mentioned above. The diagonal elements of $\hat{\boldsymbol{\rho}}$ are replaced by these communalities, and from the resulting reduced correlation matrix we estimate the loadings (see end of this section). Based on that we re-estimate the communalities:
   $$\hat{\kappa}_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 \qquad \text{for} \quad i = 1, \ldots, p \ , \tag{6.14}$$

and they are used to form the new reduced correlation matrix. The iteration continues until the communalities stabilize.

If the communalities are over-estimated, part of the uniquenesses are forced into the k-factor model, which might change the pattern and interpretability of the factors. The same may happen if the communalities are under-estimated, because variance of the factors is forced into the uniquenesses. This issue is more severe if the number of variables is small.

Estimation of the loadings

Once the communalities $\hat{\kappa}_i^2 = 1 - \hat{\psi}_{ii}$, for $i = 1, \ldots, p$, have been estimated, we can estimate the loadings based on the reduced correlation matrix

$$\hat{\boldsymbol{\rho}}_{red} = \hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\Psi}} \ , \tag{6.15}$$

where the diagonal consists of these communalities.

Using the Spectral Theorem 1.4.3, we have:

$$\hat{\boldsymbol{\rho}}_{red} = \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{A}}\hat{\boldsymbol{\Gamma}}^\top \ , \tag{6.16}$$

where $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ and $\hat{\boldsymbol{A}} = Diag(\hat{a}_1, \ldots, \hat{a}_p)$, with the eigenvalues $\hat{a}_i$ and the corresponding eigenvectors $\hat{\boldsymbol{\gamma}}_i$ of $\hat{\boldsymbol{\rho}}_{red}$ $(i = 1, \ldots, p)$.

Note that we fixed the number of factors with k, where $1 \leq k < p$, and thus the loadings matrix needs to have dimension $p \times k$. Thus, we shall only use the first k eigenvectors $\hat{\boldsymbol{\Gamma}}_{1:k}$ and eigenvalues $\hat{\boldsymbol{A}}_{1:k}$ in the k-factor model

$$\hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^\top = \hat{\boldsymbol{\Gamma}}_{1:k}\hat{\boldsymbol{A}}_{1:k}\hat{\boldsymbol{\Gamma}}_{1:k}^\top + \sum_{i=k+1}^{p} \hat{a}_i \hat{\boldsymbol{\gamma}}_i \hat{\boldsymbol{\gamma}}_i^\top \quad . \tag{6.17}$$

The estimated loadings matrix is thus naturally

$$\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Gamma}}_{1:k}\hat{\boldsymbol{A}}_{1:k}^{1/2} \quad , \tag{6.18}$$

where the diagonal elements of $\hat{\boldsymbol{A}}_{1:k}^{1/2}$ are the values $\sqrt{\hat{a}_1}, \ldots, \sqrt{\hat{a}_k}$.

Considering Equation (6.17), it is also natural to update the estimated uniquenesses as

$$\hat{\psi}_{ii} = 1 - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 \qquad \text{for} \quad i = 1, \ldots, p \quad . \tag{6.19}$$

The solution is valid if all $\hat{\psi}_{ii} \geq 0$.

Example 6.2.1 *Consider again the exam data set* `scor` *from the R package* `bootstrap`*. The empirical correlation matrix is given by*

$$\mathbf{R} = \begin{pmatrix} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 1 & 0.610 & 0.485 & 0.437 \\ & & 1 & 0.711 & 0.665 \\ & & & 1 & 0.607 \\ & & & & 1 \end{pmatrix} \quad .$$

We first determine an upper bound for the number k of factors. The choice $k = 3$ leads to $s = -1$, see Equation (6.10), which is a non-useful solution. Thus, $k = 2$ is the maximum possible number of factors.

For comparison we use two different methods for estimating the communalities: the maximum correlation, and the squared multiple correlation (SMC) – for the latter we need the inverse of the correlation matrix:

$$\mathbf{R}^{-1} = \begin{pmatrix} 1.603 & -0.558 & -0.510 & 0.001 & -0.041 \\ & 1.802 & -0.0659 & -0.152 & -0.039 \\ & & 3.047 & -1.113 & -0.864 \\ & & & 2.178 & -0.515 \\ & & & & 1.921 \end{pmatrix} .$$

The resulting estimates for the communalities are:

Schätzer	$\hat{\kappa}_1^2$	$\hat{\kappa}_2^2$	$\hat{\kappa}_3^2$	$\hat{\kappa}_4^2$	$\hat{\kappa}_5^2$
SMC	*0.376*	*0.445*	*0.672*	*0.541*	*0.479*
$max_{i\neq j} \mid r_{ij} \mid$	*0.553*	*0.610*	*0.711*	*0.711*	*0.665*

Here we use the maximum correlation estimator, resulting in the reduced correlation matrix:

$$\boldsymbol{R} - \hat{\boldsymbol{\Psi}} = \begin{pmatrix} 0.553 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 0.610 & 0.610 & 0.485 & 0.437 \\ & & 0.711 & 0.711 & 0.665 \\ & & & 0.711 & 0.607 \\ & & & & 0.665 \end{pmatrix}$$

This matrix is used for the spectral decomposition, and the resulting eigenvalues are:

$$\hat{a}_1 = 2.84 \qquad \hat{a}_2 = 0.38 \qquad \hat{a}_3 = 0.08 \qquad \hat{a}_4 = 0.02 \qquad \hat{a}_5 = -0.05$$

The size of the eigenvalues tells us that a 1- or 2-factor model makes sense. Table 6.1 presents the estimated communalities and loadings matrix columns for a 1- and 2-factor model. Here, the communalities have been re-estimated using Equation (6.19), and they change quite a lot in the 1-factor model, but not so much for the 2-factor model. Clearly, the first loadings vector in both models is identical, and since there are about equal contributions from all variables, the factor represents some kind of average result in all subjects. Factor 2 is more difficult to interpret – it is in fact similar as the second component in PCA, see also Figure 6.1 for a graphical presentation. Note, however, that we are going to rotate the factors in any case for better interpretability, see next section.

6.3 Factor rotation

The condition that either $\widehat{\mathbf{\Lambda}}^\top \widehat{\mathbf{\Psi}}^{-1} \widehat{\mathbf{\Lambda}}$ or $\widehat{\mathbf{\Lambda}}^\top \widehat{\mathbf{\Lambda}}$ are diagonal results in uniqueness of the solution, but not necessarily in interpretability. A rotation of the factors will change the loadings, and thus also the interpretation. The goal is to rotate in

Table 6.1: Estimated communalities and loadings (vectors) for a 1- and 2-factor model.

	k=1		k=2		
Merkmale	$\hat{\kappa}_i^2$	$\hat{\boldsymbol{\lambda}}_1$	$\hat{\kappa}_i^2$	$\hat{\boldsymbol{\lambda}}_1$	$\hat{\boldsymbol{\lambda}}_2$
ME	0.417	0.646	0.543	0.646	0.354
AG	0.506	0.711	0.597	0.711	0.303
LA	0.746	0.864	0.749	0.864	-0.051
AN	0.618	0.786	0.680	0.786	-0.249
ES	0.551	0.742	0.627	0.742	-0.276

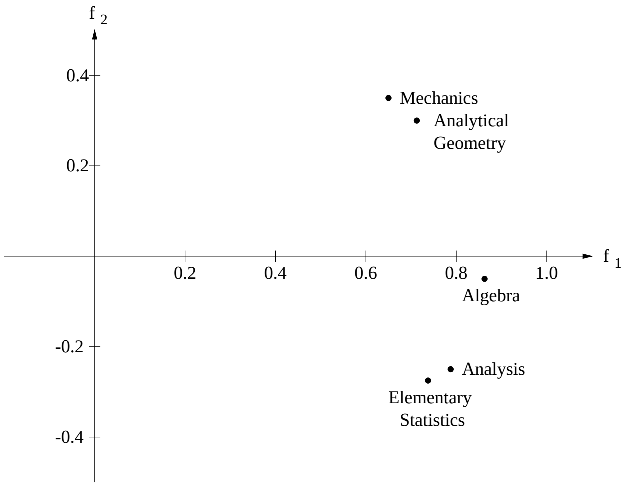


Figure 6.1: Factor loadings of the 2-factor solution for the exam data set, see also Table 6.1.

such a way that the resulting pattern of the loadings matrix is “simple” and thus interpretable. “Simple” basically means that the loadings matrix contains essentially small (absolute) values, and few values close to -1 or 1. If there is a large (absolute) value, then we know that the corresponding variable has a strong contribution on this factor, while others with loadings close to zero do not have a contribution. Many loadings close to zero would simplify the interpretation of a factor.

Figure 6.2 shows the idea of such a “simple” structure in the case of a 2-factor model. The loadings pattern in the left plot shows no specific structure, and thus the loadings from the initial factors f_1, f_2 as well as from the rotated factors $\tilde{f}_1, \tilde{f}_2$ would have no specific interpretation. This is different in the right picture, where the loadings from the initial factors are neither small nor big, while from the rotated factors we get strong contributions from some variables and weak contributions from others. The rotated factors would thus have a much clearer interpretation. In this case, the factors have been orthogonally rotated; in other situations an oblique rotation might be more successful in terms of interpretability.

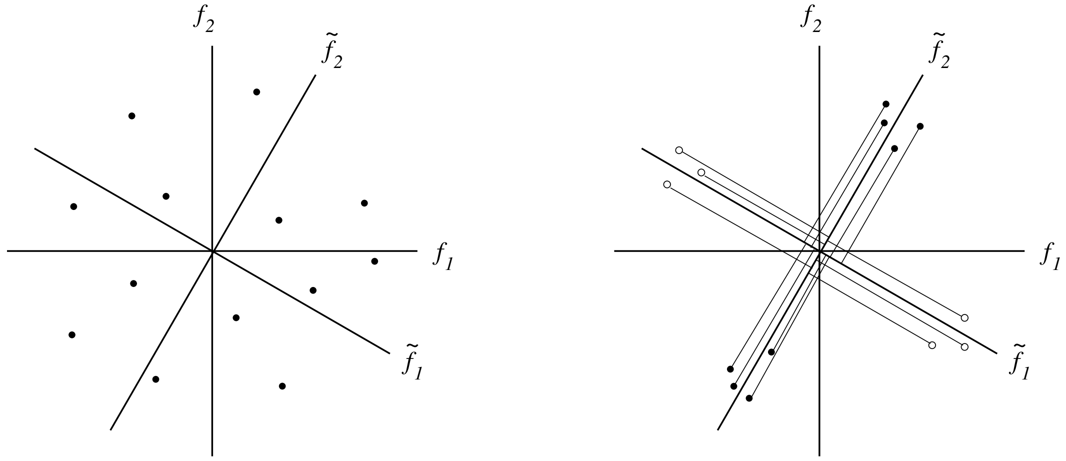


Figure 6.2: Original factors f_1, f_2 and rotated factors $\tilde{f}_1, \tilde{f}_2$. Left: random loadings pattern; right: loadings pattern where the rotated factors show a "simple" structure.

6.3.1 Orthogonal rotation

The interpretation is "simple" if a point in two dimensions is close to an axis. In that case, the product of both coordinates is small, and by taking the square one gets rid of the sign. Now we can consider the sum of all squared products as a criterion for simplicity. This should be valid for all pairs of factors, resulting in the criterion

$$\sum_{s<j=1}^{k} \sum_{i=1}^{p} \Big(\lambda_{is} \lambda_{ij} \Big)^2 \longrightarrow \min \quad . \tag{6.20}$$

A factor rotation can be achieved by a transformation of the loadings matrix. For orthogonal rotation we thus have to consider an orthogonal $k \times k$ matrix $\boldsymbol{T}$, and the rotated loadings are given by

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{T} \ . \tag{6.21}$$

Orthogonal transformations do not change the communalities, since

$$Diag(\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\Lambda}}^\top) = Diag(\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{\Lambda}^\top) = Diag(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top) \ ,$$

and thus

$$\kappa_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = \sum_{j=1}^{k} \tilde{\lambda}_{ij}^2 \qquad \text{for} \quad i = 1, \ldots, p \quad . \tag{6.22}$$

Also the squared communalities remain constant under orthogonal transformations,

$$(\kappa_i^2)^2 = \Big(\sum_{j=1}^{k} \tilde{\lambda}_{ij}^2 \Big)^2 = \sum_{j=1}^{k} \tilde{\lambda}_{ij}^4 + 2 \sum_{s<j=1}^{k} \tilde{\lambda}_{is}^2 \tilde{\lambda}_{ij}^2 = const \quad , \tag{6.23}$$

and also the sum over all variables remains constant:

$$\sum_{i=1}^{p} \sum_{j=1}^{k} \tilde{\lambda}_{ij}^4 + 2 \sum_{i=1}^{p} \sum_{s<j=1}^{k} \tilde{\lambda}_{is}^2 \tilde{\lambda}_{ij}^2 = const \quad . \tag{6.24}$$

Since the sum of both expressions is constant, one term is maximized if the other term is minimized, and vice versa.

The **Quartimax criterion** thus maximizes

$$QMAX = \sum_{i=1}^{p}\sum_{j=1}^{k}\widetilde{\lambda}_{ij}^{4}, \tag{6.25}$$

with the effect that this may lead to one dominant factor.

