9

Principal Component Analysis

> I tried one or two explanations, but, indeed, I was completely puzzled myself. Our friend's title, his fortune, his age, his character, and his appearance are all in his favour, and I know nothing against him, unless it be the dark fate which runs in his family.
> "The Hound of the Baskervilles"

This chapter addresses the issue of reducing the dimensionality of a multivariate random variable by using linear combinations (the principal components). The identified principal components are ordered in decreasing order of importance. When applied in practice to a data matrix, the principal components will turn out to be the factors of a transformed data matrix (the data will be centered and eventually standardized).

For a random vector X with $E(X) = \mu$ and $\mathrm{Var}(X) = \Sigma = \Gamma\Lambda\Gamma^\top$, the principal component (PC) transformation is defined as

$$Y = \Gamma^\top(X - \mu). \tag{9.1}$$

It will be demonstrated in Exercise 9.1 that the components of the random vector Y have zero correlation. Furthermore, it can be shown that they are also standardized linear combinations with the largest variance and that the sum of their variances, $\sum \mathit{Var}\, Y_i$, is equal to the sum of the variances of $X_1, \dots, X_p$.

In practice, the PC transformation is calculated using the estimators $\overline{x}$ and $\mathcal{S}$ instead of μ and Σ. If $\mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^\top$ is the spectral decomposition of the empirical covariance matrix $\mathcal{S}$, the principal components are obtained by

$$\mathcal{Y} = (\mathcal{X} - 1_n\overline{x}^\top)\mathcal{G}. \tag{9.2}$$

Theorem 9.1 describes the relationship between the eigenvalues of Σ and the eigenvalues of the empirical variance matrix $\mathcal{S}$.

THEOREM 9.1. *Let $\Sigma > 0$ with distinct eigenvalues and let $\mathcal{U} \sim m^{-1} W_p(\Sigma, m)$ with spectral decompositions $\Sigma = \Gamma \Lambda \Gamma^\top$ and $\mathcal{U} = \mathcal{G} \mathcal{L} \mathcal{G}^\top$. Then*

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

where $\ell = (\ell_1, \dots, \ell_p)^\top$ and $\lambda = (\lambda_1, \dots, \lambda_p)^\top$ are the diagonals of $\mathcal{L}$ and Λ.

The proof and the asymptotic distribution of $\mathcal{G}$ can be found, e.g., in Härdle & Simar (2003, theorem 9.4).

The resulting PCA (principal component analysis) or NPCA (normalized PCA) is presented in a variety of examples, including U.S. crime and health data. A PCA is also performed for an OECD data set on variables of political nature (life expectance, literacy, etc.).

EXERCISE 9.1. *Calculate the expected value and the variance of the PC transformation Y defined in (9.1). Interpret the results.*

For the expected value, EY, we have

$$EY = E\Gamma^\top (X - \mu) = \Gamma^\top E(X - \mu) = \Gamma^\top (EX - \mu) = 0_p.$$

The variance matrix, $\operatorname{Var}(Y)$, can be calculated as

$$\operatorname{Var}(Y) = \operatorname{Var}\{\Gamma^\top (X - \mu)\} = \Gamma^\top \Sigma \Gamma = \Gamma^\top \Gamma \Lambda \Gamma^\top \Gamma = \Lambda.$$

Hence, the random vector Y is centered (its expected value is equal to zero) and its variance matrix is diagonal.

The eigenvalues $\lambda_1, \dots, \lambda_p$ are variances of the principal components $Y_1, \dots, Y_p$. Notice that

$$\sum_{i=1}^{p} Var(X_i) = \operatorname{tr} \Sigma = \operatorname{tr}\{\Gamma \Lambda \Gamma^\top\} = \operatorname{tr}\{\Gamma^\top \Gamma \Lambda\} = \operatorname{tr} \Lambda = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i).$$

Hence, the variances of X_i are decomposed into the variances of Y_i which are given by the eigenvalues of Σ. The sum of variances of the first q principal components, $\sum_{i=1}^{q} \lambda_i$, thus measures the variation of the random vector X explained by $Y_1, \dots, Y_q$. The proportion of the explained variance,

$$\psi_q = \frac{\lambda_1 + \dots + \lambda_q}{\lambda_1 + \dots + \lambda_p},$$

will be important for the interpretation of results of the practical analyses presented in the following exercises.

EXERCISE 9.2. *Calculate the correlation between X and its PC transformation Y.*

The covariance between the PC vector Y and the original vector X is:

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

The correlation, $\rho_{X_i Y_j}$, between variable X_i and the PC Y_j is

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

The correlations describe the relations between the PCs and the original variables. Note that $\sum_{j=1}^{p} \lambda_j \gamma_{ij}^2 = \gamma_i^\top \Lambda \gamma_i$ is the (i, i)-element of the matrix $\Gamma\Lambda\Gamma^\top = \Sigma$, so that

$$\sum_{j=1}^{p} \rho_{X_i Y_j}^2 = \frac{\sum_{j=1}^{p} \lambda_j \gamma_{ij}^2}{\sigma_{X_i X_i}} = \frac{\sigma_{X_i X_i}}{\sigma_{X_i X_i}} = 1.$$

Hence, the correlation $\rho_{X_i Y_j}^2$ may be seen as the proportion of variance of the ith variable X_i explained by the jth principal component Y_j.

Notice that the percentage of variance of X_i explained by the first q PCs $Y_1, \dots, Y_q$ is $\sum_{j=1}^{q} \rho_{X_i Y_j}^2 < 1$. The distance of the point with coordinates $(\rho_{X_i Y_1}, \dots, \rho_{X_i Y_q})$ from the surface of the unit ball in q-dimensional space can be used as a measure of the explained variance of X_i.

EXERCISE 9.3. *Apply the PCA to the car marks data in Table A.5. Interpret the first two PCs. Would it be necessary to look at the third PC?*

The eigenvalues of the covariance matrix,

$$\lambda = (5.56, 1.15, 0.37, 0.10, 0.08, 0.05, 0.04, 0.02)^\top,$$

lead to the following proportions of the explained variance:

$$\psi = (0.76, 0.91, 0.96, 0.98, 0.99, 0.99, 1.00, 1.00)^\top.$$

Observing that the first two principal components explain more than 90% of the variability of the data set, it does not seem necessary to include also the third PC which explains only 5% of the variability. A graphical display of the eigenvalues, the screeplot, is plotted in the lower right part in Figure 9.1.

The first two eigenvectors of the covariance matrix are

$$\gamma_1 = (-0.22, 0.31, 0.44, -0.48, 0.33, 0.39, 0.42, -0.01)^\top$$

and

$$\gamma_2 = (0.54, 0.28, 0.22, 0.30, -0.14, -0.16, 0.46, 0.49)^\top.$$

Hence, the first two principal components are defined as:

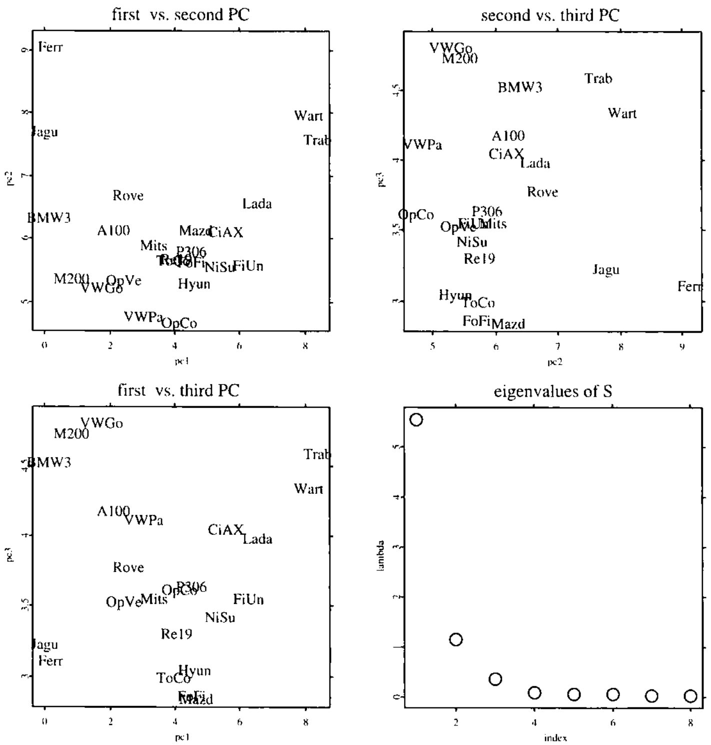

Fig. 9.1. Scatterplots of the first three principal components and a screeplot of the eigenvalues, car marks data set. SMSpcacarm

$$Y_1 = -0.22 \times \text{econ} + 0.31 \times \text{serv} + 0.44 \times \text{value} - 0.48 \times \text{price} + 0.33 \times \text{desi}$$
$$+ 0.39 \times \text{sport} + 0.42 \times \text{safe} - 0.01 \times \text{easy},$$
$$Y_2 = 0.54 \times \text{econ} + 0.28 \times \text{serv} + 0.22 \times \text{value} + 0.30 \times \text{price} - 0.14 \times \text{desi}$$
$$- 0.16 \times \text{sport} + 0.46 \times \text{safe} + 0.49 \times \text{easy}.$$

Using the coefficients of the PCs for interpretation might be misleading especially when the variables are observed on different scales. It is advisable to base the interpretations on the correlations of PCs with the original variables which are plotted in Figure 9.2.

For the car marks data set both the coefficients of the PCs and their correlations with the original variables in Figure 9.2 suggest that the first principal components distinguishes the expensive and design cars from the cheap and less sporty vehicles. This interpretation is confirmed by the plot of the first principal component, Y_1, on Figure 9.1. On the right hand side, we observe the not so cool brands such as Wartburg, Trabant, Lada or Fiat, whereas on the left hand side, we see Jaguar, Ferrari, BMW, and Mercedes-Benz.

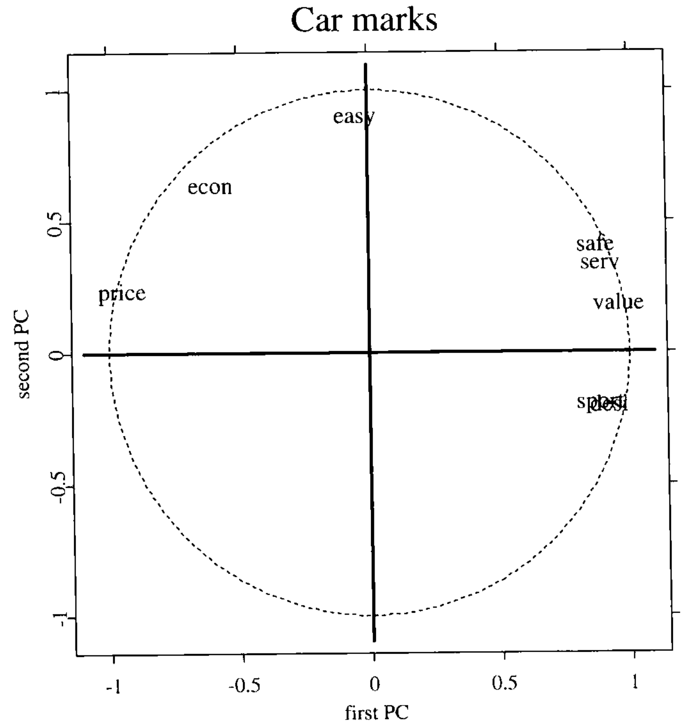

Fig. 9.2. Correlations of the first two principal components with the original variables in the car marks data set. SMSpcacarm

The second PC distinguishes economic cars that are easy to handle, such as Volkswagen and Opel, from the cars that consume a lot of gas and their handling is more problematic such as Ferrari, Wartburg, Jaguar, and Trabant.

Figure 9.2 shows that all of the original variables are very well explained by the first two PCs since all points can be found very close to the unit circle, see the explanation in Exercise 9.2.

EXERCISE 9.4. *Test the hypothesis that the proportion of variance explained by the first two PCs in Exercise 9.3 is $\psi = 0.85$.*

The variance explained by the first q PCs, $\psi_q = (\lambda_1 + \cdots + \lambda_q)/\sum_{j=1}^p \lambda_j$, is in practice estimated by $\widehat{\psi}_q = (\ell_1 + \cdots + \ell_q)/\sum_{j=1}^p \ell_j$. From Theorem 9.1 we know the distribution of $\sqrt{n-1}(\ell - \lambda)$ and, since $\widehat{\psi}_q$ is a function of asymptotically normally distributed random vector ℓ, we obtain that

$$\sqrt{n-1}(\widehat{\psi}_q - \psi_q) \xrightarrow{\mathcal{L}} N(0, \mathcal{D}^\top \mathcal{V} \mathcal{D})$$

where $\mathcal{V} = 2\Lambda^2$ from Theorem 9.1 and $\mathcal{D} = (d_1, \ldots, d_p)^\top$ with

$$d_j = \frac{\partial \psi_q}{\partial \lambda_j} = \begin{cases} \dfrac{1-\psi_q}{\operatorname{tr}(\Sigma)} & \text{if } 1 \le j \le q, \\ \dfrac{-\psi_q}{\operatorname{tr}(\Sigma)} & \text{if } q+1 \le j \le p. \end{cases}.$$

It follows that

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

where

$$\begin{aligned} \omega^2 &= \mathcal{D}^\top \mathcal{V} \mathcal{D} \\ &= \frac{2}{\{\operatorname{tr}(\Sigma)\}^2} \{(1-\psi)^2(\lambda_1^2 + \cdots + \lambda_q^2) + \psi^2(\lambda_{q+1}^2 + \cdots + \lambda_p^2)\} \\ &= \frac{2\operatorname{tr}(\Sigma^2)}{\{\operatorname{tr}(\Sigma)\}^2}(\psi^2 - 2\beta\psi_q + \beta) \end{aligned}$$

and

$$\beta = \frac{\lambda_1^2 + \cdots + \lambda_q^2}{\lambda_1^2 + \cdots + \lambda_p^2} = \frac{\lambda_1^2 + \cdots + \lambda_q^2}{\operatorname{tr}(\Sigma^2)}.$$

In practice, we work with an estimate $\widehat{\omega}^2$ based on the spectral decomposition of the empirical covariance matrix.