Another criterion for factor rotation is the **Varimax criterion** (Kaiser, 1958), where the variance of the squared factor loadings for each factor j is considered:

$$s_j^2 = \frac{1}{p}\sum_{i=1}^{p}\left(\widetilde{\lambda}_{ij}^2 - \frac{1}{p}\sum_{l=1}^{p}\widetilde{\lambda}_{lj}^2\right)^2 = \frac{1}{p}\sum_{i=1}^{p}\left(\widetilde{\lambda}_{ij}^2\right)^2 - \frac{1}{p^2}\left[\sum_{i=1}^{p}\widetilde{\lambda}_{ij}^2\right]^2 \tag{6.26}$$

This variance is summed up over all factors. Maximizing this expression means that we want to have (absolute) large and small loadings. Since variables with larger communality are dominating the criterion, one can normalize with the communalities, which yields (after multiplication with p^2) the varimax criterion

$$VMAX = p\ \sum_{j=1}^{k}\sum_{i=1}^{p}\left(\frac{\widetilde{\lambda}_{ij}}{\kappa_i}\right)^4 - \sum_{j=1}^{k}\left[\sum_{i=1}^{p}\left(\frac{\widetilde{\lambda}_{ij}}{\kappa_i}\right)^2\right]^2 \quad . \tag{6.27}$$

6.3.2 Oblique rotation

The main difference to orthogonal rotations is in the initial factor model, which is now

$$\boldsymbol{\rho} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Psi} \quad , \tag{6.28}$$

because oblique factors are no longer uncorrelated, and thus the correlation matrix $\boldsymbol{\Phi}$ of the factors needs to be considered.

The **Quartimin criterion** is defined as in Equation (6.20) by

$$QMIN = \sum_{s<j=1}^{k}\sum_{i=1}^{p}\widetilde{\lambda}_{is}^2\widetilde{\lambda}_{ij}^2 \quad . \tag{6.29}$$

The rotated loadings are

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{T}\ , \tag{6.30}$$

where $\boldsymbol{T}$ is no longer orthogonal. Plugging in the rotated loadings in the model yields

$$\begin{aligned}\boldsymbol{\rho} - \boldsymbol{\Psi} &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\top} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}^{-1}\boldsymbol{\Phi}(\boldsymbol{T}^{-1})^{\top}\widetilde{\boldsymbol{\Lambda}}^{\top} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}^{-1}\mathrm{Cov}(\boldsymbol{f})(\boldsymbol{T}^{-1})^{\top}\widetilde{\boldsymbol{\Lambda}}^{\top} \\ &= \widetilde{\boldsymbol{\Lambda}}\mathrm{Cov}(\boldsymbol{T}^{-1}\boldsymbol{f})\widetilde{\boldsymbol{\Lambda}}^{\top} = \widetilde{\boldsymbol{\Lambda}}\mathrm{Cov}(\widetilde{\boldsymbol{f}})\widetilde{\boldsymbol{\Lambda}}^{\top} \quad .\end{aligned} \tag{6.31}$$

Die rotated factors $\widetilde{\boldsymbol{f}}$ are thus

$$\widetilde{\boldsymbol{f}} = \boldsymbol{T}^{-1}\boldsymbol{f}\ . \tag{6.32}$$

The transformation matrix $\boldsymbol{T}$ thus needs to be invertible.

A more flexible criterion is the **Oblimin criterion**

$$OBMIN = \sum_{s<j=1}^{k} \left(\sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \widetilde{\lambda}_{ij}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \sum_{i=1}^{p} \widetilde{\lambda}_{ij}^2 \right) \quad . \tag{6.33}$$

The choice $\gamma = 0$ yields the Quartimin criterion, while $\gamma = 1$ leads to the so-called *Covarimin criterion*.

Example 6.3.1 *Consider once more the exam data set. The initial factor loadings from a 2-factor solution are now rotated with varimax and quartimin, which gives the results presented in Table 6.2.*

Table 6.2: Rotated factor loadings for the exam data, using the varimax (VMAX) and quartimin (QMIN) criterion.

	VMAX		*QMIN*	
Variables	$\widetilde{\boldsymbol{\lambda}}_{.1}$	$\widetilde{\boldsymbol{\lambda}}_{.2}$	$\widetilde{\boldsymbol{\lambda}}_{.1}$	$\widetilde{\boldsymbol{\lambda}}_{.2}$
ME	0.271	0.675	-0.033	0.752
AG	0.354	0.680	0.078	0.707
LA	0.734	0.513	0.694	0.250
AN	0.742	0.319	0.821	-0.019
ES	0.704	0.285	0.789	-0.041

The rotated loadings from Table 6.2 are also visualized in Figure 6.3 (varimax) and 6.4 (quartimin). These plots clearly show that an orthogonal rotation does not lead to a simpler interpretation, while oblique rotation is more successful, and we obtain two factors which can be interpreted as geometrical and algorithmical thinking.

6.4 Estimation of the factor scores

Similar to PCA we are also interested in the scores, thus in the position of the observations in the new space of the (rotated) factors. Here we still treat the factors $\boldsymbol{f}$ as random variables, and below we consider two different options to estimate the factor scores.

6.4.1 Weighted least-squares estimation

In the R function `factanal()` this method is selected by `scores="Bartlett"`.

The factor model is

$$\boldsymbol{y} = \boldsymbol{\Lambda} \boldsymbol{f} + \boldsymbol{e} \tag{6.34}$$

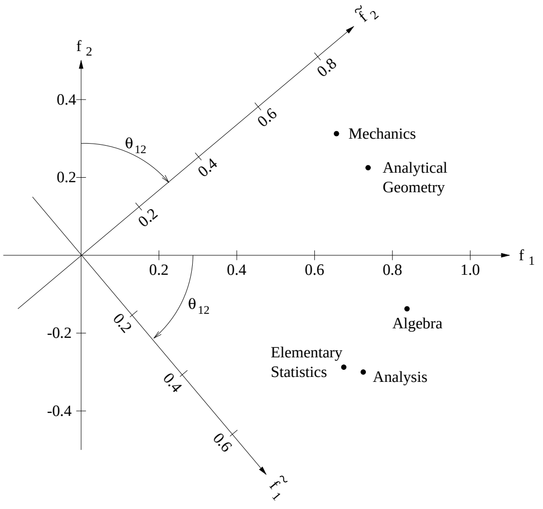


Figure 6.3: Orthogonal varimax rotation of the loadings, see Table 6.2.

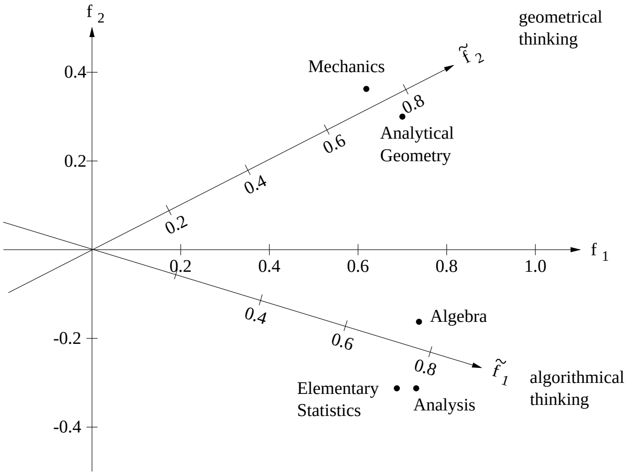


Figure 6.4: Oblique quartimin rotation of the loadings, see Table 6.2.

with the error term $\boldsymbol{e} = (e_1, \ldots, e_p)^\top$. This model could also be considered as a regression model, by regressing $\boldsymbol{y}$ on $\boldsymbol{\Lambda}$, with the regression coefficients $\boldsymbol{f}$. However, the model is heteroscedastic, since the error variances $Var(e_i) = \psi_i$ for $i = 1, \ldots, p$ are not necessarily equal. However, one can multiply the equation with weights $\boldsymbol{\Psi}^{-1/2}$, which yields

$$\boldsymbol{\Psi}^{-1/2}\boldsymbol{y} = \boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\Psi}^{-1/2}\boldsymbol{e}\ . \tag{6.35}$$

The covariance of the new error term is the identity matrix, and thus we have a homoscedastic model, and the least-squares estimator can be used, where the loadings and uniquenesses are considered as "true" values. This results in

$$\hat{\boldsymbol{f}} = \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{y} \tag{6.36}$$

for the estimated factors (random variables).

For a given $n \times p$ data matrix $\boldsymbol{X}$ we obtain the mean-centered and scaled matrix $\boldsymbol{Y}$ as a realization of $\boldsymbol{y}$, which can be substituted into Equation (6.36), to obtain the $n \times k$ matrix of estimated factor scores:

$$\hat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1} \tag{6.37}$$

6.4.2 Regression method

In the R function `factanal()` this method is selected by `scores="regression"`.

As for the previous method, the loadings $\boldsymbol{\Lambda}$ and uniquenesses $\boldsymbol{\Psi}$ are considered as known. Here we again consider a regression problem, but we regress the (unknown) factors $\boldsymbol{f}$ on $\boldsymbol{y}$ (multivariate regression):

$$\boldsymbol{f} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\delta} , \tag{6.38}$$

where $\boldsymbol{B}$ is the $k \times p$ matrix of regression coefficients, and $\boldsymbol{\delta}$ the error matrix. The least-squares estimator is

$$\widehat{\boldsymbol{B}} = \boldsymbol{f}\boldsymbol{y}^\top(\boldsymbol{y}\boldsymbol{y}^\top)^{-1} , \tag{6.39}$$

and the estimated factors are

$$\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{B}}\boldsymbol{y} = \boldsymbol{f}\boldsymbol{y}^\top(\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{y} . \tag{6.40}$$

Now we have the factors $\boldsymbol{f}$ again on the right-hand side, but we can plug in our factor model for $\boldsymbol{y}$:

$$\begin{aligned}\widehat{\boldsymbol{f}} &= \boldsymbol{f}(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{e})^\top(\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{y}\\ &= (\boldsymbol{f}\boldsymbol{f}^\top\boldsymbol{\Lambda}^\top + \boldsymbol{f}\boldsymbol{e}^\top)(\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{y}\\ &= \boldsymbol{f}\boldsymbol{f}^\top\boldsymbol{\Lambda}^\top(\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{y}\end{aligned}$$

Substituting the sample version yields

$$\begin{aligned}\widehat{\boldsymbol{F}} &= \boldsymbol{Y}(n-1)(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1}\boldsymbol{\Lambda}\frac{1}{n-1}\boldsymbol{F}^\top\boldsymbol{F}\\ &= \boldsymbol{Y}\boldsymbol{R}^{-1}\boldsymbol{\Lambda}\widehat{\boldsymbol{\Phi}} ,\end{aligned}$$

where $\widehat{\boldsymbol{\Phi}}$ is the estimated correlation matrix of the factors. In the orthogonal case, the factors are uncorrelated, and thus

$$\widehat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{R}^{-1}\boldsymbol{\Lambda} . \tag{6.41}$$

Example 6.4.1 *We use the rotated loadings after oblique quartimin rotation, see last example, and estimate the factor scores with the weighted least-squares method. The result is shown in Figure 6.5 in terms of a biplot. Note that for this biplot representation, the axes are forced to be orthogonal to each other. Large values on factor 1 refer to excellence in algorithmical thinking, and large values on factor 2 have the meaning of good geometrical thinking.*

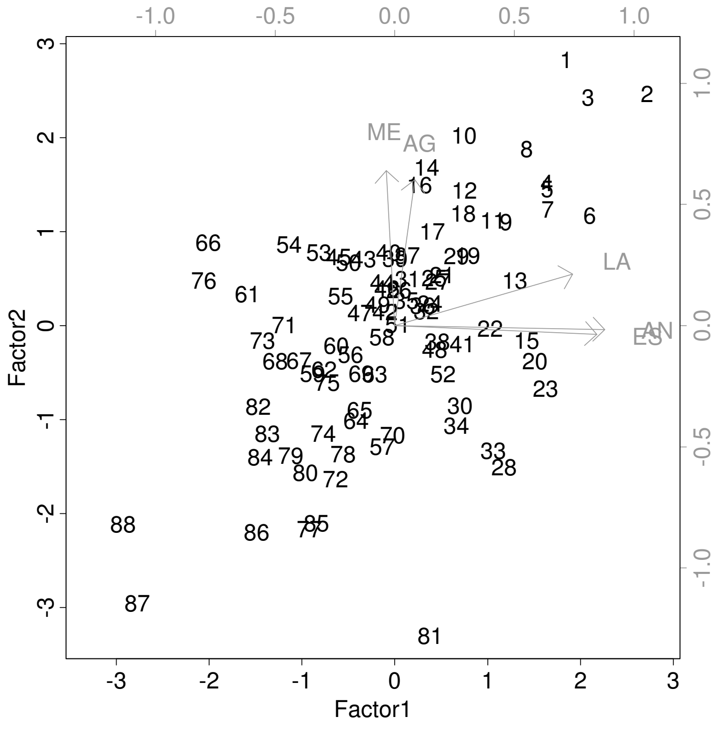


Figure 6.5: Biplot representation of the quartimin solution with estimated factor scores from weighted least-squares regression.

References

D.B. Clarkson and R.I. Jennrich. Quartic Rotation Criteria and Algorithms. *Psychometrika*, 53(2):251–259, 1988.

W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, New York, 1984.

L. Guttman. Multiple rectilinear prediction and the resolution into components I. *Psychometrika*, 5:75–99, 1940. Cynthia O, Williamsburg, Virginia.

H.H. Harman. *Modern Factor Analysis.* The University of Chicago Press, Chicago and London, 2nd edition, 1967.

R.I. Jennrich and P.F. Sampson. Rotation for Simple Loadings. *Psychometrika*, 31(3):313–323, 1966.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

H.F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

D.N. Lawley and A.E. Maxwell. *Factor Analysis as a Statistical Method.* Butterworths, London, 1963.

D. Revenstorf. *Lehrbuch der Faktorenanalyse.* Verlag W. Kohlhammer, Stuttgart, 1976.

G.A.F. Seber. *Multivariate observations.* John Wiley & Sons, New York, 1984.

L.L. Thurstone. Second-order factors. *Psychometrika*, 9:71–100, 1944. Cynthia O, Williamsburg, Virginia.

L.L. Thurstone. *Multiple factor analysis.* University Press Chicago, Chicago, sixth edition, 1961.

K. Überla. *Faktorenanalyse.* Springer, Berlin, 1971.

Chapter 7

Correlation analysis

In statistical data analysis, one is often interested in determining relationships and dependencies of features. If one also wants to measure the existence and the strength of dependencies, *correlation analysis* can be used. The correlation measures the linear relationship between features.

7.1 Multiple correlation analysis

The *multiple correlation* is a measure of the dependency of a feature x on a p - dimensional feature $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$. We assume that both x and $y_1, \ldots, y_p$ are random variables with a joint distribution. The mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of this distribution (not necessarily normal distribution) are

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \\ \boldsymbol{\sigma}_{\boldsymbol{y}x} & \boldsymbol{\Sigma}_{\boldsymbol{yy}} \end{pmatrix} ,$$

where

$$E(x) = \mu_x \qquad \mathrm{Var}(x) = \sigma_{xx}$$

$$E(\boldsymbol{y}) = \boldsymbol{\mu}_{\boldsymbol{y}} \qquad \mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{yy}}$$

and

$$\mathrm{Cov}(\boldsymbol{y}, x) = \boldsymbol{\sigma}_{\boldsymbol{y}x} \qquad \mathrm{Cov}(x, \boldsymbol{y}) = \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top .$$

If x is now predicted by $\boldsymbol{y}$ (linear), the error is analogous to regression analysis

$$x - a_0 - a_1 y_1 - \ldots - a_p y_p .$$

This error is random, and therefore one would like to choose the coefficients a_0 and $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ such that the *mean squared error* (MSE)

$$\mathrm{MSE} = E(x - a_0 - \boldsymbol{a}^\top \boldsymbol{y})^2$$

is minimal. However, this MSE depends on the joint distribution of x and $\boldsymbol{y}$, and thus on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Theorem 7.1.1 *The linear prediction function* $a_0 + \boldsymbol{a}^\top \boldsymbol{y}$ *with the coefficients*

$$\boldsymbol{a} = \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x} \quad \text{and} \quad a_0 = \mu_x - \boldsymbol{a}^\top \boldsymbol{\mu}_{\boldsymbol{y}}$$

has minimal MSE of all linear prediction functions of x. *It holds that*

$$MSE = E(x - a_0 - \boldsymbol{a}^\top \boldsymbol{y})^2 = E(x - \mu_x - \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}))^2 = \sigma_{xx} - \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x} \ .$$

Furthermore,

$$a_0 + \boldsymbol{a}^\top \boldsymbol{y} = \mu_x + \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}})$$

is the linear prediction function that has maximum correlation with x, *namely*

$$Corr(x, a_0 + \boldsymbol{a}^\top \boldsymbol{y}) = \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}} \boldsymbol{a}}{\sigma_{xx}}} = \sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x}}{\sigma_{xx}}} \ .$$

Proof: See, for example, Johnson and Wichern (1998). For the proof of the minimal MSE one needs to re-express the squared expectation, and then it is visible that with these choices of $\boldsymbol{a}$ and a_0 one obtains a minimum of the MSE. For the maximum correlation one can make use of the

Extended Cauchy-Schwarz Inequality: Let $\mathbf{b}$ and $\boldsymbol{d}$ be two vectors, and $\boldsymbol{B}$ a positive definite matrix. Then

$$(\boldsymbol{b}^\top \boldsymbol{d})^2 \leq (\boldsymbol{b}^\top \boldsymbol{B} \boldsymbol{b})(\boldsymbol{d}^\top \boldsymbol{B}^{-1} \boldsymbol{d}) \ , \tag{7.1}$$

with equality if and only if $\boldsymbol{b} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ (or $\boldsymbol{d} = c\boldsymbol{B}\boldsymbol{b}$), for a constant c.

The correlation between x and the best linear prediction function is called *multiple correlation coefficient* (of the population)

$$\rho_{x,\boldsymbol{y}} = +\sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x}}{\sigma_{xx}}} \ .$$

Its squared form $\rho_{x,\boldsymbol{y}}^2$ is called *multiple coefficient of determination* (of the population), and it indicates how well the feature x is explained by the properties $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$.

If correlations are given instead of the covariances, then the multiple correlation coefficient can also be defined as

$$\rho_{x,\boldsymbol{y}}^2 = \boldsymbol{\rho}_{\boldsymbol{y}x}^\top \boldsymbol{\rho}_{\boldsymbol{yy}}^{-1} \boldsymbol{\rho}_{\boldsymbol{y}x} \quad .$$

For a specific sample, one can write the theoretical quantities by the corresponding realizations, and hence obtain for the multiple correlation coefficient

$$r_{x,\boldsymbol{y}}^2 = \boldsymbol{R}_{\boldsymbol{y}x}^\top \boldsymbol{R}_{\boldsymbol{yy}}^{-1} \boldsymbol{R}_{\boldsymbol{y}x} \quad .$$

A test for the hypothesis that the multiple correlation is zero is equivalent to making all bivariate correlations zero. So, we test the hypothesis

$$H_0 : \rho_{x,\boldsymbol{y}} = 0 \ (= \rho_{xy_1} = \ldots = \rho_{xy_p})$$

against the alternative hypothesis

$$H_1 : \exists\, i \in \{1, \ldots, p\} \quad \text{with} \quad \rho_{xy_i} \neq 0$$

at the significance level α. If we can assume multivariate normal distribution, and n is the number of observations in the sample, the test statistic

$$F = \frac{(n-1-p)\ r^2_{x,\boldsymbol{y}}}{p\ (1-r^2_{x,\boldsymbol{y}})}$$

has an $F_{p,n-1-p}$-distribution and the null hypothesis is rejected at the significance level of α, if

$$F > F_{p,n-1-p;1-\alpha}.$$

Example 7.1.1 *The R package* ***`ISLR`*** *contains the data set* ***`Auto`*** *with different characteristics of cars. We focus on the European cars (****`origin=2`****), and on the subset of variables shown in Figure 7.1. We want to know if there is a relationship between the year of production and the car characteristics. The plot already indicates such a relationship.*

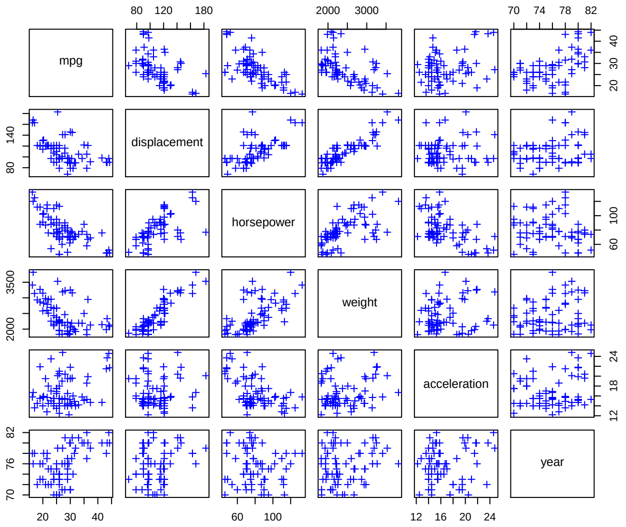


Figure 7.1: Car characteristics and year of production.

Thus, x is the variable ***`year`****, and $\boldsymbol{y}$ represents all remaining variables. We are computing the corresponding correlation matrices. For instance,*

$$\boldsymbol{R}_{x\boldsymbol{y}} = (0.50, 0.21, -0.13, 0.18, 0.18)\ ,$$

and the resulting empirical squared multiple correlation coefficient is $r^2_{x,\boldsymbol{y}} = 0.58$.

A test for uncorrelatedness gives the value 16.9 for the F-statistic, with a p-value of essentially zero. This means that the multiple correlation is significantly different from zero.

From the plot in Figure 7.1 we can see that the correlation might essentially be driven by `mpg`*. Also when looking at the linear predictor, we can see that according to the values of*

$$\boldsymbol{a} = (0.87, 0.32, -0.24, 0.54, -0.32)^\top \ ,$$

the variable `mpg` *has the strongest impact. Over the years, the car production moved towards more energy-saving cars. If we omit this variable, we obtain a squared multiple correlation coefficient of only 0.18. However, a test for uncorrelatedness still yields a p-value of 0.011.*

7.2 Canonical correlation analysis

In canonical correlation analysis we are interested in the linear dependence between two groups of variables. As it turns out, this dependence can no longer be expressed by a single correlation coefficient, but results in a subspace that describes the linear dependence between the groups.

Theorem 7.2.1 *Let $\boldsymbol{x}$ be a p-dimensional and $\boldsymbol{y}$ a q-dimensional random variable $(p \leq q)$ with the expected values*

$$E(\boldsymbol{x}) = \boldsymbol{\mu}_1 \qquad \textit{and} \qquad E(\boldsymbol{y}) = \boldsymbol{\mu}_2 \ .$$

The covariance matrices $\boldsymbol{\Sigma}_{ij}$ with $i, j = 1, 2$ are defined by

$$\begin{aligned} \boldsymbol{\Sigma}_{11} &= E[(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top] \quad , \\ \boldsymbol{\Sigma}_{22} &= E[(\boldsymbol{y}-\boldsymbol{\mu}_2)(\boldsymbol{y}-\boldsymbol{\mu}_2)^\top] \quad \textit{and} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top &= E[(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{y}-\boldsymbol{\mu}_2)^\top] \end{aligned}$$

and have full rank. We consider the linear combinations $\varphi = \boldsymbol{a}^\top\boldsymbol{x}$ and $\eta = \boldsymbol{b}^\top\boldsymbol{y}$, where $\boldsymbol{a}$ is a p-dimensional and $\boldsymbol{b}$ is a q-dimensional vector.

Then the simple correlation between φ and η is given by

$$\max_{\boldsymbol{a},\boldsymbol{b}} \textit{Corr}(\varphi, \eta) = \rho_1 \ .$$

The maximum is achieved by the linear combinations

$$\varphi_1 = \underbrace{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_1^\top} \ \boldsymbol{x} \qquad \textit{and} \qquad \eta_1 = \underbrace{\boldsymbol{f}_1^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_1^\top} \ \boldsymbol{y} \ ,$$

which are referred to as **the first pair of canonical variables**. *ρ_1 is called* **the first canonical correlation coefficient**.

The ***k*-th pair of canonical variables** *$(k = 2, 3, \ldots, p)$ is given by*

$$\varphi_k = \underbrace{\boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_k^\top} \ \boldsymbol{x} \qquad \textit{and} \qquad \eta_k = \underbrace{\boldsymbol{f}_k^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_k^\top} \ \boldsymbol{y}$$

and maximize

$$Corr(\varphi_k, \eta_k) = \rho_k$$

over all linear combinations that are uncorrelated with the previous $1, 2, \ldots, k-1$ *canonical variables.* ρ_k *is called* $\boldsymbol{k}$**-th canonical correlation coefficient**.

$\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$ *are eigenvalues of the matrix* $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ *and* $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ *are the respective eigenvectors (dimension p).*

$\rho_1^2, \rho_2^2, \ldots, \rho_p^2$ *are the p largest eigenvalues of the matrix* $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$ *with the respective eigenvectors* $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p$ *(dimension q). Every* $\boldsymbol{f}_i$ *is proportional to* $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i$.

The canonical variables have the following properties

$$\begin{aligned}
Var(\varphi_k) &= Var(\eta_k) = 1 \\
Cov(\varphi_k, \varphi_l) &= Corr(\varphi_k, \varphi_l) = 0 \qquad k \neq l \\
Cov(\eta_k, \eta_l) &= Corr(\eta_k, \eta_l) \ = 0 \qquad k \neq l \\
Cov(\varphi_k, \eta_l) &= Corr(\varphi_k, \eta_l) \ = 0 \qquad k \neq l
\end{aligned}$$

for $k, l = 1, 2, \ldots, p$*; and with the notation* $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$ *and* $\boldsymbol{\rho} = Diag(\rho_1, \ldots, \rho_p)$*, it follows that*

$$Cov\begin{pmatrix}\boldsymbol{\varphi} \\ \boldsymbol{\eta}\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}\,\boldsymbol{\rho} \\ \boldsymbol{\rho}\,\boldsymbol{I}\end{pmatrix} .$$

Proof: See, for example, Johnson and Wichern (1998). For maximizing the correlation, one can re-express the covariance by using the

Cauchy-Schwarz Inequality: Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be two vectors, then

$$(\boldsymbol{\alpha}^\top\boldsymbol{\beta})^2 \leq (\boldsymbol{\alpha}^\top\boldsymbol{\alpha})(\boldsymbol{\beta}^\top\boldsymbol{\beta}) ,$$

with equality if and only if $\boldsymbol{\beta} = c\boldsymbol{\alpha}$, for a constant c.