In Exercise 9.3 we have calculated the eigenvalues:

$$\lambda = (5.56, 1.15, 0.37, 0.10, 0.08, 0.05, 0.04, 0.02)^\top$$

and the proportions of the explained variance:

$$\psi = (0.76, 0.91, 0.96, 0.98, 0.99, 0.99, 1.00, 1.00)^\top.$$

It follows that, for $q = 2$, we obtain $\widehat{\beta} = 0.99524$ and $\widehat{\omega}^2 = 0.0140$. Under the null hypothesis, $H_0 : \psi_2 = 0.85$, the test statistic $\sqrt{n-1}(\widehat{\psi}_2 - 0.85)/\omega$ has asymptotically standard normal distribution. In our case the value of the test statistic, 2.4401, is in absolute value larger than the critical value of the normal distribution $\Phi^{-1}(0.975) = 1.96$ and we reject the null hypothesis.

Hence, on confidence level $\alpha = 0.95$, we have proved that the proportion of variance explained by the first two principal components is larger than 85%.

EXERCISE 9.5. *Take the athletic records for 55 countries given in Table A.1 and apply the NPCA. Interpret your results.*

The athletic records data set contains national records in 8 disciplines (100m, 200m, 400m, 800m, 1500m, 5km, 10km, and marathon) for $n = 55$ countries. Clearly, the times and hence also the differences between countries will be much larger for longer tracks. Hence, before running the PC analysis, the dataset is normalized by dividing each variable by its estimated standard deviation. The resulting analysis will be called Normalized PCA (NPCA).

In principle, the same results can be obtained by calculating the spectral decomposition of the empirical correlation matrix of the original data set. One only has to be very careful and keep in mind that the derived coefficients of the PCs apply to the normalized variables. Combining these coefficients with the original variables would lead to misleading results.

The eigenvalues and the proportions of explained variance are

$$\lambda = (6.04, 0.99, 0.60, 0.13, 0.10, 0.07, 0.05, 0.02)^\top$$

and

$$\psi = (0.75, 0.88, 0.95, 0.97, 0.98, 0.99, 1.00, 1.00)^\top.$$

Notice that the sum of all eigenvalues is equal to 8. This follows from the fact that the variances of the standardized variables are equal to 1 and from the relationship $\sum_{i=1}^p \lambda_i = \operatorname{tr} \mathcal{S} = \sum_{i=1}^p 1 = p = 8$.

Considering the above eigenvalues and proportions of explained variance, it would be reasonable to investigate only 1 principal component, see also the screeplot in Figure 9.3. A commonly accepted rule says that it suffices to keep only PCs that explain larger than the average number of the total variance. For NPCA, it is easy to see that larger than average proportion of variance is explained by PCs with corresponding eigenvalue larger than 1.

However, the second eigenvalue $\lambda_2 = 0.99$ is so close to 1 that we have decided to discuss also the second PC. The coefficients of the linear combinations are given by the eigenvectors

$$\gamma_1 = (0.32, 0.16, 0.37, 0.38, 0.39, 0.39, 0.39, 0.37)^\top$$

and

$$\gamma_2 = (0.39, 0.85, 0.03, -0.04, -0.13, -0.16, -0.17, -0.22)^\top.$$

In this exercise, it is very important to keep in mind the meaning of the measurements. Larger values correspond here to longer, i.e., worse times. The first PC is positively related to all original variables and it can be interpreted as the arithmetic average of the records with slightly smaller weight of the record on 200m track, see also the correlations in Figure 9.4. In Figure 9.3,

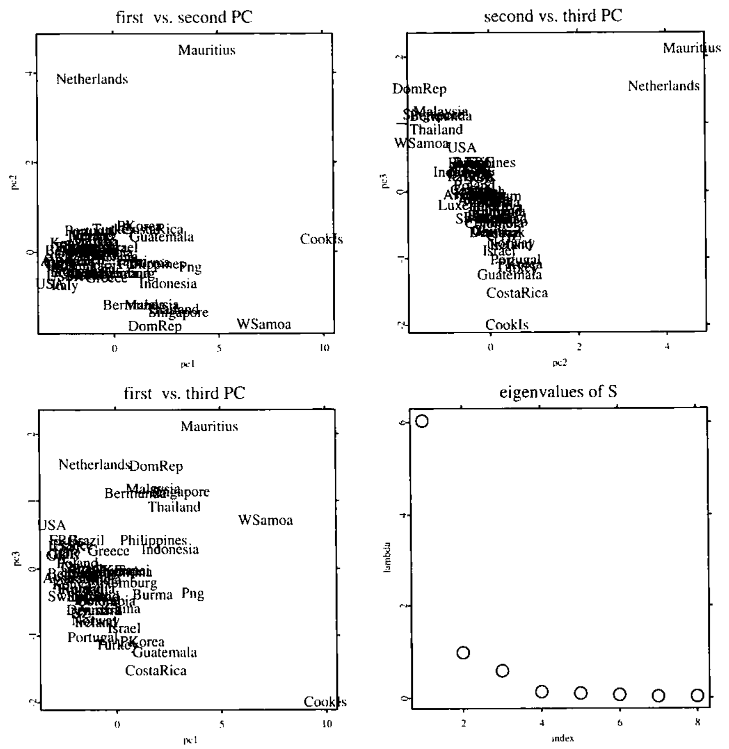

Fig. 9.3. Scatterplots of the first three principal components and a screeplot of the eigenvalues, athletic records data set. SMSnpcathletic

we can see that large values of this "average time" component are achieved in Cook Islands, West Samoa, and Mauritius. On contrary, fastest times are achieved in USA.

The second principal component is strongly positively related to 200m and important positive component is also the 100m record whereas longer tracks show mostly negative relationship. The second principal components separates Mauritius and Netherlands which shows poor records in 200m.

In Figure 9.4, we see that two principal components explain very well all original variables. Using only one PC would lead to much worse explanation of the 200m records.

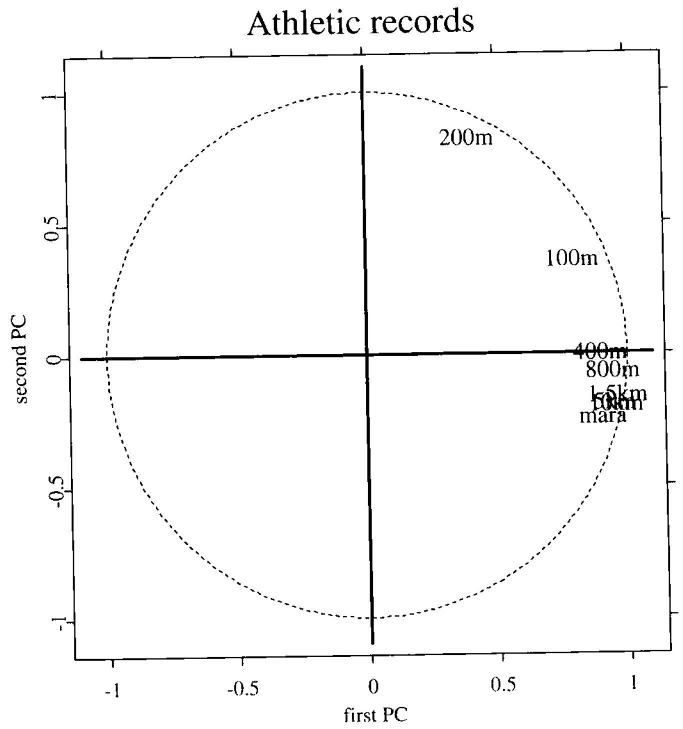

Fig. 9.4. Correlations of the first two principal components with the original variables in the athletic records data set. SMSnpcathletic

EXERCISE 9.6. *Apply a PCA to* $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, *where* $0 < \rho < 1$. *Now change the scale of* X_1, *i.e., consider the covariance of* cX_1 *and* X_2, *where* $c > 1$. *How do the PC directions change with the screeplot?*

The spectral decomposition of matrix Σ has already been investigated in Exercise 2.7. Recall that we have

$$\Sigma = \Gamma \Lambda \Gamma^\top = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Since $\rho > 0$, the PCs are $Y_1 = (X_1 + X_2)/\sqrt{2}$ and $Y_1 = (X_1 - X_2)/\sqrt{2}$.

Multiplying X_1 by constant $c > 0$ leads the covariance matrix:

$$\mathrm{Var}\{(cX_1, X_2)^\top\} = \Sigma(c) = \begin{pmatrix} c^2 & c\rho \\ c\rho & 1 \end{pmatrix}.$$

The spectral decomposition of $\Sigma(c)$ can be derived similarly as in Exercise 2.7. The eigenvalues of $\Sigma(c)$ are solutions to:

$$\begin{vmatrix} c^2 - \lambda & c\rho \\ c\rho & 1 - \lambda \end{vmatrix} = 0.$$

Hence the eigenvalues are

$$\lambda_{1,2}(c) = \frac{1}{2}\left(c^2 + 1 \pm \sqrt{(c^2-1)^2 + 4c^2\rho^2}\right).$$

The eigenvector corresponding to λ_1 can be computed from the system of linear equations:

$$\begin{pmatrix} c^2 & c\rho \\ c\rho & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

which implies that $x_1 = x_2(\lambda_1 - 1)/c\rho$ and the first PC is pointing in the direction $(cX_1)(\lambda_1 - 1)/c\rho + X_2$.

Next, observe that $\lambda_1 > 1$ and the function $\lambda_1(c)/c$ is increasing in c. Hence, $x_1 > x_2$ and, furthermore, the ratio of x_1 and x_2 is an increasing function of c.

Summarizing the above results, we can say that as c increases, the first eigenvalue λ_1 becomes larger and the rescaled random variable cX_1 gains more weight in the first principal component.

The choice of scale can have a great impact on the resulting principal components. If the scales differ, it is recommended to perform the Normalized PCA (NPCA), i.e., to standardize each variable by its standard deviation.

EXERCISE 9.7. *Suppose that we have standardized some data using the Mahalanobis transformation. Would it be reasonable to apply a PCA?*

Standardizing any given data set $\mathcal{X}$ by the Mahalanobis transformation leads to a data set $\mathcal{Z} = \mathcal{X}\mathcal{S}^{-1/2}$ with the covariance matrix

$$\mathcal{S}_\mathcal{Z} = \mathcal{S}^{-1/2}\mathcal{S}\mathcal{S}^{-1/2} = \mathcal{I}_p.$$

It immediately follows that all eigenvalues of $\mathcal{S}_\mathcal{Z}$ are equal to 1 and that the principal components of $\mathcal{Z}$ have exactly the same variances as the original variables. Hence, such analysis would be entirely useless.

Principal components analysis of $\mathcal{Z}$ leads always to this same uninteresting result.

EXERCISE 9.8. *Apply a NPCA to the U.S. crime data set in Table A.18. Interpret the results. Would it be necessary to look at the third PC? Can you see any difference between the four regions?*

The U.S. crime data set consists of the reported number of crimes in the 50 U.S. states in 1985. The crimes were classified according to 7 categories: murder, rape, robbery, assault, burglary, larceny, and auto theft. The dataset also contains identification of the region: Northeast, Midwest, South, and West.

The Normalized PCA means that, before running the analysis, all observed variables are put on the same scale.

The eigenvalues of the correlation matrix are:

$$\lambda = (4.08, 1.43, 0.63, 0.34, 0.25, 0.14, 0.13)^\top$$

and we obtain the proportions of explained variance:

$$\psi = (0.58, 0.79, 0.88, 0.93, 0.96, 0.98, 1.00)^\top.$$

The data set is well described by the first two NPCs, each of the first two NPCs describes larger than average amount of variance. The first two NPCs describe together 79% of the total variability, see also the screeplot in Figure 9.5.

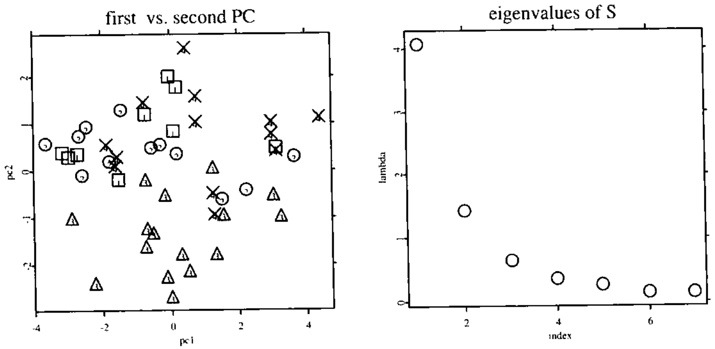

Fig. 9.5. Scatterplot of the first two principal components and a screeplot of the eigenvalues, U.S. crime data set. `SMSnpcacrime`

The first two eigenvectors are:

$$\gamma_1 = (0.28, 0.42, 0.39, 0.39, 0.44, 0.36, 0.35)^\top,$$
$$\gamma_2 = (-0.64, -0.12, 0.05, -0.46, 0.26, 0.40, 0.37)^\top.$$

The first principal component combines the numbers of all crimes with approximately constant (0.28–0.44) weights and we can interpret it as the overall crime rate, see also the correlations in Figure 9.6. The second principal component is negatively correlated with 1st and 4th variable (murder and assault) and positively correlated with the 5th till 7th variable (burglary, larceny, auto theft). The second NPC can be interpreted as "type of crime" component.

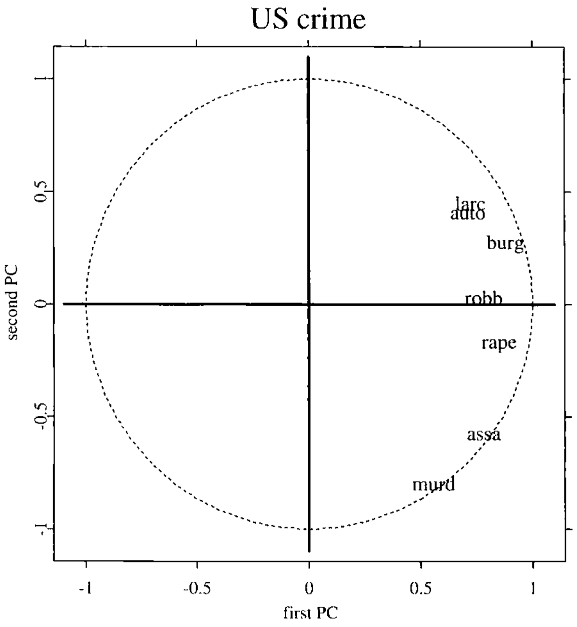

Fig. 9.6. Correlations of the first two principal components with the original variables in the U.S. crime data set. Q SMSnpcacrime

In Figure 9.5, we denote each of the four regions by a different plotting symbol. It looks as if the symbol changes in the direction of the second, type of crime, principal component. In the upper part of the graph, we see mainly circles, squares, and crosses corresponding to the regions 1, 2, and 4. In the lower part, we observe mainly triangles corresponding to the third South region. Hence, it seems that in region 3 occur more murders and assaults and less burglaries, larcenies and auto thefts than in the rest of USA.

EXERCISE 9.9. *Repeat Exercise 9.8 using the U.S. health data set in Table A.19.*

The U.S. health data set consists of reported number of deaths in the 50 U.S. states classified according to 7 categories: accident, cardiovascular, cancer, pulmonary, pneumonia flu, diabetes, and liver.

Here, we have decided to run the usual PC analysis. Normalizing the data set would mean that, in certain sense, all causes of death would have the same importance. Without normalization, we can expect that the variables responsible for the largest number of deaths will play the most prominent role in our analysis, see also Exercise 9.6 for theoretical justification.

The eigenvalues of the covariance matrix are:

$$\lambda = (8069.40, 189.22, 76.03, 25.21, 10.45, 5.76, 3.47)^\top$$

and the huge first eigenvalue stresses the importance of the first principal component. Calculating the proportions of the explained variance,

$$\psi = (0.96, 0.99, 0.99, 1.00, 1.00, 1.00, 1.00)^\top,$$

we see that the first PC explains 96% of the total variability. The screeplot is plotted in Figure 9.7.

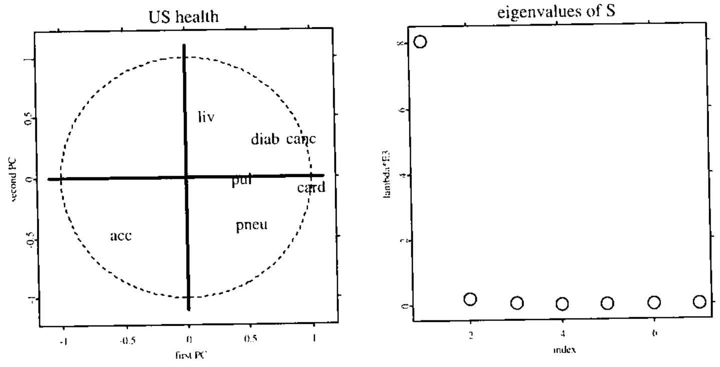

Fig. 9.7. Correlations of the first two principal components with the original variables and the screeplot for the U.S. health data set. Q SMSpcahealth

The first (most important) eigenvectors is:

$$\gamma_1 = (-0.06, 0.94, 0.34, 0.03, 0.02, 0.03, 0.01)^\top$$

and we see that the first PC reflects the most common causes of death: cardiovascular diseases and, with smaller weight, cancer. The second eigenvector,

$$\gamma_2 = (-0.34, -0.34, 0.86, 0.01, -0.11, 0.09, 0.11)^\top,$$

is strongly positively correlated with cancer and less strongly negatively correlated with cardiovascular and pulmonary diseases, see also Figure 9.7. The first principal component explains satisfactorily only variables cardiovascular and cancer.

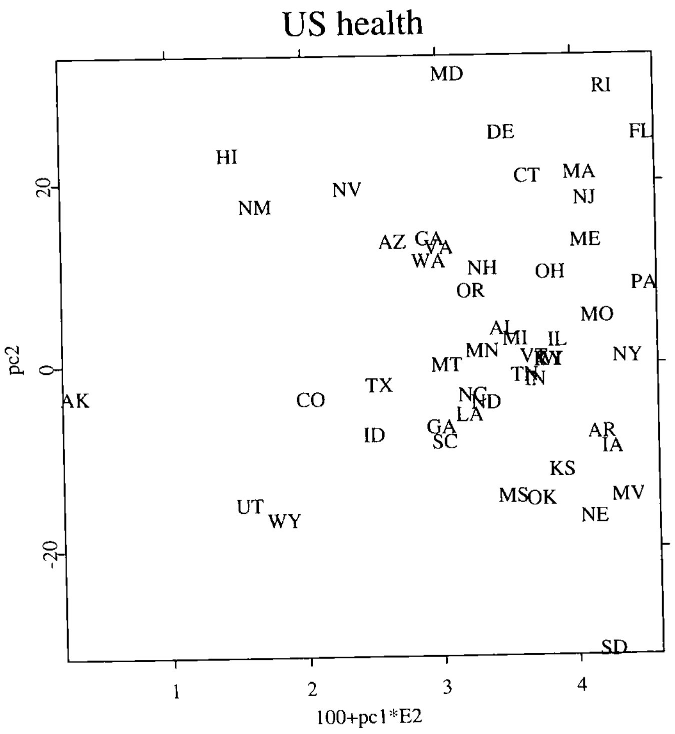

Fig. 9.8. Scatterplot of the first two principal components for U.S. health data set. SMSpcahealth

In Figure 9.8, we show the values of the first two PCs for the 50 observed U.S. states. Keeping in mind the meaning of the principal components, we should see the states with large number of deaths due to cardiovascular diseases and cancer on the right hand side (Florida, New York, Pennsylvania). From the point of view of the first PC, the best quality of life can be found in Arkansas, Hawaii, New Mexico, Wyoming, and Colorado. The much less important second PC suggests that cancer is more common cause of death in Maryland than in South Dakota.

EXERCISE 9.10. *Do a NPCA on the Geopol data set, Table A.10, which compares 41 countries with respect to different aspects of their development. Why or why not would a PCA be reasonable here?*

The Geopol data set contains a comparison of 41 countries according to 10 political and economic parameters. We will perform the analysis without the first variable, size of population. The variables to be analyzed, X_2–X_9 are: gross internal product per habitant (giph), rate of increase of the population (ripo), rate of urban population (rupo), rate of illiteracy (rlpo), rate of students (rspo), expected lifetime (eltp), rate of nutritional needs realized (rnnr), number of newspaper and magazines per 1000 habitants (nunh), and number of televisions per 1000 inhabitants (nuth).

Clearly, these variables are measured on very different scales and, in order to produce trustworthy results, the data set has to be normalized. In this exercise, we have to perform NPCA.

The eigenvalues of the correlation matrix are:

$$\lambda = (5.94, 0.87, 0.70, 0.54, 0.43, 0.18, 0.15, 0.12, 0.08)^\top$$

and we obtain the percentages of explained variance:

$$\psi = (0.66, 0.76, 0.83, 0.89, 0.94, 0.96, 0.98, 0.99, 1.00)^\top.$$

The screeplot is plotted in Figure 9.9. It would suffice to keep only one NPC, but we decide to keep the first three principal components although Y_2 and Y_3 contribute only little to the total variability.

The coefficients of the first three normalized principal components are given by the first three eigenvectors:

$$\gamma_1 = (0.34, -0.34, 0.29, -0.36, 0.30, 0.37, 0.28, 0.33, 0.37)^\top,$$
$$\gamma_2 = (0.41, 0.38, 0.23, 0.20, 0.16, -0.20, -0.61, 0.36, 0.19)^\top,$$
$$\gamma_3 = (-0.18, 0.37, 0.34, -0.02, 0.66, -0.05, 0.14, -0.49, 0.06)^\top.$$

The correlations of $Y_1, \dots, Y_3$ with the original variables are plotted in Figure 9.10.

From the correlations plotted in Figure 9.10, we can interpret the first PC as the overall quality of life component: notice that it is positively related to the

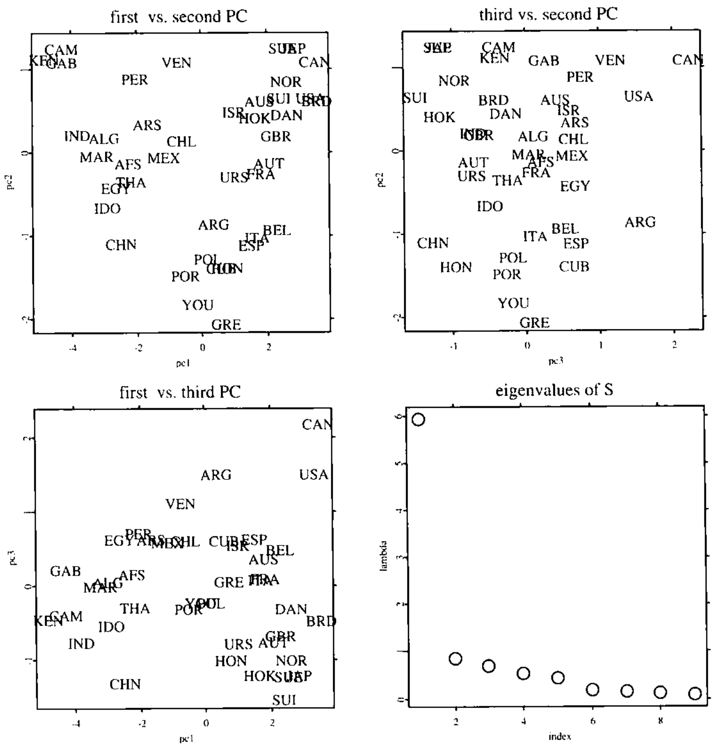

Fig. 9.9. Scatterplots of the first three principal components and a screeplot of the eigenvalues, Geopol data set. SMSnpcageopol

all variables apart of rate of increase of the population and rate of illiteracy. In Figure 9.9, we can see that large values of this component are achieved in the former West Germany (BRD), Canada, and USA. Smallest values of this component are observed in Kenya, Cameroon, Gabon, and India.

The second PC seems to point mainly in the direction opposite to the rnnr (rate of nutritional needs realized). The third PC is positively correlated to the rate of students and negatively correlated to the number of newspapers. From Figure 9.9, we can see that already one PC is enough to explain substantial part of the variability of all variables.