Then, the maximum can be obtained by the

Theorem: Consider a positive definite matrix $\boldsymbol{B}$, and let λ_1 be the largest eigenvalue of $\boldsymbol{B}$ to the normed eigenvector $\boldsymbol{e}_1$. Then

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}} = \lambda_1$$

and the maximum will be attained for $\boldsymbol{x} = \boldsymbol{e}_1$.

Remark: The coefficients $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ that solve the maximization problem of Theorem 7.2.1 can also be determined in the following way:

ρ_i^2 in Theorem 7.2.1 is also eigenvalue of $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ with eigenvector $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i = \boldsymbol{a}_i$,

ρ_i^2 in Theorem 7.2.1 is also eigenvalue of $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ with eigenvector $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{f}_i = \boldsymbol{b}_i$ with $i = 1, \ldots, p$.

The canonical correlation coefficients are invariant to linear transformations, which will be shown below.

Theorem 7.2.2 *Let $\boldsymbol{x}^* = \boldsymbol{U}^\top\boldsymbol{x}+\boldsymbol{u}$ and $\boldsymbol{y}^* = \boldsymbol{V}^\top\boldsymbol{y}+\boldsymbol{v}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are regular matrices of dimension $p \times p$ and $q \times q$, respectively, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are fixed vectors of length p and q, respectively. Then, it holds:*

(a) The canonical correlation coefficients between $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ are the same as those between $\boldsymbol{x}$ and $\boldsymbol{y}$.

(b) The linear combinations that maximize the linear dependence between $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ are given by

$$\boldsymbol{a}_i^* = \boldsymbol{U}^{-1}\boldsymbol{a}_i \qquad \textit{and} \qquad \boldsymbol{b}_i^* = \boldsymbol{V}^{-1}\boldsymbol{b}_i \qquad \textit{for} \quad i = 1, \ldots, p\ ,$$

where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the respective linear combinations for $\boldsymbol{x}$ and $\boldsymbol{y}$.

Proof: If, instead of the matrices $\boldsymbol{\Sigma}_{ij}$ for $\boldsymbol{x}$ and $\boldsymbol{y}$, the matrices $\boldsymbol{\Sigma}_{ij}^*$ for $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ $(i = 1, 2)$ are calculated, one obtains for $\boldsymbol{M}_1^* = \boldsymbol{\Sigma}_{11}^{*-1}\boldsymbol{\Sigma}_{12}^*\boldsymbol{\Sigma}_{22}^{*-1}\boldsymbol{\Sigma}_{21}^*$:

$$\boldsymbol{M}_1^* = (\boldsymbol{U}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Sigma}_{12}\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Sigma}_{21}\boldsymbol{U} = \boldsymbol{U}^{-1}\boldsymbol{M}_1\boldsymbol{U}$$

with $\boldsymbol{M}_1 = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Using Theorem 1.4.1 with $\boldsymbol{C} = \boldsymbol{U}^{-1}$ resp. $\boldsymbol{C} = \boldsymbol{V}^{-1}$ proves the statement. □

Application: If one chooses $\boldsymbol{U} = (diag\boldsymbol{\Sigma}_{11})^{-1/2}$ and $\boldsymbol{V} = (diag\boldsymbol{\Sigma}_{22})^{-1/2}$, one obtains the correlation matrices $\boldsymbol{\rho}_{ij}$ instead of $\boldsymbol{\Sigma}_{ij}$. Instead of $\boldsymbol{M}_1$, one obtains $\boldsymbol{M}_1^* = \boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}$.
Therefore, in case the covariance matrices and not the correlation matrices $\boldsymbol{\rho}_{ij}$ are given, then the following holds:

(a) the canonical correlation coefficients are determined by the roots of the eigenvalues of $\boldsymbol{M}_1^*$ (or $\boldsymbol{M}_2^*$),

(b) the linear combinations $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are determined by the transformations $\boldsymbol{a}_i = \boldsymbol{U}\boldsymbol{a}_i^*$ resp. $\boldsymbol{b}_i = \boldsymbol{V}\boldsymbol{b}_i^*$, where $\boldsymbol{a}_i^*$ resp. $\boldsymbol{b}_i^*$ are the eigenvectors of $\boldsymbol{M}_1^*$ resp. $\boldsymbol{M}_2^*$. However, the variances of x_i and y_i must be known.

Example 7.2.1 *Of course, all previously formulated theorems can be transferred to a sample version. To make the difference between population and sample more clear, we will write $\boldsymbol{S}_{ij}$ instead of $\boldsymbol{\Sigma}_{ij}$, $\boldsymbol{R}_{ij}$ instead of $\boldsymbol{\rho}_{ij}$ and r_i instead of ρ_i.*

As an example, we consider the data set `schooldata` from the R package `FRB`. It consists of 70 observations (persons) on the following 8 variables:

- *education: education level of mother as measured in terms of percentage of high school graduates among female parents*
- *occupation: highest occupation of a family member according to a pre-arranged rating scale*
- *visit: parental visits index representing the number of visits to the school site*

- *counseling: parent counseling index calculated from data on time spent with child on school-related topics such as reading together, etc.*
- *teacher: number of teachers at a given site*
- *reading: total reading score as measured by the Metropolitan Achievement Test*
- *mathematics: total mathematics score as measured by the Metropolitan Achievement Test*
- *selfesteem: Coopersmith Self-Esteem Inventory, intended as a measure of self-esteem*

Observations 44 and 59 haev quite extreme values in some variables, and thus they are excluded. Figure 7.3 shows a scatter plot of the data set.

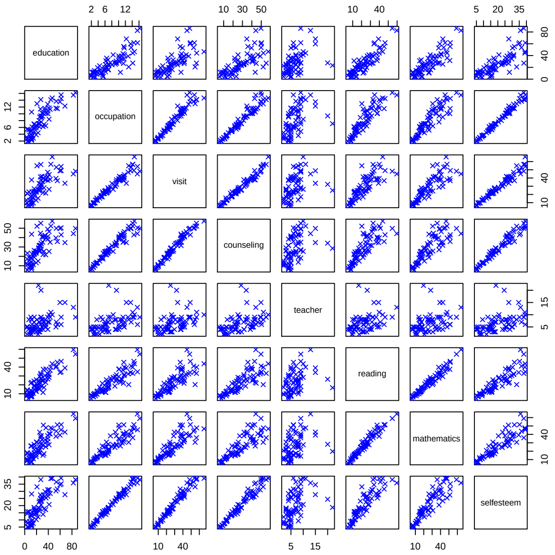


Figure 7.2: Scatter plot of the school data set.

We are interested in the relationship between the last 3 and the first 5 variables. The corresponding data sets are centered and scaled first. This has the advantage that the coefficients for the linear combination can be directly compared (similar to loadings in PCA).

The R function `cancor()` allows to compute the coefficients for the canonical variables, as well as the canonical correlation coefficients. For example, the coefficients for the first canonical variable φ_1 are $\boldsymbol{a}_1^* = (-0.011, 0.004, 0.129)^\top$, and those for η_1 are $\boldsymbol{b}_1^* = (-0.006, 0.085, 0.037, 0.005, 0.001)^\top$. The resulting first canonical correlation coefficient is 0.995. This means that there is a very strong linear relationship, and this is driven essentially by the relationship between selfesteem and occupation/visits.

Figure 7.3 shows the 3 pairs of canonical variables, Also for the second and the third pair there is a clear relationship ($r_2 = 0.744$, $r_3 = 0.270$).

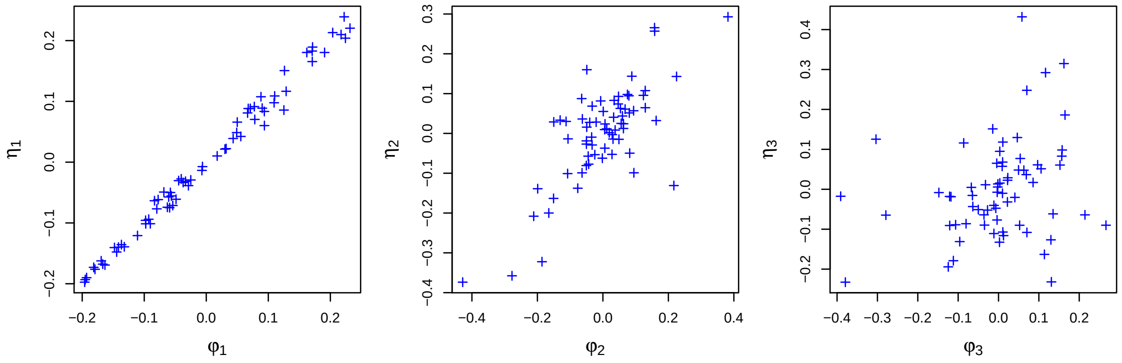


Figure 7.3: All three pairs of canonical variables for the school data.

In canonical correlation analysis, there are a number of **tests**, of which only selected ones are mentioned here, because the distributions of the estimators are very complicated. Assuming that the data are normally distributed, the following applies:

(a) A likelihood ratio test for the hypothesis $H_0 : \boldsymbol{\Sigma}_{12} = \boldsymbol{O}$, i.e. for the hypothesis that $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated, is given by the test statistic

$$\lambda^{2/n} = |\boldsymbol{I} - \boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}| = \prod_{i=1}^{p}(1 - r_i^2).$$

This test statistic has a $\Lambda(q, n-1-p, p)$ - Wilks distribution. n is the sample size and $r_1, \ldots, r_p$ are the sample canonical correlation coefficients.

(b) Since the Wilks distribution as a product of beta distributions is relatively complex, it can be approximated (with Bartlett's approximation) for large sample sizes n by

$$-\left[n - \frac{1}{2}\Big(p+q+3\Big)\right] \ln \prod_{i=1}^{p}(1 - r_i^2) \sim \chi^2_{pq} \quad .$$

(c) A similar test statistic can also be formulated for the hypothesis that only s of the canonical correlation coefficients are non-zero, namely by

$$-\left[n - \frac{1}{2}\Big(p+q+3\Big)\right] \ln \prod_{i=s+1}^{p} (1 - r_i^2) \sim \chi^2_{(p-s)(q-s)} \quad .$$

Example 7.2.2 *For the school data example we have $n = 68$, $p = 3$ and $q = 5$. The test statistic for the hypothesis $\rho_1 = \rho_2 = \rho_3 = 0$ according to (b) is given by*

$$-\left(68 - \frac{11}{2}\right) \ln \left[(1 - 0.995^2)(1 - 0.774^2)(1 - 0.270^2)\right] \approx 862 \ .$$

This value is well above $\chi^2_{15;0.95} \approx 25$, so the hypothesis is rejected.

If only the hypothesis $\rho_3 = 0$ is tested, then the test statistic according to (c) is given by

$$-\left(68 - \frac{11}{2}\right) \ln(1 - 0.270^2) \approx 39.3 \ .$$

This is still significant when tested against $\chi^2_{8;0.95} \approx 15.5$, so the hypothesis is again rejected.

Example 7.2.3 *The maximization problem stated in Theorem 7.2.1 of finding linear combinations which maximize a bivariate correlation measure can also be viewed as a projection-pursuit problem, similar as in PCA. The task is to find a projection direction in the x-space and a projection direction in the y-space, such that the resulting correlation gets as large as possible. An efficient algorithm allows to "scan" the corresponding spaces in order to identify the coefficients for the directions. The second pair can be found by imposing the restrictions of uncorrelatedness.*

This idea is implemented in the R package ccaPP, see Alfons et al. (2016, 2017). The advantage of this approach is that any bivariate correlation measure can be considered, also a robust or a non-parametric measure.

We demonstrate this with the school data set, but not deleting any observations, and the considered correlation measure is the non-parametric Spearman correlation coefficient.

```
library(ccaPP)
X <- schooldata[,6:8]
Y <- schooldata[,1:5]
res2 <- CCAgrid(X,Y,k=3,method="spearman",standardize=TRUE)
```

Figure 7.4 shows the resulting 3 pairs of canonical variables, with the canonical correlations $r_1 = 0.993$, $r_2 = 0.642$, $r_3 = 0.417$. It seems that there is no big difference to the previous analysis, although one would have to go more into detail with a comparison.