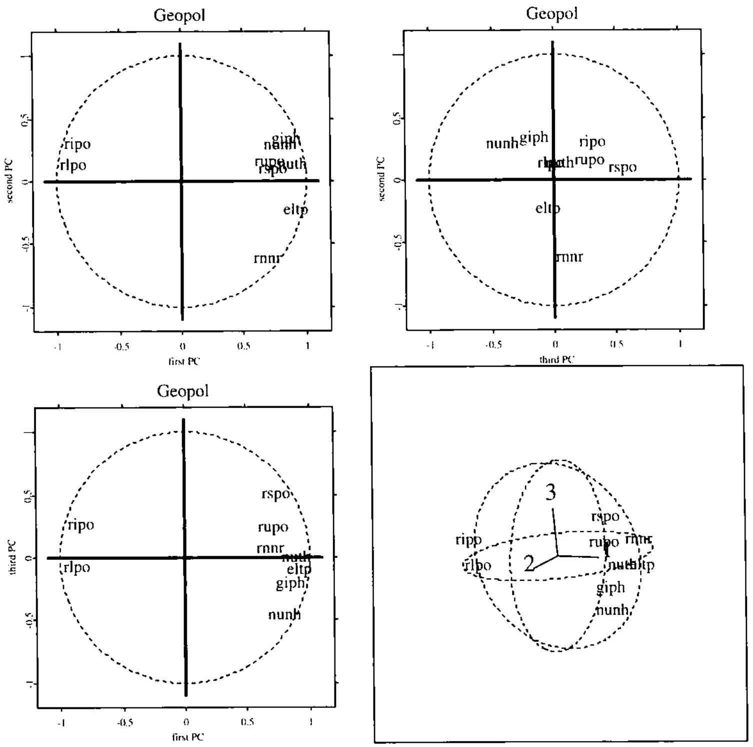

Fig. 9.10. Correlations of the first three principal components with the original variables in the Geopol data set. SMSnpcageopol

EXERCISE 9.11. *Let U be an uniform random variable on $[0,1]$. Let $a = (a_1, a_2, a_3)^\top \in \mathbb{R}^3$ be a vector of constants. Suppose that $X = (X_1, X_2, X_3)^\top = aU$. What do you expect the NPCs of X to be?*

Let us assume that $a_i \neq 0$, $i = 1, 2, 3$. Next, normalizing the random vector X by subtracting its expected value and by dividing it by its standard deviation leads to the normalized random vector

$$Z = \left\{\operatorname{diag}\left(a^2\sigma_U^2\right)\right\}^{-1/2}(X - EX) = \left\{\operatorname{diag}\left(a^2\sigma_U^2\right)\right\}^{-1/2} a(U - EU)$$

with the variance matrix

$$\begin{aligned}\operatorname{Var}(Z) &= \left\{\operatorname{diag}\left(a^2\sigma_U^2\right)\right\}^{-1/2}\operatorname{Var}(X)\left\{\operatorname{diag}\left(a^2\sigma_U^2\right)\right\}^{-1/2}\\ &= \left\{\operatorname{diag}\left(a^2\sigma_U^2\right)\right\}^{-1/2} a\sigma_U^2 a^\top \left\{\operatorname{diag}\left(a^2\sigma_U^2\right)\right\}^{-1/2}\\ &= \left(\frac{a_i a_j}{\operatorname{abs} a_i \operatorname{abs} a_j}\right)_{i,j=1,2,3}\\ &= \{\operatorname{sign}(a_i a_j)\}_{i,j=1,2,3}\,.\end{aligned}$$

Clearly, the rank of the variance matrix $\operatorname{Var}(Z)$ is equal to 1 and it follows that it has only one nonzero eigenvalue. Hence, the spectral decomposition of $\operatorname{Var}(Z)$ leads to only one principal component explaining 100% of total variability of Z.

The NPC can be written as

$$\begin{aligned}Y_1 &= \frac{1}{\sqrt{3}}\{\operatorname{sign}(a_1)Z_1 + \operatorname{sign}(a_2)Z_2 + \operatorname{sign}(a_3)Z_3\}\\ &= \frac{1}{\sqrt{3}}\{\operatorname{sign}(a_1)a_1U + \operatorname{sign}(a_2)a_2U + \operatorname{sign}(a_3)a_3U\}\\ &= U\frac{\operatorname{abs}(a_1)+\operatorname{abs}(a_2)+\operatorname{abs}(a_3)}{\sqrt{3}},\end{aligned}$$

i.e., the normalized principal components analysis of $X = aU$ leads us back to the one-dimensional random variable U.

EXERCISE 9.12. *Let U_1 and U_2 be two independent uniform random variables on $[0,1]$. Suppose that $X = (X_1, X_2, X_3, X_4)^\top$ where $X_1 = U_1$, $X_2 = U_2$, $X_3 = U_1 + U_2$ and $X_4 = U_1 - U_2$. Compute the correlation matrix P of X. How many PCs are of interest? Show that $\gamma_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 1, 0\right)^\top$ and $\gamma_2 = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 1\right)^\top$ are eigenvectors of P corresponding to the non trivial λ's. Interpret the first two NPCs obtained.*

For random variables U_1 and $U_2 \sim U[0,1]$, we have $EU_1 = 1/2$ and $Var\,U_1 = Var\,U_2 = 1/12$. It follows that also $Var\,X_1 = Var\,X_2 = 1/12$.

For the variance of $X_3 = U_1 + U_2$ and $X_4 = U_1 - U_2$, we obtain

$$Var(X_3) = Var(X_4) = Var(U_1) + Var(U_2) = \frac{1}{6}$$

since U_1 and U_2 are independent. The covariances can be calculated as

$$\operatorname{Cov}(X_1, X_3) = \operatorname{Cov}(U_1, U_1 + U_2) = Var(U_1) + \operatorname{Cov}(U_1, U_2) = \frac{1}{12}$$

and

$$\operatorname{Cov}(X_3, X_4) = \operatorname{Cov}(U_1 + U_2, U_1 - U_2) = Var(U_1) - Var(U_2) = 0.$$

The remaining elements of the variance matrix can be calculated in the same way leading to

$$\operatorname{Var}(X) = \frac{1}{12}\begin{pmatrix}1 & 0 & 1 & 1\\ 0 & 1 & 1 & -1\\ 1 & 1 & 2 & 0\\ 1 & -1 & 0 & 2\end{pmatrix}.$$

Dividing each row and each column by the square root of the corresponding diagonal element gives the correlation matrix

$$P = \begin{pmatrix}1 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ 0 & 1 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 & 0\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 1\end{pmatrix}.$$

Now it is easy to verify that γ_1 and γ_2 are indeed eigenvectors of the correlation matrix P since

$$P\gamma_1 = \begin{pmatrix}1 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ 0 & 1 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 & 0\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 1\end{pmatrix} \cdot \begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\\ 1\\ 0\end{pmatrix} = \begin{pmatrix}\sqrt{2}\\ \sqrt{2}\\ 2\\ 0\end{pmatrix} = 2\gamma_1.$$

and, similarly, $P\gamma_2 = 2\gamma_2$. This, by the way, implies that also $P(\gamma_2 + \gamma_1) = 2(\gamma_1 + \gamma_2)$ and hence, any linear combination of γ_1 and γ_2 is also an eigenvector of P with the same eigenvalue.

Thus, we have the eigenvalues $\lambda_1 = \lambda_2 = 2$. The remaining two eigenvalues, λ_3 and λ_4 are equal to 0 because the rank of the correlation matrix is equal to 2.

The first two NPCs are not determined uniquely. Choosing the coefficients as γ_1 and γ_2 and keeping in mind that these coefficients correspond to the normalized variables we have:

$$\begin{aligned}Y_1 &= \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2 + \frac{X_3}{\sqrt{2}} = \sqrt{2}(U_1 + U_2)\\ Y_2 &= \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2 + \frac{X_4}{\sqrt{2}} = \sqrt{2}(U_1 - U_2).\end{aligned}$$

The NPCs, Y_1 and Y_2, can be now interpreted respectively as the sum and the difference of U_1 and U_2.

EXERCISE 9.13. *Simulate a sample of size $n = 50$ for the r.v. X in Exercise 9.12 and analyze the results of a NPCA.*

Performing the NPCA for the simulated data set, we obtain the eigenvalues:

$$\widehat{\lambda} = (2.11, 1.89, 0.00, 0.00)^\top$$

and the proportions of the explained variance:

$$\widehat{\psi} = (0.53, 1.00, 1.00, 1.00)^\top .$$

These numbers correspond well to the theoretical values $\lambda_1 = \lambda_2 = 2$ derived in Exercise 9.12. The remaining two eigenvalues are equal to zero because of the linear dependencies in the data set. The screeplot is plotted in Figure 9.11 and we see that the first two NPCs explain each approximately 50% of the variability whereas the other two NPCs do not explain anything.

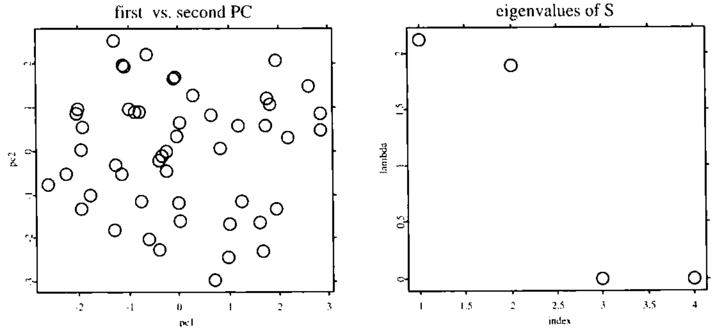

Fig. 9.11. Scatterplots of the first two principal components and a screeplot of the eigenvalues, simulated data set. Q SMSnpcasimu

The first two eigenvectors are

$$\widehat{\gamma}_1 = (0.32, -0.64, -0.26, 0.65)^\top$$

and

$$\widehat{\gamma}_2 = (0.65, 0.28, 0.67, 0.23)^\top$$

and the resulting values for the 50 NPCs are plotted in Figure 9.11. Rewriting the resulting NPCs in terms of the original variables and rounding the coefficients leads that the first NPC points approximately in the direction $U_1 - 2U_2$ and the second NPC in the direction $2U_1 + U_2$. This result differs from the eigenvectors γ_1 and γ_2 calculated in Exercise 9.12 because γ_1 and γ_2 are not uniquely defined.

In Figure 9.12, we plot the correlation of the NPCs with the normalized variables $X_1, \ldots, X_4$. The correlations correspond to the coefficients of the NPCs.

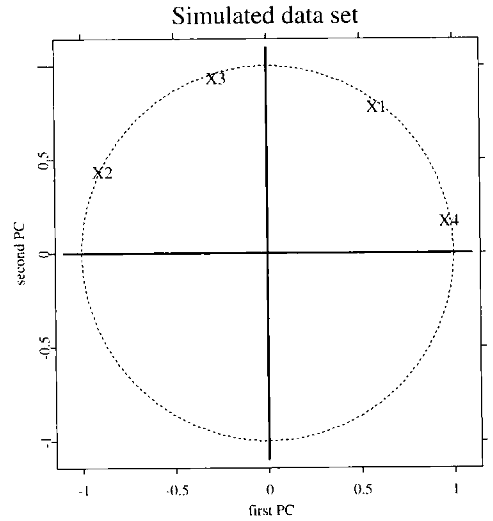

Fig. 9.12. Correlations of the first two principal components with the original variables in the simulated data set. Q SMSnpcasimu

All of the original variables are perfectly explained by two NPCs because all four points are lying on the unit circle.

The simulated data set changes with every simulation. One can observe that the eigenvalues $\widehat{\lambda}$ do not vary a lot for different runs of the simulation. However, the eigenvectors can vary a lot due to the fact that they are not defined uniquely.

10

Factor Analysis

> A certain selection and discretion must be used in producing a realistic effect.
> Sherlock Holmes in "A Case of Identity"

In factor analysis, we address the same problem of reducing the dimension of a multivariate random variable, but we want to fix, from the start, the number of factors. Each factor will then be interpreted as a latent characteristic of the individuals revealed by the original variables.

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

The ultimate goal is to find underlying reasons that explain the data variation. In achieving this goal we need to check the relation of the factors and original variables and give them an interpretation in the framework of how the data were generated.

Factor Analysis Model

The factor analysis model used in practice is:

$$X = \mathcal{Q}F + U + \mu, \tag{10.1}$$

where $\mathcal{Q}$ is a $(p \times k)$ matrix of the (nonrandom) loadings of the common factors $F(k \times 1)$ and U is a $(p \times 1)$ matrix of the (random) specific factors. It is assumed that the common factors F are uncorrelated random variables and that the specific factors are uncorrelated and have zero covariance with the common factors. More precisely, it is assumed that: $EF = 0$, $\text{Var}(F) = \mathcal{I}_k$, $EU = 0$, $\text{Cov}(U_i, U_j) = 0$, $i \neq j$, and $\text{Cov}(F, U) = 0$.

The random vectors F and U are unobservable. Define $\text{Var}(U) = \Psi = \text{diag}(\psi_{11}, \dots, \psi_{pp})$; then the variance matrix of X can be written as $\text{Var}(X) = \Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$, and we have for the ith component of the random vector X that $\sigma_{X_jX_j} = Var(X_j) = \sum_{\ell=1}^k q_{j\ell}^2 + \psi_{jj}$. The quantity $h_j^2 = \sum_{\ell=1}^k q_{j\ell}^2$ is called the communality and ψ_{jj} the specific variance. The objective of factor analysis is to find a small number, k, of common factors leading to large communalities and small specific variances.