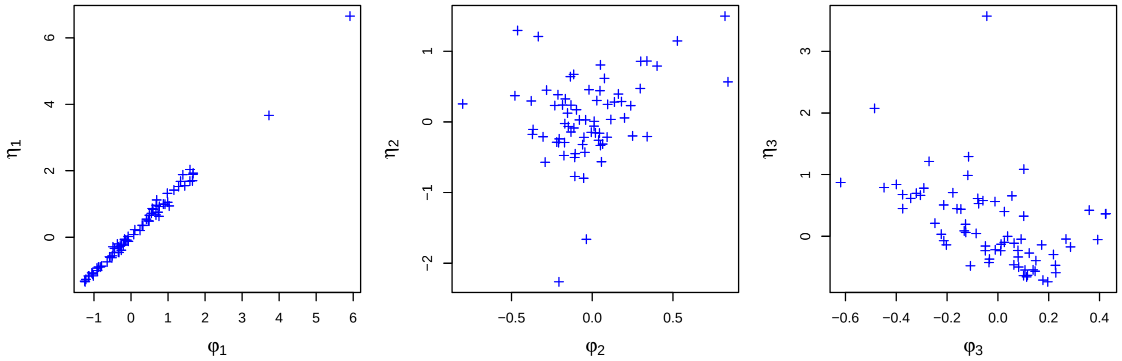


Figure 7.4: All three pairs of canonical variables, using the Spearman correlation.

Finally, this package also allows to test for uncorrelatedness. Here, a permutation test is carried out, which does not rely on a distributional assumption. The idea is to *permute the observations of one data set, by keeping the other data set unpermuted. Then the canonical correlations are estimated, and this is done many times. Finally, a p-value is determined as the fraction of bootstrap correlation results exceeding the canonical correlation of the unpermuted data.*

```
library(ccaPP)
resp <- permTest(X,Y,R=1000,nCores=8,method="pearson")
```

Here, the p-value is zero. This means that non of the first canonical correlation coefficients of the permuted data was higher than that of the original data – which is not a surprise.

References

A. Alfons, C. Croux and P. Filzmoser. Robust maximum association between data sets: The R Package ccaPP. *Austrian Journal of Statistics*, 45(1), 71–79, 2016

A. Alfons, C. Croux and P. Filzmoser. Robust maximum association estimators. *Journal of the American Statistical Association*, 112 (517), 436–445, 2017.

W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, New York, 1984.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

G.A.F. Seber. *Multivariate observations.* John Wiley & Sons, New York, 1984.

Chapter 8

Discriminant Analysis

8.1 Introduction

The term discriminant analysis was coined by Fisher (1938). It is a multivariate method that deals on the one hand with the classification of different object groups and on the other hand with the assignment of new objects to previously determined groups. In the former case, the attempt is made to capture the differences of the objects which are known to originate from two or more populations, either graphically or algebraically. One is thus looking for a discriminant function that allows the best possible separation. In the second case, one would like to divide the objects into two or more groups. The goal is then to classify new objects by means of defined rules. The above cases are often directly related, because a function that separates objects can also be used to classify new objects or vice versa.

8.2 Reflections on classification rules

The objects are classified on the basis of measurements of p underlying random variables $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. Assuming that there are two groups, the objects to be measured are then divided into the classes π_1 and π_2 respectively. Let the sum of the values of the first class be the population of the $\boldsymbol{x}$-values of π_1, and that of the second class the population of the $\boldsymbol{x}$-values of π_2. The two populations are then described by the probability distributions $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$.

For example, the two populations π_1 and π_2 could be given by compositions by J.S. Bach and L.v. Beethoven. The variables $\boldsymbol{X}$ to be measured are frequencies of different chords or tone sequences. It is then of interest to find a discriminatory function that describes how the two composers differ on the basis of their works.

Let Ω be the entire sampling space, i.e. the space containing all observations $\boldsymbol{x}$. Each object $\boldsymbol{x}$ must come from either population π_1 or π_2. Furthermore, let R_1 be the space of observations $\boldsymbol{x}$ to which we assign the objects of π_1, and R_2 the space to which the remaining objects of π_2 are assigned. Ω is the union of R_1 and R_2.

It may happen in the case of group assignment that objects which actually belong to population π_1 are falsely classified as π_2. If the probability functions $f_1(\boldsymbol{x})$ and

$f_2(\boldsymbol{x})$ are known, this probability of incorrect assignment can be calculated as a conditional probability $P(2|1)$ by

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.1}$$

Conversely, it may be the case that objects originating from population π_2 are erroneously assigned to π_1. The corresponding probability is

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.2}$$

The integral in (8.1) describes the volume of the density function $f_1(\boldsymbol{x})$ over the region R_2, or analogously for (8.2).

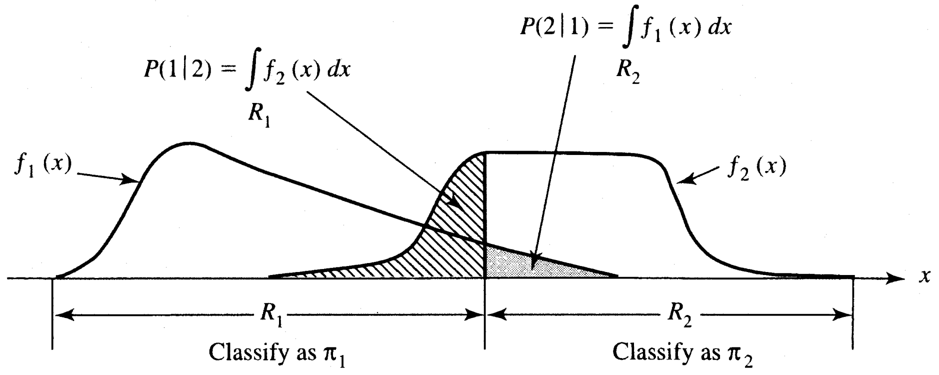


Figure 8.1: Probabilities of misclassification

Let p_1 be the probability that the objects come from π_1 (*prior probability*), and p_2 those for π_2, where $p_1 + p_2 = 1$ must hold. Then, the following probabilities can be calculated by applying the formula for conditional probabilities:

$$\begin{aligned}
P(\text{Observation correctly classified as } \pi_1) &= P(\boldsymbol{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \\
P(\text{Observation incorrectly classified as } \pi_1) &= P(\boldsymbol{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2 \\
P(\text{Observation correctly classified as } \pi_2) &= P(\boldsymbol{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2 \\
P(\text{Observation incorrectly classified as } \pi_2) &= P(\boldsymbol{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1
\end{aligned}$$

Missclassification is often directly associated with costs. Of course, the costs are 0 if correctly classified. They are $c(1|2)$ if an observation of π_2 is wrongly classified as π_1. And the cost is $c(2|1)$ if observations of π_1 are erroneously classed as π_2. Together with the probabilities for misclassification, the expected costs for misclassification (ECM) can now be calculated as

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \ . \tag{8.3}$$

The goal of a classification rule is to keep ECM as small as possible.

Theorem 8.2.1 *A classification rule that minimizes ECM is as follows: The set R_1 is defined for observations $\boldsymbol{x}$, for which the following applies:*

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{8.4}$$

The set R_2 is defined for observations $\boldsymbol{x}$, for which the following applies:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{8.5}$$

Proof: Since $\Omega = R_1 \cup R_2$, we have:

$$1 = \int_\Omega f_1(\boldsymbol{x})d\boldsymbol{x} = \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x} + \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}$$

Thus,

$$\begin{aligned} ECM &= c(2|1)p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + c(1|2)p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{R_1} \Big[c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x})\Big] d\boldsymbol{x} + c(2|1)p_1 \end{aligned}$$

The expressions $p_1, p_2, c(1|2), c(2|1)$ within $\Big[\ \cdot\ \Big]$ are all non-negative, and $f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$ are also non-negative for all $\boldsymbol{x}$. Thus, ECM is minimal if R_1 includes those $\boldsymbol{x}$ for which $\Big[\ \cdot\ \Big] \leq 0$, and excudes those $\boldsymbol{x}$ for which $\Big[\ \cdot\ \Big] > 0$. Thus, R_1 contains those $\boldsymbol{x}$ for which $c(1|2)p_2 f_2(\boldsymbol{x}) \leq c(2|1)p_1 f_1(\boldsymbol{x})$. □

Theorem 8.2.1 leads to the following special cases:

(a) $p_1 = p_2$:

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2) = c(2|1)$:

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1} \qquad R_2 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{p_2}{p_1} \tag{8.6}$$

(c) $p_1 = p_2$ and $c(1|2) = c(2|1)$:

$$R_1 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1 \qquad R_2 : \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < 1$$

However, other criteria can also be used to create a classification rule. For example, one could choose R_1 and R_2 so that the total probability of misclassification (TPM) is minimal, i.e.

$$\begin{aligned} TPM &= P(\text{Misclassification of an observation from } \pi_1 \text{ or } \pi_2) \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \qquad (8.7) \\ &= p_1 P(2|1) + p_2 P(1|2) \ . \qquad (8.8) \end{aligned}$$

Assuming that the cost of misclassification is the same, one immediately realizes that minimizing (8.8) is equivalent to minimizing (8.3).

8.3 The two-group case

We limit ourselves here to multivariate normally distributed populations. Thus, $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are density functions of multivariate normal distributions with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively.

8.3.1 The special case $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

The joint density function of the random variable $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ for the populations π_1 and π_2 are given by

$$f_i(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right\} \quad \text{for } i = 1, 2 \ . \qquad (8.9)$$

If the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known, then according to (8.4), the region that minimizes ECM is:

$$\begin{aligned} R_1 :\ & \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right\} \\ & \qquad \geq \left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \\ R_2 :\ & \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right\} \\ & \qquad < \left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \qquad (8.10) \end{aligned}$$

The following classification rule can be specified using these regions R_1 and R_2:

Theorem 8.3.1 *Let π_1 and π_2 be normally distributed populations with the same covariance $\boldsymbol{\Sigma}$. Then the classification rule for minimizing ECM is:*
An observation $\boldsymbol{x}_0$ is assigned to π_1, if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right) \ . \qquad (8.11)$$

Otherwise $\boldsymbol{x}_0$ is assigned to π_2.

Proof: Since all quantities in (8.10) are non-negative, the rules do not change by logarithms. Further, it holds:

$$\begin{aligned}
&-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\\
&=\tfrac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\tfrac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2-\tfrac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\tfrac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\\
&=\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\tfrac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2
\end{aligned}$$

The above transformation can be done because the individual summands are scalars and therefore $a = a^\top$. If one now expands with $\frac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$, we get:

$$\begin{aligned}
&(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\tfrac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\\
&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\tfrac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\tfrac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\\
&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\tfrac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)
\end{aligned}$$

This immediately implies the statement. □

Note that the discriminant function defined in Theorem 8.3.1 is *linear* in $\boldsymbol{x}$.

The quantities $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are usually not given, and they must be estimated from the sample. Let n_1 be the sample size of the random variable $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ for the population π_1, and n_2 the number of measurements for π_2. The corresponding data matrices of the dimensions $(n_1 \times p)$ and $(n_2 \times p)$ are denoted by

$$\boldsymbol{X}_1 = \begin{pmatrix} \boldsymbol{x}_{11}^\top \\ \boldsymbol{x}_{12}^\top \\ \vdots \\ \boldsymbol{x}_{1n_1}^\top \end{pmatrix} \qquad \boldsymbol{X}_2 = \begin{pmatrix} \boldsymbol{x}_{21}^\top \\ \boldsymbol{x}_{22}^\top \\ \vdots \\ \boldsymbol{x}_{2n_2}^\top \end{pmatrix} \tag{8.12}$$

The arithmetic mean vectors and empirical covariance matrices are then given by

$$\begin{aligned}
\bar{\boldsymbol{x}}_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{x}_{1j} & \boldsymbol{S}_1 &= \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)^\top\\
\bar{\boldsymbol{x}}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{x}_{2j} & \boldsymbol{S}_2 &= \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)^\top
\end{aligned}$$

However, since it was assumed that π_1 and π_2 are populations with the same covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are combined into one pooled covariance matrix

$$\begin{aligned}
\boldsymbol{S}_{pooled} &= \left(\frac{n_1-1}{(n_1-1)+(n_2-1)}\right)\boldsymbol{S}_1+\left(\frac{n_2-1}{(n_1-1)+(n_2-1)}\right)\boldsymbol{S}_2\\
&= \frac{1}{n_1+n_2-2}\sum_{l=1}^{2}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)^\top\ .
\end{aligned}$$

$\boldsymbol{S}_{pooled}$ is an unbiased estimate of $\boldsymbol{\Sigma}$, and it corresponds to a covariance estimation of the group-centered data.

If the estimates are used in (8.11), then the following sample classification rule results:

Theorem 8.3.2 *An observation $\boldsymbol{x}_0$ is assigned to π_1, if*

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right) \ . \tag{8.13}$$

Otherwise $\boldsymbol{x}_0$ is assigned to π_2.

Since the summands in (8.13) are scalars, the rule for the case that in (8.13) it holds that $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$ ($\ln(1) = 0$) can be simplified in the following way:

$$\hat{y} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} \tag{8.14}$$

is evaluated at $\boldsymbol{x}_0$ and then compared with the number

$$\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \ ,$$

where

$$\bar{y}_1 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \bar{\boldsymbol{x}}_1 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1$$

and

$$\bar{y}_2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \bar{\boldsymbol{x}}_2 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2 \ .$$

Thus, the rule in case of p-dimensional features reduces the decision to a 1-dimensional variable y, which results from the corresponding linear combinations of the observations of π_1 and π_2.