Estimation of the Factor Model

In practice, we have to find estimates $\widehat{\mathcal{Q}}$ of the loadings $\mathcal{Q}$ and estimates $\widehat{\Psi}$ of the specific variances Ψ such that $\mathcal{S} = \widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top + \widehat{\Psi}$, where $\mathcal{S}$ denotes the empirical covariance of $\mathcal{X}$. The most commonly used methods are the following:

> The maximum likelihood method is based on the assumption of normality. The equations resulting from the maximization of the log-likelihood under the assumption $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$ are complicated and have to be solved by iterative numerical algorithms.
>
> The method of principal factors starts with a preliminary estimate of $\widehat{h}_j^2$ and the specific variances $\widehat{\psi}_{jj} = 1 - \widehat{h}_j$. In the next step, the matrix of loadings is estimated from the spectral decomposition of the reduced covariance matrix $\mathcal{S} - \widehat{\Psi}$. This procedure can be iterated until convergence is reached.
>
> The principal component method starts by obtaining estimated loadings $\widehat{\mathcal{Q}}$ from a spectral decomposition of the matrix $\mathcal{S}$. The specific variances are then estimated by the diagonal elements of the matrix $\mathcal{S} - \widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top$.

Rotation

Suppose that $\mathcal{G}$ is an orthogonal matrix. Then X in (10.1) can also be written as $X = (\mathcal{Q}\mathcal{G})(\mathcal{G}^\top F) + U + \mu$. This implies that the factors are not defined uniquely because equivalent models with factors $\mathcal{G}^\top F$ and loadings $\mathcal{Q}\mathcal{G}$ are valid for an arbitrary orthogonal matrix $\mathcal{G}$. In practice, the choice of an appropriate rotation $\mathcal{G}$ of the loadings $\mathcal{Q}$ results in a matrix of loadings $\mathcal{Q}^* = \mathcal{Q}\mathcal{G}$ that are easier to interpret.

A well-known algorithm for choosing a reasonable rotation of the factor loadings is given by the varimax rotation method proposed by Kaiser (1985). The idea of this popular method is to find the angles that maximize the sum of the variances of the squared loadings q_{ij}^* within each column of $\mathcal{Q}^*$. The varimax criterion attempts to split the variables automatically into disjoint sets, each associated with one factor.

Strategy for Factor Analysis

1. Perform a principal component factor analysis, look for suspicious observations, try varimax rotation.
2. Perform maximum likelihood factor analysis, including varimax rotation.
3. Compare the factor analyses: do the loadings group in the same manner?
4. Repeat the previous steps for other numbers of common factors.

After the estimation and interpretation of factor loadings and communalities, estimate the factor values. The estimated values of the factors are called the factor scores and may be useful in the interpretation as well as in the diagnostic analysis. To be more precise, the factor scores are estimates of the unobserved k-dimensional random vectors F for each individual x_i, $i = 1, \dots, n$. Johnson & Wichern (1998) describe three methods that in practice yield very similar results. The regression method (see Exercise 10.6) is also described in Härdle & Simar (2003, section 10.3).

EXERCISE 10.1. *Compute the orthogonal factor model for*

$$\Sigma = \begin{pmatrix} 1.0 & 0.9 & 0.7 \\ 0.9 & 1.0 & 0.4 \\ 0.7 & 0.4 & 1.0 \end{pmatrix}.$$

We have to find loadings $\mathcal{Q}$ and specific variances Ψ satisfying the decomposition $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$. The problem is difficult to solve due to the non-uniqueness of the solutions. An acceptable technique is to impose some additional constraints such as: $\mathcal{Q}^\top\Psi^{-1}\mathcal{Q}$ is diagonal.

The factor analysis without any constraints has $pk + k$ unknown parameters of the matrix $\mathcal{Q}$ and the diagonal of Ψ. The diagonality of $\mathcal{Q}^\top\Psi^{-1}\mathcal{Q}$ introduces $\frac{1}{2}\{k(k-1)\}$ constraints. Therefore, the degrees of freedom of a model with k factors is $d = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$.

If $d < 0$, then there are infinitely many solutions. If $d = 0$ the there is an unique solution to the problem (except for rotation). In practice we usually have that $d > 0$ and an exact solution does not exist. Evaluating the degrees of freedom, d, is particularly important, because it already gives an idea of the upper bound on the number of factors we can hope to identify in a factor model.

If $p = 3$, we can identify at most $k = 1$ factor. This factor is then given uniquely since $d = \frac{1}{2}(3-1)^2 - \frac{1}{2}(3+1) = 0$. Implementing a simple iterative procedure, i.e., the principal factor method described in the introduction, we arrive to the following exact solution:

$$\Sigma = \begin{pmatrix} 1.0 & 0.9 & 0.7 \\ 0.9 & 1.0 & 0.4 \\ 0.7 & 0.4 & 1.0 \end{pmatrix}$$
$$= \begin{pmatrix} 1.2549 \\ 0.7172 \\ 0.5578 \end{pmatrix} (1.2549, 0.7172, 0.5578) + \begin{pmatrix} -0.5748 & 0.0000 & 0.0000 \\ 0.0000 & 0.4857 & 0.0000 \\ 0.0000 & 0.0000 & 0.6889 \end{pmatrix}.$$

The obvious disadvantage of this unique solution is that it cannot be interpreted as a factor analysis model since the specific variance ψ_{11} cannot be negative.

Hence, the ability to find a unique solution of the orthogonal factor model does not have to lead to the desired result. `SMSfactsigma`

EXERCISE 10.2. *Using the bank data set in Table A.2, how many factors can you find with the method of principal factors?*

The number of variables is $p = 6$. For $k = 3$ factors, the orthogonal factor model would have

$$d = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) = 4.5 - 4.5 = 0$$

degrees of freedom, see Exercise 10.1. It follows that for 3 factors, we would have an exact solution. Unfortunately, as we have seen in Exercise 10.1, the unique exact solution does not have to be interpretable. In this situation, it is advisable to work with at most $k = 2$ factors.

The empirical correlation analysis calculated from the given 6-dimensional data set is:

$$\mathcal{R} = \begin{pmatrix} 1.0000 & 0.2313 & 0.1518 & -0.1898 & -0.0613 & 0.1943 \\ 0.2313 & 1.0000 & 0.7433 & 0.4138 & 0.3623 & -0.5032 \\ 0.1518 & 0.7433 & 1.0000 & 0.4868 & 0.4007 & -0.5165 \\ -0.1898 & 0.4138 & 0.4868 & 1.0000 & 0.1419 & -0.6230 \\ -0.0613 & 0.3623 & 0.4007 & 0.1419 & 1.0000 & -0.5940 \\ 0.1943 & -0.5032 & -0.5165 & -0.6230 & -0.5940 & 1.0000 \end{pmatrix}.$$

The communalities h_j^2, $j = 1, \dots, 6$, measure the part of variance of each variable that can be assigned to the common factors. One possibility to define a reasonable starting estimates is to set $\widehat{h}_j^2 = \max_{i \neq j, i=1,\dots,6} |r_{X_j X_i}|$. For the Swiss bank notes, we obtain

$$\widehat{h}^2 = (\widehat{h}_1^2, \dots, \widehat{h}_6^2)^\top = (0.2313, 0.7433, 0.7433, 0.6230, 0.5940, 0.6230)^\top.$$

The estimates of the specific variances ψ_{jj}, $j = 1, \dots, 6$ are

$$\widehat{\psi} = (\widehat{\psi}_{11}, \dots, \widehat{\psi}_{66})^\top = (0.7687, 0.2567, 0.2567, 0.3770, 0.4060, 0.3770)^\top$$

and the reduced correlation matrix $\mathcal{R} - \widehat{\Psi}$ is

$$\mathcal{R} - \text{diag}(\widehat{\psi}) = \begin{pmatrix} 0.2313 & 0.2313 & 0.1518 & -0.1898 & -0.0613 & 0.1943 \\ 0.2313 & 0.7433 & 0.7433 & 0.4138 & 0.3623 & -0.5032 \\ 0.1518 & 0.7433 & 0.7433 & 0.4868 & 0.4007 & -0.5165 \\ -0.1898 & 0.4138 & 0.4868 & 0.6230 & 0.1419 & -0.6230 \\ -0.0613 & 0.3623 & 0.4007 & 0.1419 & 0.5940 & -0.5940 \\ 0.1943 & -0.5032 & -0.5165 & -0.6230 & -0.5940 & 0.6230 \end{pmatrix}.$$

The vector of the eigenvalues of the reduced correlation matrix is:

$$\lambda = (2.6214, 0.7232, 0.4765, 0.0054, -0.0845, -0.1841)^\top.$$

At this step, some of the eigenvalues can be negative. The possibility that the reduced correlation matrix does not have to be positive definite has to be taken into account in the computer implementation of the factor analysis.

The matrix of eigenvectors of the reduced correlation matrix is:

$$\Gamma = \begin{pmatrix} -0.0011 & -0.6225 & 0.0488 & -0.1397 & 0.7663 & 0.0582 \\ 0.4832 & -0.4510 & -0.0727 & -0.5783 & -0.4575 & -0.1185 \\ 0.5019 & -0.3314 & -0.1077 & 0.7670 & -0.1328 & 0.1438 \\ 0.3974 & 0.3489 & -0.6039 & -0.0434 & 0.3510 & -0.4802 \\ 0.3543 & 0.1661 & 0.7768 & 0.0604 & 0.1328 & -0.4714 \\ -0.4807 & -0.3872 & -0.1125 & 0.2285 & -0.2123 & -0.7135 \end{pmatrix}.$$

With $k = 2$ factors, we obtain the factor loadings

$$\widehat{\mathcal{Q}} = \begin{pmatrix} -0.0011 & -0.6225 \\ 0.4832 & -0.4510 \\ 0.5019 & -0.3314 \\ 0.3974 & 0.3489 \\ 0.3543 & 0.1661 \\ -0.4807 & -0.3872 \end{pmatrix} \begin{pmatrix} \sqrt{2.6214} & 0 \\ 0 & \sqrt{0.7232} \end{pmatrix} = \begin{pmatrix} -0.0018 & -0.5294 \\ 0.7824 & -0.3835 \\ 0.8127 & -0.2819 \\ 0.6435 & 0.2967 \\ 0.5736 & 0.1412 \\ -0.7783 & -0.3293 \end{pmatrix}.$$

If the variables are normalized, i.e., if the analysis is based on the correlation matrix, the factor loadings $\mathcal{Q}$ are the correlations between the original variables and the unobserved factors.

The final estimates of the two factor model, given in Table 10.1, were obtained by several iterations of the described algorithm. It is interesting to notice that the final estimates are rather different from the starting values.

The next step in the analysis is a rotation of the two factor loadings leading to better interpretable results. In Figure 10.1 you can see both the original factor loadings as given in Table 10.1 and the same factor loadings rotated by the angle $5\pi/12$ counterclockwise. The rotation, i.e., multiplication of the factor loadings by the rotation matrix

	Estimated factor loadings		Communalities	Specific variances
	$\hat{q}_1$	$\hat{q}_2$	$\hat{h}_j^2$	$\hat{\psi}_{jj} = 1 - \hat{h}_j^2$
1 length	−0.0046	−0.5427	0.2946	0.7054
2 height measured left	0.7888	−0.4107	0.7910	0.2090
3 height measured right	0.7996	−0.2982	0.7283	0.2717
4 lower frame distance	0.5929	0.1953	0.3896	0.6104
5 upper frame distance	0.5109	0.1068	0.2724	0.7276
6 length of the diagonal	−0.8784	−0.4436	0.9683	0.0317

Table 10.1. Estimated factor loadings, communalities, and specific variances, PFM, Swiss bank notes data set. `SMSfactbank`

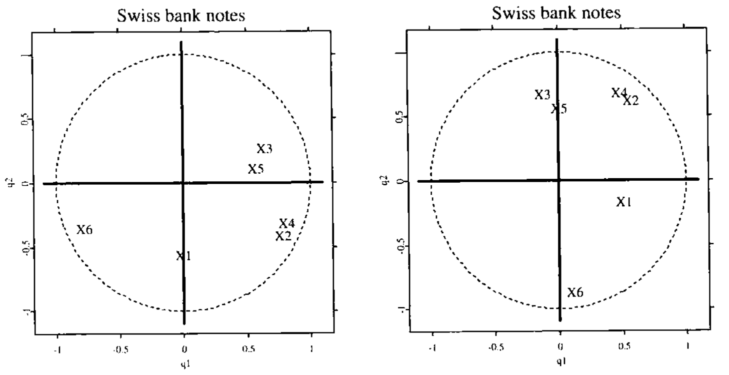

Fig. 10.1. Rotation of the factor loadings in the Swiss bank notes data set. The original and rotated factor loadings are on the left and right hand side, respectively. `SMSfactbank`

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

where $\theta = 5\pi/12$ changes only the factor loadings and their interpretation. In Figure 10.1, we suggest rotation leading to one factor positively correlated to X_1, X_2, and X_4 whereas the second factor is strongly positively related to X_2, X_3, X_4, and X_5 and strongly negatively related to X_6.

Further insight into the factors might be achieved by estimating their values for our observations. This part of the factor analysis will be demonstrated in detail in Exercise 10.6.

EXERCISE 10.3. *An example of an orthogonal matrix in two-dimensions is the so-called rotation matrix*

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

representing a clockwise rotation of the coordinate axes by the angle θ. Generalize the two-dimensional rotation matrix $\mathcal{G}(\theta)$ to 3-dimensional space.

The two-dimensional rotation matrix $\mathcal{G}(\theta)$ rotates two-dimensional coordinates counterclockwise by angle θ with respect to the origin $(0,0)^\top$, see Figure 10.1 for an illustration.

In 3-dimensional space, we can fix three angles, θ_1, θ_2, and θ_3 specifying three two-dimensional rotations. In the first step, we can rotate the given three-dimensional points in the first two coordinates and keep the third coordinate fixed, this can be achieved by the rotation matrix:

$$\mathcal{G}_{12}(\theta_3) = \begin{pmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Rotating the points only in the first coordinates can be described as a rotation of the thee-dimensional cloud of points around the third axis by angle θ_3.

The rotation in the first and third coordinate (around the second axis) is achieved by:

$$\mathcal{G}_{13}(\theta_2) = \begin{pmatrix} \cos\theta_2 & 0 & \sin\theta_2 \\ 0 & 1 & 0 \\ -\sin\theta_2 & 0 & \cos\theta_2 \end{pmatrix}$$

and for the rotation in the second and third coordinate (around the first axis), we have:

$$\mathcal{G}_{23}(\theta_1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_1 & \sin\theta_1 \\ 0 & -\sin\theta_1 & \cos\theta_1 \end{pmatrix}.$$

Arbitrary rotation in three-dimensional space can now be written as a combination of the two-dimensional rotations $\mathcal{G}_{23}(\theta_1)$, $\mathcal{G}_{13}(\theta_2)$, and $\mathcal{G}_{12}(\theta_3)$. We define the general three-dimensional rotation matrix:

$$\mathcal{G}_{123}(\theta_1, \theta_2, \theta_3) = \mathcal{G}_{23}(\theta_1)\mathcal{G}_{13}(\theta_2)\mathcal{G}_{12}(\theta_3).$$

Similarly, the two-dimensional rotation matrices can be used to define a rotation in n-dimensional space.

EXERCISE 10.4. *Perform a factor analysis on the type of families in the French food data set A.9. Rotate the resulting factors in a way which provides a reasonable interpretation. Compare your result to the varimax method.*

The French food data set contains average expenditures on seven types of food for different types of families (manual workers, employees, managers) in France. The abbreviations MA, EM, and CA denote respectively manual workers, employees, and managers. The number denotes the number of children. In this exercise, we consider the dataset as consisting of 7 measurement of the 12 type of family variables.

A first look at the data set reveals that the structure of expenditures strongly depends on the type of food. Hence, before running the factor analysis, we put all measurements on the same scale by standardizing the expenditures for each type of food separately.

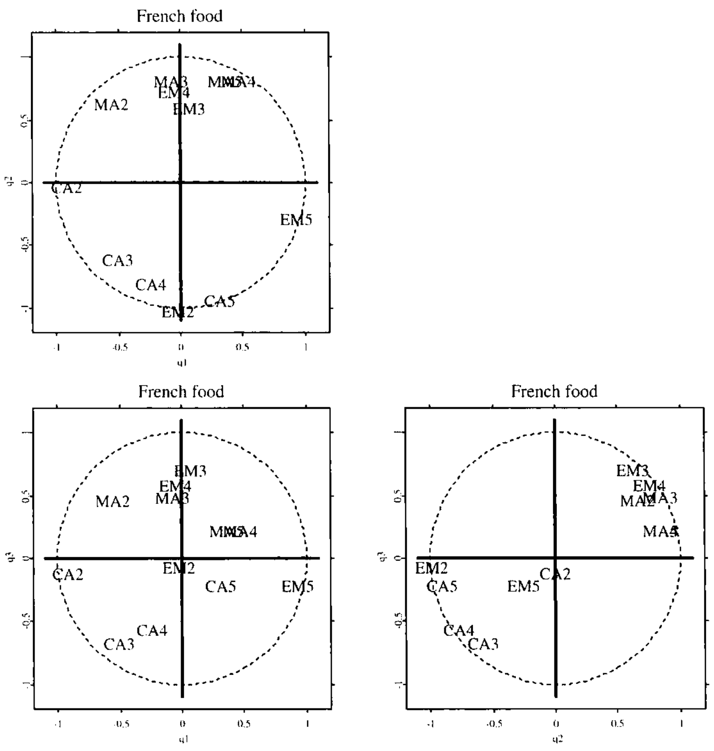

Fig. 10.2. Factor loadings for the French food data set after manual rotation of the factor loading obtained by PFM method. SMSfactfood

We choose $k = 3$ factors. The corresponding factor loadings were estimated by the principal factors method. In order to obtain more interpretable results, we have rotated the factor loadings in Figure 10.2. After the manual rotation of the factor loadings, the first factor seems to be related to the number of children. The second and the third factor are related to the type of family. The main disadvantage of this approach are that a manual rotation of the factor loadings is rather time consuming and that the final result might be strongly influenced by prior beliefs of the data analyst.

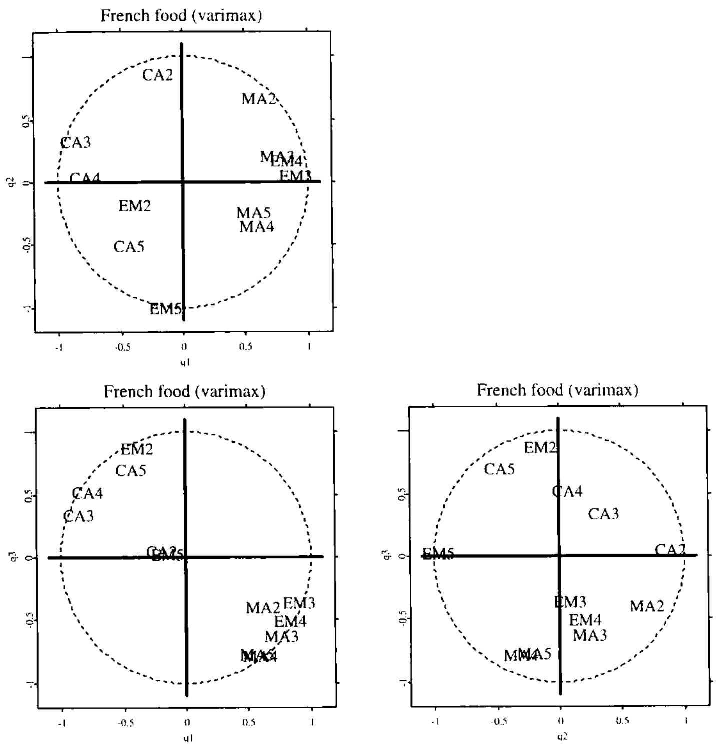

Fig. 10.3. Varimax rotation for French food data set. SMSfactfood

Hence, in practice, we recommend to use the varimax rotation which in this case leads to very similar result, see Figure 10.3. A comparison of Figures 10.2 and 10.3 shows that the varimax methods find automatically a rotation which is very similar to the result obtain by manual rotation of factor loadings. The main difference seems to be the order and the signs of the factors.

EXERCISE 10.5. *Perform a factor analysis on the variables X_4 to X_{10} in the U.S. health data set in Table A.19. Would it make sense to use all of the variables for the factor analysis?*

From the discussion of the degrees of freedom of the factor analysis model in Exercises 10.1 and 10.2 it follows that we can estimate at most $k = 3$ factors in this 7-dimensional data set. The results of the factor analysis are given in Table 10.2 and Figure 10.4. The factor analysis model was estimated by the maximum likelihood method with varimax rotation.

	Estimated factor loadings			Communalities	Specific variances
	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{h}_j^2$	$\hat{\psi}_{jj} = 1 - \hat{h}_j^2$
1 accident	−0.5628	0.0220	−0.1958	0.3556	0.6448
2 cardiovascular	0.7354	0.1782	0.5955	0.9271	0.0735
3 cancer	0.8381	−0.1166	0.5246	0.9913	0.0087
4 pulmonary	0.1709	−0.0682	0.5476	0.3337	0.6666
5 pneumonia flu	0.0098	0.4338	0.7631	0.7706	0.2252
6 diabetes	0.8046	−0.0488	0.0569	0.6531	0.3477
7 liver	0.1126	−0.8082	0.3321	0.7762	0.2173

Table 10.2. Estimated factor loadings after varimax rotation, communalities, and specific variances, MLM, U.S. health data set. SMSfactushealth

Table 10.2 shows that the three factor model explains very well most of the original variables. Only variables accident and pulmonary have lower communalities.

The plots of the factor loadings in Figure 10.4 suggests that the first factor corresponds to causes of death related by cardiovascular problems, cancer, and diabetes. The second factor seems to be positively related to pneumonia flu and negatively related to liver. The third factor combines all causes of death apart of accidents and diabetes. The discussion of the meaning of the factors will be continued in Exercise 10.7, where we present the estimation of the corresponding factor scores for each state.

Let us now investigate the question whether the three factors derived in this exercise describe sufficiently the dependencies within the U.S. health data set. This question can be answered by formal statistical test based on the likelihood ratio approach that has been demonstrated in Chapter 7.

Assuming that $\widehat{\mathcal{Q}}$ and $\widehat{\Psi}$ are the estimates obtained by the maximum likelihood method, the likelihood ratio (LR) test statistic for the null hypothesis H_0 : $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$ can be derived as:

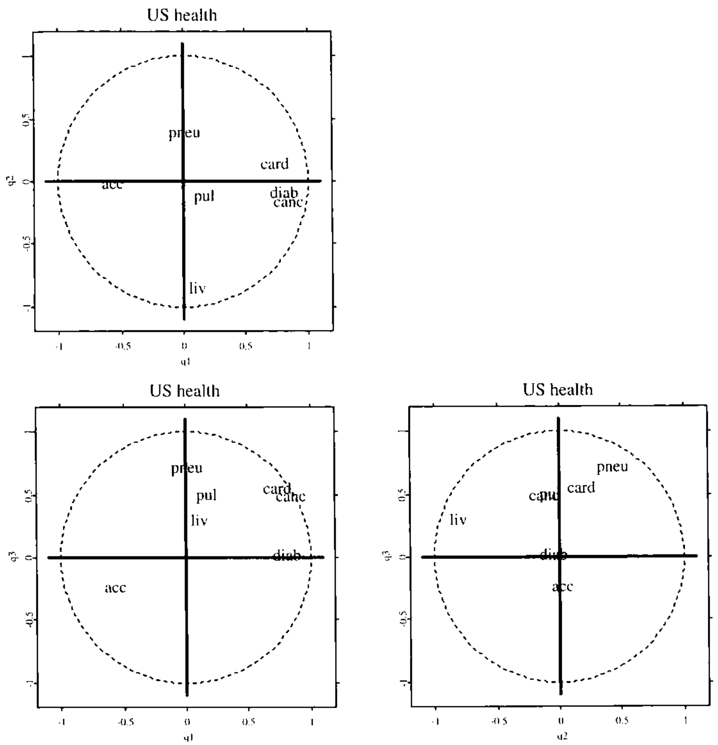

Fig. 10.4. Factor loadings for the U.S. health data set after varimax rotation. SMSfactushealth

$$-2\log\left(\frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood}}\right) = n\log\left(\frac{|\widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top + \widehat{\Psi}|}{|\mathcal{S}|}\right). \qquad (10.2)$$

Under the null hypothesis, the LR test statistic has asymptotically the $\chi^2_{\frac{1}{2}\{(p-k)^2-p-k\}}$ distribution. Bartlett (1954) suggested a correction which improves the above χ^2 approximation by replacing n by $n - 1 - (2p + 4k + 5)/6$ in (10.2). The LR test can be applied only if the degrees of freedom are positive, see also the discussion of the degrees of freedom in Exercise 10.1.

Let us now test the null hypothesis $H_0 : k = 3$. The value of the LR test statistic with Bartlett correction is 3.66 and we cannot reject the null hypothesis $H_0 : \Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$ since the observed value of the test statistic is smaller than the critical value $\chi^2_{0.95;3} = 7.81$. It seems that the factor analysis model with $k = 3$ factors is appropriate for the U.S. health data set.

EXERCISE 10.6. *Perform a factor analysis on the U.S. crime data set in Table A.18 and estimate the factor scores.*

The U.S. crime data set states the reported number of crimes in the 50 states of the USA classified according to 7 categories. Hence, at most $k = 3$ factors can be considered for the factor analysis.

The factor loadings presented in Table 10.3 and plotted in Figure 10.5 were obtained by the maximum likelihood method and varimax rotation.