Example 8.3.1 *This example is from a study by Bouma et al. (1975), which deals with the detection of hemophilia. There are several types of this disease, and the study aims to identify carriers of hemophilia A. Two variables were measured*

$$\begin{aligned} X_1 &= \log_{10}(\textit{AHF activity}) \\ X_2 &= \log_{10}(\textit{AHF-similar antigen}) \ , \end{aligned}$$

where AHF denotes the anti-hemophilia factor. The values of two groups of women were measured, with the first group with $n_1 = 30$ women not carrying the hemophilia gene and the second group with $n_2 = 22$ women carrying hemophilia A. The measurements are illustrated in Figure 8.2. In addition, the estimated contour lines with center $\bar{\boldsymbol{x}}_1$ resp. $\bar{\boldsymbol{x}}_2$ are shown, which contain 50% and 95% of the observations, respectively. The estimated means are

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} \quad \textit{and} \quad \bar{\boldsymbol{x}}_2 = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix}$$

and the pooled covariance matrix

$$\boldsymbol{S}_{pooled}^{-1} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix} \ .$$

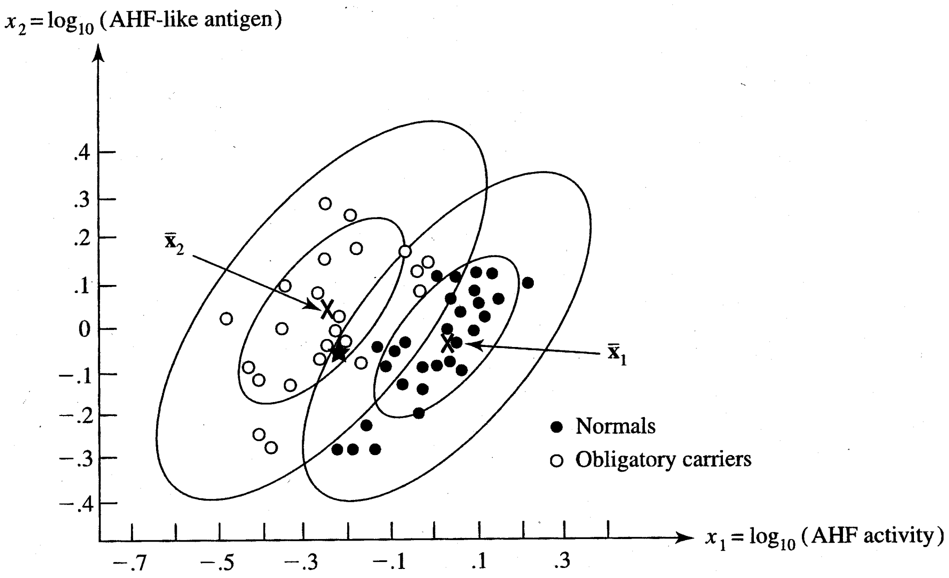


Figure 8.2: Plot of the hemophilia data

Assuming that the costs of misclassification and the prior probabilities are the same, we obtain by inserting into (8.14)

$$\begin{aligned}\hat{y} &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\boldsymbol{x} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} \\ &= (0.2418 \quad -0.0652)\begin{pmatrix}131.158 & -90.423 \\ -90.423 & 108.147\end{pmatrix}\begin{pmatrix}x_1 \\ x_2\end{pmatrix} \\ &= 37.61x_1 - 28.92x_2 \ . \end{aligned} \tag{8.15}$$

Further, we obtain

$$\bar{y}_1 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 = (37.61 \quad -28.92)\begin{pmatrix}-0.0065 \\ -0.0390\end{pmatrix} = 0.88$$

$$\bar{y}_2 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2 = (37.61 \quad -28.92)\begin{pmatrix}-0.2483 \\ 0.0262\end{pmatrix} = -10.10$$

and the midpoint between these means through

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(0.88 - 10.10) = -4.61 \ .$$

It is now suspected that a woman is carrier of hemophilia A. Her values are

$$x_1 = -0.210 \quad \textit{and} \quad x_2 = -0.044 \ .$$

In order to determine whether she belongs to group π_1 (normal) or to group π_2 (carrier), we insert her values in (8.15) and use the classification rule (8.13):

$$\textit{Assign } \boldsymbol{x}_0 \textit{ to group } \pi_1, \textit{ if } \ \hat{y}_0 = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 \geq \hat{m} = -4.61$$
$$\textit{Assign } \boldsymbol{x}_0 \textit{ to group } \pi_2, \textit{ if } \ \hat{y}_0 = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 < \hat{m} = -4.61$$

with $\boldsymbol{x}_0 = (-0.210, -0.044)^\top$. *Since*

$$\hat{y}_0 = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 = (37.61 \quad -28.92)\begin{pmatrix}-0.210\\-0.044\end{pmatrix} = -6.62 < -4.61,$$

the woman is assigned to π_2, *i.e. to the group of carriers of hemophilia A. This is shown as a star in Figure 8.2, and it is seen to be within the estimated 50% range of* π_2 *and approximately on the 95% contour line of* π_1. *So the groups are not clearly separated.*

Now assume we know the prior probabilities p_1 *and* p_2 *for the respective group memberships. For example, consider the previously classified woman to be a maternal cousin of a carrier of hemophilia A. Then, the genetic likelihood of this woman being a carrier as well is* 0.25. *Thus,* $p_1 = 0.75$ *(normal) and* $p_2 = 0.25$ *(carrier). Furthermore, if the cost of misclassification is the same (unrealistic), we obtain from inserting into (8.13)*

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = -2.01\ .$$

Since

$$-2.01 < \ln\left(\frac{p_2}{p_1}\right) = \ln\left(\frac{0.25}{0.75}\right) = -1.10,$$

the woman is again assigned to group π_2 *corresponding to the carriers of hemophilia A.*

Remark: The coefficients $\hat{\boldsymbol{a}} = \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ are unique except for a multiplicative constant, because every vector $c\hat{\boldsymbol{a}}$ with $c \neq 0$ can also be used as a coefficient vector for discrimination. Therefore, one will usually make the following standardization

$$\hat{\boldsymbol{a}}^* = \frac{\hat{\boldsymbol{a}}}{\sqrt{\hat{\boldsymbol{a}}^\top \hat{\boldsymbol{a}}}}$$

which makes $\hat{\boldsymbol{a}}^*$ have a length of 1.

8.3.2 The special case $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

The previously formulated classification rules were based on (8.4) and looked at the relation of the density functions $f_1(\boldsymbol{x})/f_2(\boldsymbol{x})$. In the case of the same covariance matrices, this ratio is reduced to a relatively simple term, which is usually expressed by (8.11). In the case of unequal covariance matrices (and unequal means), however, the ratio of density functions becomes a more complicated expression. As with (8.11), the logarithm of the relation can also be considered here:

$$\begin{aligned}\ln\left(\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}\right) &= \ln\left(\frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\right) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\\ &= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}\end{aligned}$$

$$-\frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

This results in the following classification rule:

$$\begin{aligned} R_1: &\ -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - k \geq \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) \\ R_2: &\ -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - k < \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) \end{aligned}$$

with

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) .$$

The constant k now only depends on the mean and covariance of the two distributions. In the case of $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the above rule is reduced to (8.11). The following statement now follows directly:

Theorem 8.3.3 *Let π_1 and π_2 be normally distributed populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Then, the classification rule for minimizing ECM is:*
An observation $\boldsymbol{x}_0$ is assigned to π_1, if

$$-\frac{1}{2}\boldsymbol{x}_0^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0 + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0 - k \geq \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) . \quad (8.16)$$

Otherwise $\boldsymbol{x}_0$ is assigned to π_2.

The above discriminant function is *quadratic* in $\boldsymbol{x}$. Since the population parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are not known, they need to be obtained by the respective estimators, e.g. by the empirical sample estimators $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$

8.3.3 Evaluation of the classification

The above classification rules are based on normally distributed data. If this assumption is not met, the rules that express themselves by linear or quadratic discriminant functions can strongly fail. One could therefore first transform the data so that they correspond better to the (multivariate) normal distribution. In any case, it should be checked how good the discrimination of both populations really was. One instrument for this is the *apparent error rate* (AER),

$$AER = p_1 \int_{\hat{R}_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 \int_{\hat{R}_1} f_2(\boldsymbol{x})d\boldsymbol{x} , \quad (8.17)$$

which is derived directly from the total probability of misclassification (TPM) (8.7). $\hat{R}_1$ and $\hat{R}_2$ are the areas obtained by the sample, if e.g. rule (8.13) is used (assuming the same covariance matrix).

You could also just state the proportion of misclassified objects compared to the total number of objects. Based on a sample of size n_1 of π_1 and n_2 of π_2, this proportion can be determined by applying a classification rule.

However, these procedures always assume that at least one "training data set" knows which objects belong to which group. From this, the discriminant function can be determined and the error rate calculated. New objects can then be classified according to the previously determined rule with a known error rate.

Lachenbruch and Mickey (1968) proposed the following "jackknife" procedure for estimating the error rate:

1. Start with the objects of group π_1. An observation of this group is omitted and a classification function is determined with the remaining $n_1 - 1$ and n_2 observations.

2. The observation omitted in step 1 is classified based on the determined classification function.

3. Repeat steps 1 and 2 until all objects of π_1 are classified. $\bar{n}_1$ is the number of misclassified objects.

4. Repeat steps 1 to 3 for the objects of π_2. The result is the number $\bar{n}_2$ of the wrongly classified objects of group π_2.

5. Now the conditional probabilities for misclassification (8.1) and (8.2) can be estimated by
$$\hat{P}(2|1) = \frac{\bar{n}_1}{n_1} \qquad \hat{P}(1|2) = \frac{\bar{n}_2}{n_2}$$
and from that the estimated error rate
$$\frac{\bar{n}_1 + \bar{n}_2}{n_1 + n_2}$$
can be obtained.

A more useful rule may be based on a k-fold cross-validation scheme, in which the observations are randomly split into k parts (e.g. $k = 5$) of about equal size, and where the discriminant rule in turn is computed for $k - 1$ parts and evaluated on the k-th part. If many observations are available, one could split the data randomly into a training set (e.g. 2/3), compute the rule, and evaluate it on the remaining test data. The proportion of misclassified observations determines the accuracy of the classifier.

Remark: When using the linear discriminant function of Theorem 8.3.1, it should be noted that the variances of the two groups are indeed assumed to be the same. If this is not the case, it can happen that the probabilities for misclassification are very different. This is shown in Figure 8.3. The linear discriminant function (8.14) was used. The dotted curves show the distributions of the two populations around

$\bar{y}_1$ and $\bar{y}_2$ with the pooled variance. Based on the midpoint $\hat{m}$ between $\bar{y}_1$ and $\bar{y}_2$ it is now decided for each observation to which population it is counted. However, if the variance of the populations differs greatly, as indicated by the actual distributions in the figure, there may be a strong imbalance in the actual error probabilities. It is much more likely that values from π_2 are wrongly classified than values from π_1.

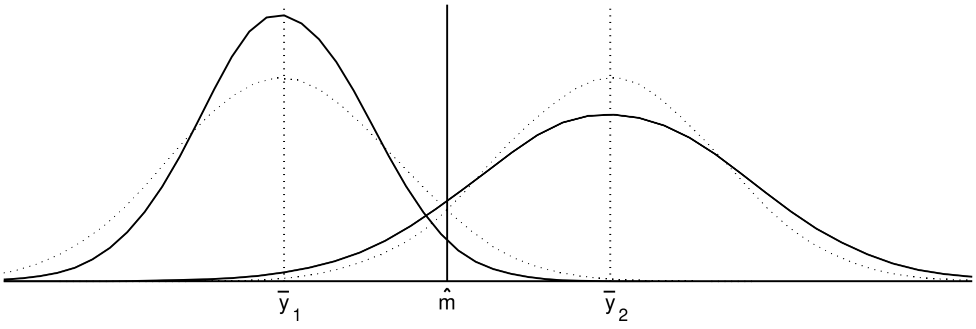


Figure 8.3: Probabilities of misclassification when using the linear discriminant function

8.3.4 Fisher's linear discriminant function

Fisher (1938) developed a linear discriminant function analogously to (8.14), but the idea behind it was different. He tried to transform multivariate observations to univariate, so that the two transformed groups are as strongly separated as possible. The idea was thus to find a direction $\boldsymbol{a} \in I\!R^p$, and to project the observations $\boldsymbol{a}^\top \boldsymbol{x}$ in order to obtain univariate values $y_{11}, y_{12}, \ldots, y_{1n_1}$ for the observations of the first group and values $y_{21}, y_{22}, \ldots, y_{2n_2}$ for the observations of the second group. The separation of the two populations then occurs in such a way that the arithmetic means $\bar{y}_1$ and $\bar{y}_2$ of the univariate y values deviate as much as possible. This difference is expressed in units of the standard deviation and is therefore used as a criterion for the separation

$$\frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \longrightarrow \max$$

with the pooled variance of the y-values

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \ .$$

An attempt is now made to find a linear combination $\boldsymbol{a}$ of $\boldsymbol{x}$, which allows a maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

Theorem 8.3.4 *The linear combination*

$$\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} \tag{8.18}$$

maximizes the ratio

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 - \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled} \hat{\boldsymbol{a}}} \tag{8.19}$$

over all possible vectors $\hat{\boldsymbol{a}}$. $\boldsymbol{S}_{pooled}$ *is defined analogously to the two-group case. The maximum ratio of (8.19) is*

$$D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \ .$$

Proof: Using the extended Cauchy-Schwarz inequality (7.1) with the notation $\boldsymbol{b} = \hat{\boldsymbol{a}}$, $\boldsymbol{d} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, and $\boldsymbol{B} = \boldsymbol{S}_{pooled}$, it follows:

$$\big(\hat{\boldsymbol{a}}^\top (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\big)^2 \leq \hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled} \hat{\boldsymbol{a}} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

Thus,

$$\max_{\hat{\boldsymbol{a}}} \frac{\big(\hat{\boldsymbol{a}}^\top (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\big)^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled} \hat{\boldsymbol{a}}} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = D^2$$

is reached with $\hat{\boldsymbol{a}} = const \cdot \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$. □

The result is thus the following classification rule (cf. (8.14)):

Theorem 8.3.5 *An observation* $\boldsymbol{x}_0$ *is assigned to* π_1*, if*

$$\hat{y}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}_0 \geq \hat{m} = \frac{1}{2} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{8.20}$$

Otherwise, $\boldsymbol{x}_0$ *is assigned to* π_2*.*

This classification rule is shown in Figure 8.4 for $p = 2$. The observations are projected onto a straight line. The direction $\hat{\boldsymbol{a}}$ of the line is varied until the two groups are maximally separated.

In contrast to rule (8.13), Fisher's classification rule does not explicitly require the assumption of normal distribution of both populations. However, in case of violations from this assumption, we know from the previous results that the rule would not be optimal in terms of minimizing TPM (or ECM). Moreover, we also have to assume that the populations have the same covariance matrix, since a pooled estimate of the covariance is used. It can be seen immediately that Fisher's linear discriminant function in (8.20) is a special case of (8.13). If the prior probabilities and the expected costs for misclassification are the same for rule (8.13), which results from minimizing the expected costs of misclassification, we get exactly (8.20).

Example 8.3.2 *We consider the hemophilia data again. Assuming equal prior probabilities and equal cost of misclassification, we obtained in (8.15)*

$$\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} = 37.61 x_1 - 28.92 x_2 \ .$$

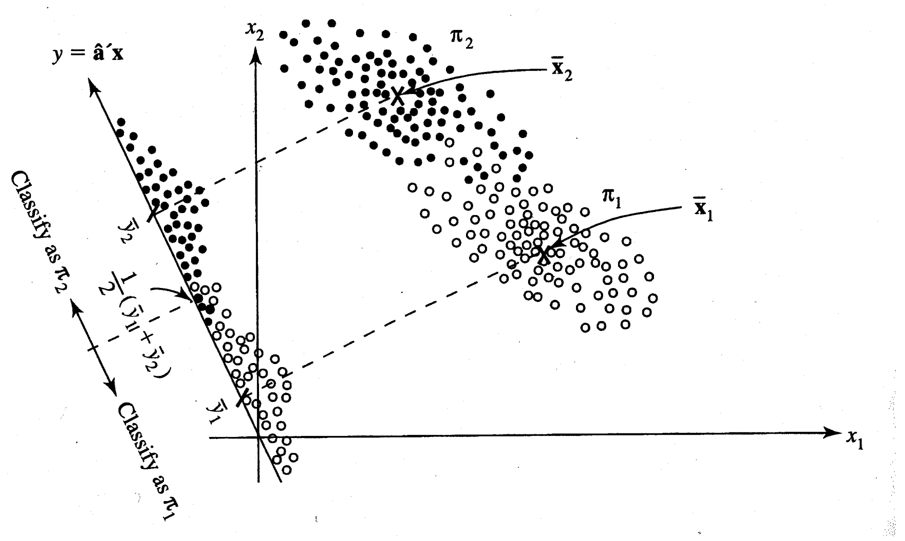


Figure 8.4: Schematic representation of Fisher's classification rule.

This function corresponds to the linear discriminant function of Fisher in (8.18) and allows for maximum separation of the two groups. The maximum separation is given by

$$\begin{aligned} D^2 &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\ &= (0.2418 \quad -0.0652)\begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix}\begin{pmatrix} 0.2418 \\ -0.0652 \end{pmatrix} \\ &= 10.98 \ . \end{aligned}$$

This means that the value $D^2 = 10.98$ indicates the maximum separation of the two groups, which is made possible by linear combination of the multivariate observations.

8.4 Classification of multiple populations

In this section we will now expand the concepts for the case of two populations to $g > 2$ groups. This extension is mathematically plausible, but so far little is known about the properties of the corresponding classification rules of the sample. Deviations from the normal distribution of the groups or different covariances could strongly distort the good theoretical properties.

8.4.1 The method to minimize ECM

Let $f_i(\boldsymbol{x})$ be the density of the observations of population π_i with $i = 1, \ldots, g$. Mostly, it is assumed that $f_i(\boldsymbol{x})$ is the density of a multivariate normal distribution, but this is not a requirement of the following method.

The notation is based on the two-group case. Thus, p_i denotes the prior probability for the population π_i $(i = 1 \ldots, g)$. Further, let R_k be the space of observations

$\boldsymbol{x}$ to which the objects from π_k are assigned. $c(k|i)$ is the cost of misclassification when objects from π_i are mistakenly mapped to π_k $(k = 1, \ldots, g)$, where $k = i$ is of course $c(i|i) = 0$. The probability of this wrong assignment is

$$P(k|i) = P(\boldsymbol{X} \in R_k|\pi_i) = \int_{R_k} f_i(\boldsymbol{x})d\boldsymbol{x} \; . \tag{8.21}$$

The conditional expected cost of misclassifying $\boldsymbol{x}$ from π_1 to $\pi_2, \ldots, \pi_g$ are analogous to (8.3)

$$ECM(1) = \sum_{k=2}^{g} P(k|1)c(k|1) \; . \tag{8.22}$$

These expected costs arise with prior probability p_1. $ECM(2), \ldots, ECM(g)$ are analogously defined, and one obtains by multiplying with the prior probabilities and adding up all contributions, the total expected cost of misclassification (ECM) as

$$ECM = \sum_{i=1}^{g} p_i ECM(i) = \sum_{i=1}^{g} p_i \left(\sum_{\substack{k=1 \\ k \neq i}}^{g} P(k|i)c(k|i) \right) \; . \tag{8.23}$$

An optimal classification rule should now yield such ranges $R_1, \ldots, R_g$ (disjoint and complete decomposition of the sample space Ω) so that (8.23) is minimized.

Theorem 8.4.1 *The areas for minimizing ECM (8.23) are given by assigning $\boldsymbol{x}$ to the population π_k $(k = 1, \ldots, g)$ for which the expression*

$$\sum_{\substack{i=1 \\ k \neq i}}^{g} p_i f_i(\boldsymbol{x}) c(k|i) \tag{8.24}$$

is minimal.

Proof: see Anderson (1984); as well as (8.4)

For simplicity, we assume in the remainder of this chapter that the costs of misclassification are the same for all groups, and thus we can ignore them (or set them equal to 1). Then, minimizing criterion (8.24) corresponds to maximizing the omitted term $p_k f_k(\boldsymbol{x})$, which simplifies the classification rule to:

An observation $\boldsymbol{x}$ is assigned to π_k, if

$$p_k f_k(\boldsymbol{x}) > p_i f_i(\boldsymbol{x}) \quad \text{for all } i \neq k \; . \tag{8.25}$$

Note that the above classification rules can only be applied if the prior probabilities, the misclassification costs, and the density functions are known.

8.4.2 Classification in case of the normal distribution

As an important special case, we consider multivariate normally distributed populations π_i with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$,

$$f_i(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \quad \text{for } i=1,\ldots,g \ . \tag{8.26}$$

If we now use rule (8.25) or the logarithm of this rule, we get:
An observation $\boldsymbol{x}$ is assigned to π_k, if

$$\ln p_k f_k(\boldsymbol{x}) = \ln p_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) = \max_i \ln p_i f_i(\boldsymbol{x}) \ . \tag{8.27}$$

The constant $(p/2)\ln(2\pi)$ can be omitted in (8.27) since it is the same for all π_i. We therefore define the *quadratic discriminant values* for the i-th population as

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln p_i \quad \text{for } i=1,\ldots,g \ . \tag{8.28}$$

Thus one obtains the following classification rule:

An observation $\boldsymbol{x}$ is assigned to π_k, if:

$$d_k^Q(\boldsymbol{x}) \quad \text{is the largest of} \quad d_1^Q(\boldsymbol{x}),\ldots,d_g^Q(\boldsymbol{x}) \ . \tag{8.29}$$

Mostly, the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are unknown and must be estimated. If a “training set” with known class memberships of the observations is available, then the sample means $\bar{\boldsymbol{x}}_i$ and sample covariance matrices $\boldsymbol{S}_i$ can be used to estimate the corresponding parameters of the i-th population ($i=1,\ldots,g$). This results in the estimated quadratic discriminant values

$$\hat{d}_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i) + \ln p_i \ , \tag{8.30}$$

with the help of which objects are classified analogous to (8.29).

A simplification results if the covariance matrices of the populations are the same, that is $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ for $i=1,\ldots,g$. In this case, the quadratic discriminant values from (8.28) are

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \ . \tag{8.31}$$

Since the first two terms in (8.31) are the same for all populations, they can be omitted, and we get the *linear discriminant values*

$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad \text{for} \quad i=1,\ldots,g \ . \tag{8.32}$$

One obtains an estimate of the linear discriminant values from the sample by first estimating a pooled covariance matrix

$$\begin{aligned}\boldsymbol{S}_{pooled} &= \frac{1}{n_1+\cdots+n_g-g}\Big((n_1-1)\boldsymbol{S}_1+\cdots+(n_g-1)\boldsymbol{S}_g\Big)\\ &= \frac{1}{\sum_{i=1}^g n_i - g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)^\top \ .\end{aligned}$$

This gives the following estimates

$$\hat{d}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{pooled}^{-1}\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{pooled}^{-1}\bar{\boldsymbol{x}}_i + \ln p_i \ . \tag{8.33}$$

The resulting classification rule is:

An observation $\boldsymbol{x}$ is assigned to π_k, if:

$$\hat{d}_k(\boldsymbol{x}) \quad \text{is the largest of} \quad \hat{d}_1(\boldsymbol{x}), \ldots, \hat{d}_g(\boldsymbol{x}) \ . \tag{8.34}$$

Remark: The expression (8.32) is a linear function of $\boldsymbol{x}$. However, one could obtain an analogous rule by ignoring the first term in (8.28), which is the same for all populations in the case of equal covariance matrices. With the corresponding estimated values, one obtains a squared distance

$$D_i^2(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_{pooled}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i) \ , \tag{8.35}$$

and the classification rule is:

Assign $\boldsymbol{x}$ to population π_i for which $-1/2 \ D_i^2(\boldsymbol{x}) + \ln p_i$ is the largest. (8.36)

This rule is analogous to rule (8.34), it assigns $\boldsymbol{x}$ to the population that is "closest", with the distance measure penalized with $\ln p_i$. If the a-priori probabilities are not known, one could estimate them by $p_i = 1/g$.

Example 8.4.1 *We consider the data of Ruspini (1970) from the R package `cluster`. These are 2-dimensional data with 75 observations forming 4 groups, see Figure 8.5. Here, the division into the groups took place with k -means clustering, which also corresponds to a visual classification. The numbers of objects in each group are*

$$n_1 = 17 \qquad n_2 = 20 \qquad n_3 = 23 \qquad n_4 = 15 \ .$$

The group means are marked by + in Figure 8.5 and amount to

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 98.18 \\ 114.88 \end{pmatrix} \qquad \bar{\boldsymbol{x}}_2 = \begin{pmatrix} 20.15 \\ 64.95 \end{pmatrix} \qquad \bar{\boldsymbol{x}}_3 = \begin{pmatrix} 43.91 \\ 146.04 \end{pmatrix} \qquad \bar{\boldsymbol{x}}_4 = \begin{pmatrix} 68.93 \\ 19.40 \end{pmatrix} \ .$$

The covariance matrices of each group are

$$\boldsymbol{S}_1 = \begin{pmatrix} 164.53 & 64.27 \\ 64.27 & 120.36 \end{pmatrix} \qquad \boldsymbol{S}_2 = \begin{pmatrix} 92.661 & -5.045 \\ -5.045 & 101.524 \end{pmatrix}$$

$$\boldsymbol{S}_3 = \begin{pmatrix} 91.81 & -23.27 \\ -23.27 & 52.59 \end{pmatrix} \qquad \boldsymbol{S}_4 = \begin{pmatrix} 51.2095 & 0.2429 \\ 0.2429 & 52.8286 \end{pmatrix}$$

This results in the pooled covariance matrix and its inverse as

$$\boldsymbol{S}_{pooled} = \begin{pmatrix} 100.419 & 5.972 \\ 5.972 & 81.004 \end{pmatrix} \qquad \boldsymbol{S}_{pooled}^{-1} = \begin{pmatrix} 0.0100021 & -0.0007374 \\ -0.0007374 & 0.0123995 \end{pmatrix}$$

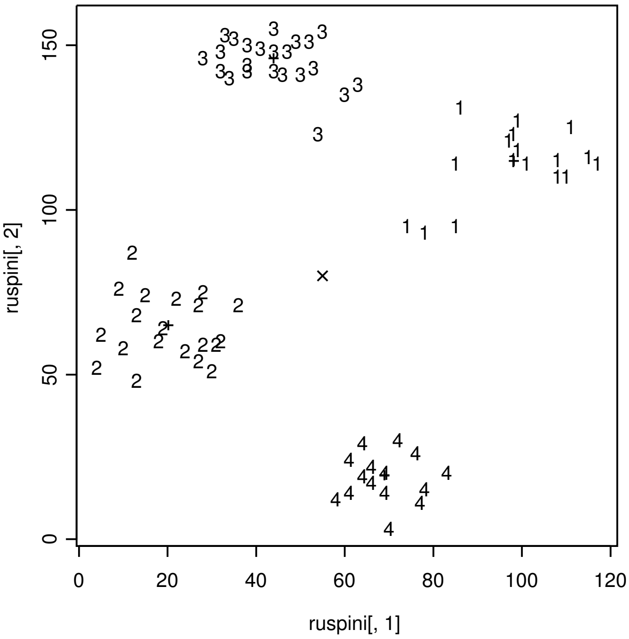


Figure 8.5: Data `ruspini` from the R package `cluster`.