	Estimated factor loadings			Communalities	Specific variances
	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{h}_j^2$	$\hat{\psi}_{jj} = 1 - \hat{h}_j^2$
1 murder	0.4134	−0.7762	−0.0651	0.7777	0.2225
2 rape	0.7938	−0.2438	−0.0006	0.6895	0.3108
3 robbery	0.6148	−0.1866	0.4494	0.6147	0.3855
4 assault	0.6668	−0.6940	0.0368	0.9275	0.0723
5 burglary	0.8847	0.1073	0.2302	0.8472	0.1534
6 larceny	0.8753	0.3834	−0.0625	0.9172	0.0808
7 auto theft	0.6132	0.1435	0.5995	0.7561	0.2432

Table 10.3. Estimated factor loadings after varimax rotation, communalities, and specific variances, MLM, U.S. crime data set. SMSfactuscrime

The LR test of the hypothesis that three factors are enough to described the dependencies within the U.S. crime data set leads p-value 0.8257 and the null hypothesis $H_0 : k = 3$ cannot be rejected.

The first factor could be described as the overall criminality factor. The second factor is positively related to larceny and negatively related to more violent crimes such as murder and assault. The third factor is related mainly to robbery and auto theft.

In order to describe the differences between different states, we have to estimate the values of the factor scores for individual observations. The idea of the commonly used regression method is based on the joint distribution of $(X - \mu)$ and F. The joint covariance matrix of $(X - \mu)$ and F is:

$$Var\begin{pmatrix} X - \mu \\ F \end{pmatrix} = \begin{pmatrix} \mathcal{Q}\mathcal{Q}^\top + \Psi & \mathcal{Q} \\ \mathcal{Q}^\top & \mathcal{I}_k \end{pmatrix} = \begin{pmatrix} \Sigma & \mathcal{Q} \\ \mathcal{Q}^\top & \mathcal{I}_k \end{pmatrix}. \tag{10.3}$$

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

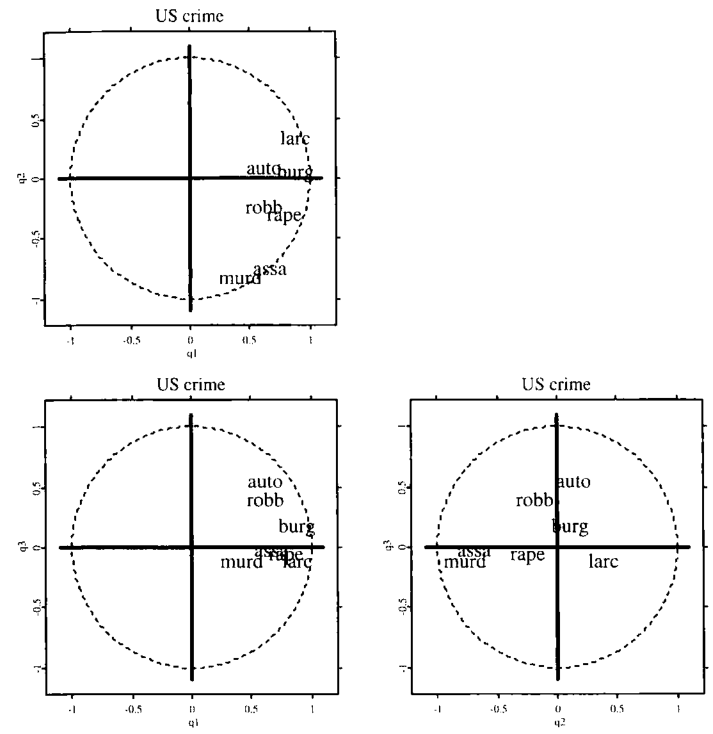

Fig. 10.5. Factor loadings for the U.S. crime data set after varimax rotation. SMSfactuscrime

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

The same rule can be followed when using $\mathcal{R}$ instead of $\mathcal{S}$. Then (10.3) remains valid when standardized variables, i.e., $Z = \mathcal{D}_\Sigma^{-1/2}(X - \mu)$, are considered if $\mathcal{D}_\Sigma = \text{diag}(\sigma_{11}, \dots, \sigma_{pp})$. In this case the factors are given by

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

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

The factor scores corresponding to the factor loadings given in Table 10.3 are plotted in Figure 10.6. The estimated factor scores for the first factor,

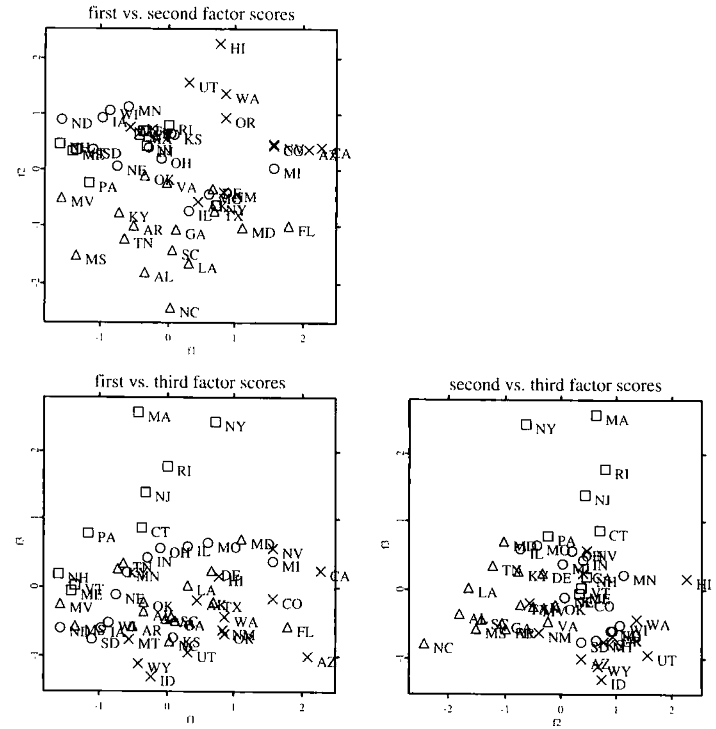

Fig. 10.6. Factor scores for the U.S. crime data set estimated by the regression method. Northeast (squares), Midwest (circles), South (triangles) and West (crosses). Q `SMSfactuscrime`

overall criminality, seem to be largest in California, Arizona, and Florida. The second factor suggests that murder and assault are common mainly in North Carolina. The third factor, auto theft and robbery, reaches the highest estimated factor scores in Massachusetts and New York.

EXERCISE 10.7. *Estimate the factor scores for the U.S. health data set analyzed in Exercise 10.5 and compare the estimated factor scores to the scores obtained for the U.S. crime data set in Exercise 10.6.*

The factor scores for the U.S. health data set, corresponding to the factor loadings obtained in Exercise 10.5, are plotted in Figure 10.7.

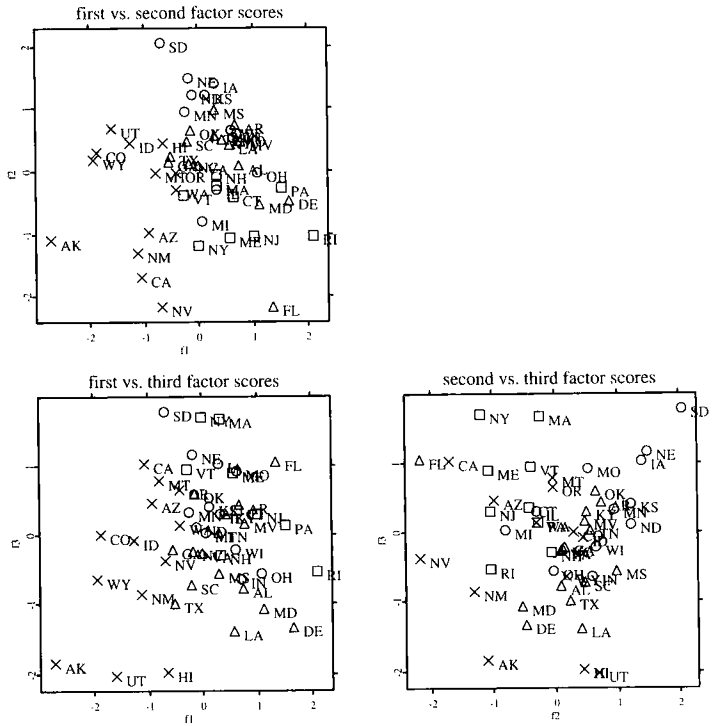

Fig. 10.7. Factor scores for the U.S. health data set estimated by the regression method. Northeast (squares), Midwest (circles), South (triangles) and West (crosses). Q `SMSfactushealth`

The first factor, corresponding to diabetes, cancer, and cardiovascular problems, leads to higher factor scores in Richmond, Delaware, and Pennsylvania. On the other side, these causes of death are less common mainly in Arkansas, Wyoming, Colorado, and Utah. This factor looks a bit like the third factor obtained for the U.S. crime data set in Exercise 10.6.

The second health factor, strongly negatively related to liver and positively related to pneumonia flu has highest values in South Dakota and Nebraska and smallest values in Nevada, Florida, and California. The third health factor has high values in South Dakota, New York, and Massachusetts and small values in Utah, Hawaii, and Arkansas.

Apart of the partial similarity of the first health and third crime factor, there does not seem to be any other relation between the health and crime factors. However, in both plots of the factor scores, we obtain similar factor scores for states coming from the same region.

The factor analysis is not designed to investigate the similarities between two sets of variables. Such comparisons ought to be carried out by the method of canonical correlations described in Chapter 14.

EXERCISE 10.8. *Analyze the vocabulary data given in Table A.20.*

The vocabulary data set contains test scores of 64 pupils from the eighth through eleventh grade levels. For each pupil we have one test score per grade which leads to a 4-dimensional data set. Recalling the considerations presented in Exercises 10.1 and 10.2, we see that in this exercise we can estimate only one factor.

Performing the LR test (10.2) of the hypothesis $H_0 : k = 1$, we obtain the value of the LR test statistic 1.6101, which is smaller than the corresponding critical value $\chi^2_{0.95;2} = 5.9915$ (p-value 0.4470). Hence, one factor seems to be appropriate for the factor analysis of this 4-dimensional data set.

	Estimated factor loadings	Communalities	Specific variances
	$\hat{q}_1$	$\hat{h}_j^2$	$\hat{\psi}_{jj} = 1 - \hat{h}_j^2$
1 Grade 8	0.9284	0.8620	0.1380
2 Grade 9	0.8611	0.7415	0.2585
3 Grade 10	0.9306	0.8659	0.1341
4 Grade 11	0.8618	0.7427	0.2573

Table 10.4. Estimated factor loadings, communalities, and specific variances, MLM, vocabulary data set. Q `SMSfactvocab`

The results obtained by maximum likelihood method are summarized in Table 10.4. The rotation on the one-dimensional factor loadings would not have any meaning. The resulting factor can be interpreted as an overall vocabulary score strongly positively related to the test score in all four grades.

The estimated one-dimensional factor scores are plotted in Figure 10.8 by means of a dot-plot. The position of each observation on the horizontal axis is given by the estimated factor score. The values on the vertical axis are randomly chosen so that the plotted numbers are readable. The best values were achieved in observations 36 and 38 whereas the 5th observation seems to be extremely bad.

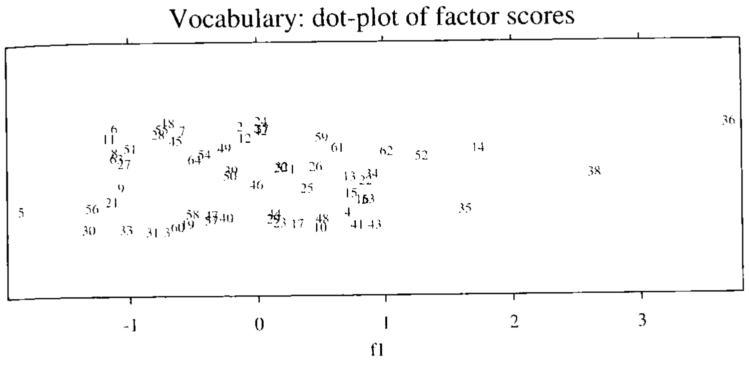

Fig. 10.8. Dot-plot of the one-dimensional factor scores for the vocabulary data set estimated by the regression method. Q `SMSfactvocab`

EXERCISE 10.9. *Analyze the athletic records data set in Table A.1. Can you recognize any patterns if you sort the countries according to the estimates of the factor scores?*

The athletic records data set provides data on athletic records in 100m up to a marathon for 55 countries.

Performing the estimation of the factor loadings by the maximum likelihood method allows us to test the hypothesis $H_0 : k = 3$ by means of the likelihood ratio test statistic (10.2). In this exercise, we obtain the test statistic 7.5207 which is smaller than the critical value $\chi^2_{0.95;7} = 14.0671$. The p-value of the test is 0.3767. The hypothesis that 3 factors are enough to describe the athletic records data set thus cannot be rejected.

The estimated factor loadings obtained by maximum likelihood method and varimax rotation are given in Table 10.5 and plotted in Figure 10.9. The communalities and specific variances show that three factors explain very well all of the original variables up to the record in 200m.

The first factor is most strongly related to times achieved in 100 and 200m, the second factor is positively related mainly to the records in longer distances. The third factor has positive relationship to the records in middle distances and 100m. It is important to keep in mind that high numbers here correspond to worse times. Hence, the athletic nations should exhibit small values of the factor scores.