We now want to classify a new observation $\boldsymbol{x}_0 = (55, 80)^\top$, which is plotted in Figure 8.5 as ×. We assume that the prior probabilities are the same, so $p_i = 0.25$. We estimate the linear discriminant values by (8.33) and get

$$\hat{d}_1 = 34.42 \qquad \hat{d}_2 = 43.08 \qquad \hat{d}_3 = 21.98 \qquad \hat{d}_4 = 25.81 \ .$$

Using the classification rule (8.34) that assigns the new observation to the nearest population, $\boldsymbol{x}_0$ is assigned to population π_2, because 43.08 is the largest discriminant value.

If one does not want to assume that the covariance matrices are the same, one uses the (estimated) quadratic discriminant values from (8.30) and obtains

$$\hat{d}_1^Q = -13.59 \qquad \hat{d}_2^Q = -13.93 \qquad \hat{d}_3^Q = -49.06 \qquad \hat{d}_4^Q = -42.07 \ .$$

The largest value is -13.59, and thus $\boldsymbol{x}_0$ is assigned to population π_1. However, the decision is extremely close.

Similar to the two-group case, the ”jackknife” procedure of Lachenbruch and Mickey (1968) can be used to estimate the error rate. If $\bar{n}_i$ denotes the number of misclassified objects of the i-th group ($i = 1, \ldots, g$), then

$$\frac{\sum_{i=1}^{g} \bar{n}_i}{\sum_{i=1}^{g} n_i}$$

is the estimated error rate. As mentioned above, a cross-validation procedure, or a training/test scenario is more advisable.

Remark: The method of Fisher can be also be extended to the case of multiple groups. The discriminant function is a ratio of variation between the groups to the variation within the groups (similar to analysis of variance). Details can be found e.g. in Johnson and Wichern (1998).

8.4.3 Discriminant analysis by Fisher for the multi-group case

Fisher's discriminant analysis for $g = 2$ can be extended to the multi-group case ($g > 2$), see Rao (1948). For this purpose, we again consider univariate projections of the form $y = \mathbf{a}^T\mathbf{x}$ with $\mathbf{a} \in I\!R^p$ and $\mathbf{a} \neq \mathbf{0}$, but this time a single projection direction $\mathbf{a}$ will not be enough to describe the solution.

Let $\mathbf{a} \neq \mathbf{0}$ be the projection direction we are looking for. The expected value for population π_i (for $i = 1, \ldots, g$) of the random variable $y = \mathbf{a}^T\mathbf{x}$ is then

$$\mu_{i,y} = E(y|\mathbf{x} \in \pi_i) = \mathbf{a}^T E(\mathbf{x}|\mathbf{x} \in \pi_i) = \mathbf{a}^T \boldsymbol{\mu}_i$$

and the variance is

$$\sigma^2_{i,y} = \mathrm{Var}(y|\mathbf{x} \in \pi_i) = \mathbf{a}^T \mathrm{Cov}(\mathbf{x}|\mathbf{x} \in \pi_i)\mathbf{a} = \mathbf{a}^T \boldsymbol{\Sigma}_i \mathbf{a} \ .$$

The total weighted average of the populations is denoted by $\bar{\boldsymbol{\mu}} = \sum_{i=1}^g p_i \boldsymbol{\mu}_i$, and the corresponding projection into the univariate space by $\bar{\mu}_y = \mathbf{a}^T \bar{\boldsymbol{\mu}}$.

We now make the assumption that the covariances of all groups are the same, that is $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$. Then it also applies to the above variance that

$$\sigma^2_{1,y} = \ldots = \sigma^2_{g,y} = \sigma^2_y = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \ .$$

In Fisher's two-group case, an expression $(\bar{y}_1 - \bar{y}_2)^2/s_y^2$ was maximized. In our notation with random variables, this corresponds to $(\mu_{1,y} - \mu_{2,y})^2/\sigma_y^2$. Additionally, we now consider prior probabilities, and the weighted mean is

$$\bar{\mu}_y = p_1 \mu_{1,y} + p_2 \mu_{2,y} = \mathbf{a}^T (p_1 \boldsymbol{\mu}_1 + p_2 \boldsymbol{\mu}_2) \ .$$

Since $p_1 + p_2 = 1$, it is straightforward to see that

$$p_1(\mu_{1,y} - \bar{\mu}_y)^2 + p_2(\mu_{2,y} - \bar{\mu}_y)^2 = p_1 p_2 (\mu_{1,y} - \mu_{2,y})^2 \ . \tag{8.37}$$

The latter expression (8.37) should therefore be maximized using the Fisher rule (in terms of variance), and this expression describes the weighted sum of the squared distances of the group means to the total mean.

The generalization of (8.37) to the multi-group case is then immediately obvious; now

$$\frac{\sum_{i=1}^g p_i (\mu_{i,y} - \bar{\mu}_y)^2}{\sigma_y^2} \tag{8.38}$$

should be maximized. The denominator is according to above $\sigma_y^2 = \mathbf{a}^T \mathbf{\Sigma} \mathbf{a}$. If the group covariances are not equal, the resulting classification rule will no longer be optimal. $\mathbf{\Sigma}$ would then be best replaced by a pooled version, that is

$$\mathbf{W} = \sum_{i=1}^{g} p_i \mathbf{\Sigma}_i \ .$$

The matrix $\mathbf{W}$ describes the *variation within groups.*

The numerator of (8.38) can be represented as

$$\sum_{i=1}^{g} p_i (\mu_{i,y} - \bar{\mu}_y)^2 = \sum_{i=1}^{g} p_i (\mathbf{a}^T (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}))^2 = \mathbf{a}^T \mathbf{B} \mathbf{a} \ ,$$

with the matrix

$$\mathbf{B} = \sum_{i=1}^{g} p_i (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^T \ ,$$

which describes the *variation between groups.*

All in all, the maximization problem (8.38) can be expressed as

$$\frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}} \qquad \text{for } \mathbf{a} \in I\!R^p, \ \mathbf{a} \neq \mathbf{0} \ . \tag{8.39}$$

Theorem 8.4.2 *The solution of the maximization problem (8.39) is given by the eigenvectors* $\mathbf{a}_1, \ldots, \mathbf{a}_l$ *of the matrix* $\mathbf{W}^{-1}\mathbf{B}$*, which must be scaled, so that* $\mathbf{a}_j^T \mathbf{W} \mathbf{a}_j = 1$ *for* $i = j, \ldots, l$*. The number* l *of the strictly positive eigenvalues of the eigenvalue decomposition of* $\mathbf{W}^{-1}\mathbf{B}$ *is then* $l \leq \min(g-1, p)$*.*

Proof: Problem (8.39) is invariant with respect to the rescaling of $\mathbf{a}$. Meaning, for every $\tilde{\mathbf{a}} = \alpha \mathbf{a}$, with $\alpha \neq 0$, you get the same maximum. Therefore, one can scale the denominator to $\mathbf{a}^T \mathbf{W} \mathbf{a} = 1$. This simplifies the optimization problem (8.39) to

$$\min_{\mathbf{a}} (-\mathbf{a}^T \mathbf{B} \mathbf{a}) \qquad \text{so that} \quad \mathbf{a}^T \mathbf{W} \mathbf{a} = 1 \ .$$

The minimization takes place by means of Langrange's expression

$$\phi = -\frac{1}{2} \mathbf{a}^T \mathbf{B} \mathbf{a} + \frac{1}{2} \lambda (\mathbf{a}^T \mathbf{W} \mathbf{a} - 1)$$

with the Lagrange multiplier λ. The terms 1/2 are convenient if we now form the derivative:

$$\frac{\partial \phi}{\partial \mathbf{a}} = -\mathbf{B}\mathbf{a} + \lambda \mathbf{W} \mathbf{a}$$

Zeroing the derivative yields

$$\mathbf{B}\mathbf{a} = \mathbf{W} \lambda \mathbf{a} \qquad \text{or} \qquad \mathbf{W}^{-1} \mathbf{B} \mathbf{a} = \lambda \mathbf{a} \ .$$

So, again we get an eigenvalue problem and the solutions for $\mathbf{a}$ are the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_l$ of $\mathbf{W}^{-1}\mathbf{B}$ to the eigenvalues $\lambda_1, \ldots, \lambda_l$.

The eigenvalues are sorted in descending order, and then $\mathbf{a}_1$ represents the direction along which the group means are best separated. Thus, the order of the directions obtained is important, and with a projection in the first direction, as in principal component analysis, one obtains the most informative dimension reduction of the entire information, with the difference that in addition to the multivariate information of the observations, one also has information about the class.

The number $l \leq \min(g-1, p)$ can be explained directly from the ranks of $\mathbf{W}$, namely p at most, and $\mathbf{B}$ (maximum $g-1$).

Note that $\mathbf{W}^{-1}\mathbf{B}$ is not necessarily symmetric, and thus eigenvalues and -vectors may have imaginary parts. This can easily be circumvented by representing the symmetric matrix $\mathbf{B}$ as $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$.

With definition $\mathbf{b} = \mathbf{B}^{1/2}\mathbf{a}$, or equivalent $\mathbf{a} = \mathbf{B}^{-1/2}\mathbf{b}$, the above eigenvalue problem is $\mathbf{B}^{1/2}\mathbf{W}^{-1}\mathbf{B}^{1/2}\mathbf{b} = \lambda\mathbf{b}$, so we have an eigenvalue problem of the symmetric matrix $\mathbf{B}^{1/2}\mathbf{W}^{-1}\mathbf{B}^{1/2}$. □

We can now define the *Fisher discriminant functions* as $y_j = \mathbf{a}_j^T\mathbf{x}$, for $j = 1, \ldots, l$, which is the projection of the random variable $\mathbf{x}$ in the direction $\mathbf{a}_j$. If specific data is available, visualizing the first two discriminant functions is particularly interesting because it represents the projection of the data in which the group means appear best separated. Note that in the case of $g = 3$ groups it holds that $l \leq 2$, regardless of whether p is large or not.

Finally, we also want to get a classification rule. For that, consider the *Fisher discriminant values*

$$d_i^F(\mathbf{x}) = \sum_{j=1}^{l}(y_j - \mu_{i,y_j})^2 - 2\log p_i \tag{8.40}$$

for $i = 1, \ldots, g$. Here $\mu_{i,y_j} = \mathbf{a}_j^T\boldsymbol{\mu}_i$, and one thus has a measure of the deviation from $\mathbf{x}$ to the i-th group mean in the discriminant space, adjusted with the prior probability (analogous to earlier). Note that here in the discriminant space the distance measure is simply the Euclidean distance. A new observation $\mathbf{x}$ is then assigned to population π_k, if $d_k^F(\mathbf{x})$ is the smallest (!) value of all the values of the groups $d_1^F(\mathbf{x}), \ldots, d_g^F(\mathbf{x})$.

By arranging the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_l$ in the columns of the matrix $\mathbf{A}$, we can also write the *Fisher discriminant values* as

$$d_i^F(\mathbf{x}) = \sum_{j=1}^{l}(\mathbf{a}_j^T(\mathbf{x} - \boldsymbol{\mu}_i))^2 - 2\log p_i = (\mathbf{x} - \boldsymbol{\mu}_i)^T\mathbf{A}\mathbf{A}^T(\mathbf{x} - \boldsymbol{\mu}_i) - 2\log p_i \ ,$$

which corresponds to a (squared) Mahalanobis distance in the original space.

References

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, 1984.

B.N. Bouma, et al. Evaluation of the detection rate of hemophilia carriers. *Statistical Methods for Clinical Decision Making*, 7(2):339–350, 1975.

R.A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376-386, 1938.

D.J. Hand. *Discrimination and Classification.* John Wiley & Sons, New York, 1981.

C.J. Huberty. *Applied Discriminant Analysis.* John Wiley & Sons, New York, 1994.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

P.A. Lachenbruch and M.R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968.

C.R. Rao. The utilization to multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

E.H. Ruspini. Numerical methods for fuzzy clustering. *Inform. Sci.*, 2:319-350, 1970.

G.A.F. Seber. *Multivariate observations.* John Wiley & Sons, New York, 1984.