	Estimated factor loadings			Communalities	Specific variances
	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{h}_j^2$	$\hat{\psi}_{jj} = 1 - \hat{h}_j^2$
1 100 m	0.7642	0.1803	0.6192	1.0000	0.0000
2 200 m	0.5734	0.0711	0.0474	0.3361	0.6642
3 400 m	0.4617	0.4869	0.6468	0.8686	0.1315
4 800 m	0.3442	0.6530	0.6060	0.9120	0.0878
5 1.5 km	0.3391	0.7655	0.4894	0.9404	0.0596
6 5 km	0.3771	0.8612	0.2842	0.9647	0.0354
7 10 km	0.4022	0.8636	0.2768	0.9842	0.0157
8 marathon	0.3231	0.8813	0.1843	0.9151	0.0850

Table 10.5. Estimated factor loadings after varimax rotation, communalities, and specific variances, MLM, athletic records data set. SMSfacthletic

Rank	1	2	3
1	Italy	Portugal	GB
2	Colombia	NZ	Bermuda
3	USA	Ireland	DomRep
4	USSR	Netherlands	Thailand
5	Canada	Kenya	USA
6	Poland	Norway	FRG
⋮	⋮	⋮	⋮
50	Kenya	Bermuda	Colombia
51	PKorea	Malaysia	PNG
52	Netherlands	Singapore	WSamoa
53	Philippines	DomRep	Guatemala
54	Mauritius	Thailand	CostaRica
55	CookIs	WSamoa	CookIs

Table 10.6. Countries sorted according to the factor scores estimated for the athletic records data set. SMSfacthletic

The factor scores estimated by the regression method are plotted in Figure 10.10. Furthermore, Table 10.6 lists the best and the worst countries according to each factor.

Keeping in mind the interpretation of the factors, we can say that Italy, Colombia, USA, USSR, Canada, and Poland possess the best sprinters. On long distances, the best countries are Portugal, New Zealand, Ireland, Netherlands, and Kenya. The best times on 100m, 400m, and 800m are on average achieved by Great Britain, Bermuda, Dominican Republic, Thailand, and USA.

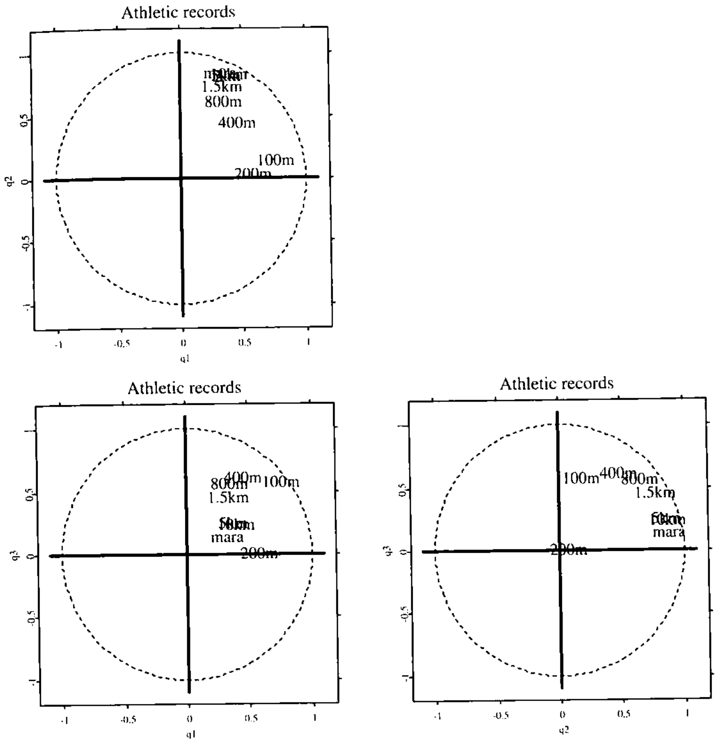

Fig. 10.9. Factor loadings for the athletic records data set after varimax rotation. SMSfacthletic

It is also interesting to notice that some of the countries which have very good factor scores for third or second factor, have, at the same time, very bad first or second factor scores. See, for example, Dominican Republic, Netherlands, and Kenya.

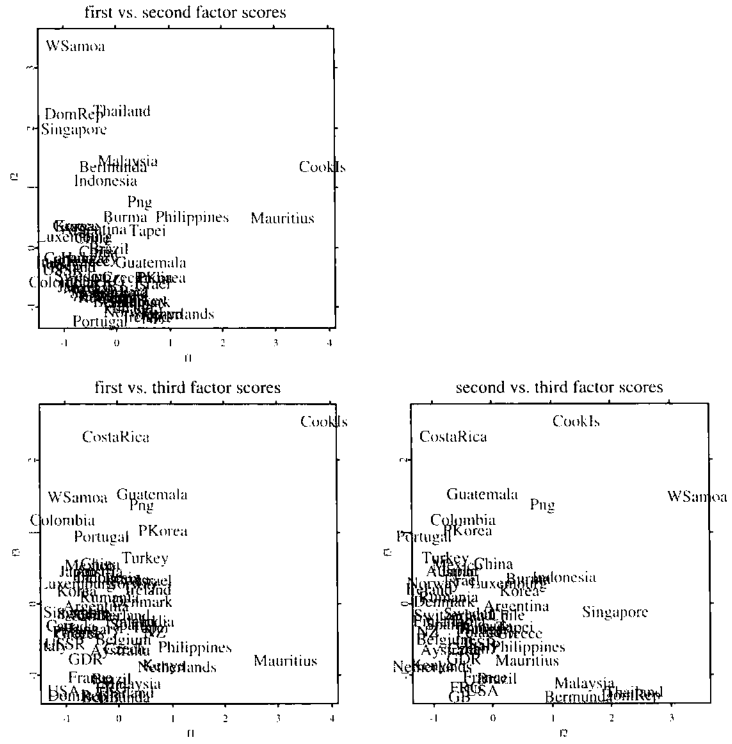

Fig. 10.10. Factor scores for the athletic records data set estimated by the regression method. Q SMSfacthletic

11

Cluster Analysis

From a drop of water, a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard of one or the other. So all life is a great chain, the nature of which is known whenever we are shown a single link of it.
Sherlock Holmes in "Study in Scarlet"

When considering groups of objects in a multivariate data set, two situations can arise. Given a data set containing measurements on individuals, in some cases we want to see if some natural groups or classes of individuals exist, and in other cases, we want to classify the individuals according to a set of existing groups. Cluster analysis develops tools and methods concerning the former case, that is, given a data matrix containing multivariate measurements on a large number of individuals (or objects), the objective is to build subgroups or clusters of individuals. This is done by grouping individuals that are "similar" according to some appropriate criterion.

Cluster analysis is applied in many fields, including the natural sciences, the medical sciences, economics, and marketing. In marketing, for instance, it is useful to build and describe the different segments of a market from a survey of potential consumers. An insurance company, on the other hand, might be interested in the distinction among classes of potential customers so that it can derive optimal prices for its services. Other examples are provided in this chapter.

In this chapter we will concentrate on the so-called agglomerative hierarchical algorithms. The clustering algorithms start by calculating the distances between all pairs of observations, followed by stepwise agglomeration of close observations into groups.

Agglomerative Algorithm

1. Compute the distance matrix $\mathcal{D} = (d_{ij})_{i,j=1,\ldots,n}$.
2. Find two observations with the smallest distance and put them into one cluster.
3. Compute the distance matrix between the $n-1$ clusters.
4. Find two clusters with the smallest intercluster distance and join them.
5. Repeat step 4 until all observations are combined in one cluster.

The properties of the clustering algorithm are driven mainly by the choice of distance.

Intercluster Distance

Assume that two observations or clusters, P and Q, are combined in a cluster denoted by $P \cup Q$. Let $d(P,Q)$ denote the distance between clusters P and Q and n_P and n_Q the number of observations belonging to clusters P and Q, respectively. Some common methods for defining the distance between the cluster $P \cup Q$ and some other cluster, say R, are:

Single linkage: $d(P \cup Q, R) = \min\{d(P,R), d(Q,R)\}$.

Complete linkage: $d(P \cup Q, R) = \max\{d(P,R), d(Q,R)\}$.

Average linkage: $d(P \cup Q, R) = \{d(P,R) + d(Q,R)\}/2$.

Average linkage (weighted):

$$d(P \cup Q, R) = \{n_P d(P,R) + n_Q d(Q,R)\}/(n_P + n_Q).$$

Median: $d^2(P \cup Q, R) = \{d^2(P,R) + d^2(Q,R)\}/2 - d^2(P,Q)/4$.

Centroid: $d^2(P \cup Q, R)$ is defined as the squared distance between R and the weighted (coordinatewise) average of P and Q; see Exercise 11.1.

Ward method: the heterogeneity of group R is measured by the inertia $I_R = \sum_{i=1}^{n_R} d^2(x_i, \overline{x}_R)$ (Ward 1963). In each step, we join the groups P and Q that give the smallest increase, $\Delta(P,Q)$, of the overall inertia; see Exercises 11.2 and 11.3.

Dendrogram

The successive joining of observations to the clusters is finally plotted in the so-called dendrogram. The construction of the dendrogram is explained in detail in Exercise 11.4.

EXERCISE 11.1. *Prove that the centroid distance $d^2(R, P \cup Q)$, defined as the (squared) distance between $R = (r_1, \ldots, r_p)^\top$ and the weighted average $\{n_P(p_1,\ldots,p_p)^\top + n_Q(q_1,\ldots,q_p)^\top\}/(n_P+n_Q)$ of P and Q, can be calculated as*

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

Let us calculate the Euclidean distance between the center $(r_1,\ldots,r_p)^\top$ of the cluster R and the weighted "center of gravity" of clusters P and Q:

$$\begin{aligned}
&d^2(P \cup Q, R)\\
&= \sum_{i=1}^{p}\left\{r_i - \frac{p_i n_P + q_i n_Q}{n_Q + n_P}\right\}^2\\
&= \sum_{i=1}^{p}\left[r_i^2 - 2r_i\frac{p_i n_P + q_i n_Q}{n_Q + n_P} + \left\{\frac{p_i n_P + q_i n_Q}{n_Q + n_P}\right\}^2\right]\\
&= \sum_{i=1}^{p}\left[\frac{n_P}{n_P+n_Q}(r_i - p_i)^2 + \frac{n_Q}{n_P+n_Q}(r_i - q_i)^2 - \frac{n_P n_Q}{(n_P+n_Q)^2}(q_i - p_i)^2\right]\\
&= \frac{n_P}{n_P+n_Q}\,d^2(R,P) + \frac{n_Q}{n_P+n_Q}\,d^2(R,Q) - \frac{n_P n_Q}{(n_P+n_Q)^2}\,d^2(P,Q).
\end{aligned}$$

Hence, the intercluster distance between R and $P \cup Q$ can be calculated from the distance between R, P, and Q. This property greatly simplifies the software implementation of the clustering algorithm since all calculations can be carried out using only the distance matrix between the n observations.

EXERCISE 11.2. *Derive the formula for the increase of the inertia $\Delta(P,Q)$ in the Ward method.*

In the Ward method, the heterogeneity of group R is measured by the inertia defined as:

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

where $\overline{x}_R$ is the arithmetic average and n_R the number of observations within group R. If the usual Euclidean distance is used, then I_R represents the sum of the variances of the p components of x_i inside group R, see Exercise 11.3.

The Ward algorithm joins the groups P and Q that give the smallest increase, $\Delta(P,Q)$, of the inertia. The common inertia of the new group $P \cup Q$ can be written as:

$$
\begin{aligned}
I_{P\cup Q} &= \sum_{i=1}^{n_P+n_Q} d^2(x_i - \overline{x}_{P\cup Q}) = \sum_{i=1}^{n_P+n_Q}\sum_{j=1}^{p}(x_{ij} - \overline{x}_{P\cup Q,j})^2 \\
&= \sum_{j=1}^{p}\left\{\sum_{i=1}^{n_P}(x_{P,ij} - \overline{x}_{P\cup Q,j})^2 + \sum_{i=1}^{n_Q}(x_{Q,ij} - \overline{x}_{P\cup Q,j})^2\right\} \\
&= \sum_{j=1}^{p}\left\{\sum_{i=1}^{n_P}(x_{P,ij} - \overline{x}_{P,j})^2 + n_P(\overline{x}_{P,j} - \overline{x}_{P\cup Q,j})^2\right. \\
&\qquad \left. + \sum_{i=1}^{n_Q}(x_{Q,ij} - \overline{x}_{Q,j})^2 + n_Q(\overline{x}_{Q,j} - \overline{x}_{P\cup Q,j})^2\right\} \\
&= I_P + I_Q + \sum_{j=1}^{p}\left\{n_P(\overline{x}_{P,j} - \overline{x}_{P\cup Q,j})^2 + n_Q(\overline{x}_{Q,j} - \overline{x}_{P\cup Q,j})^2\right\}
\end{aligned}
$$

Hence, the inertia of $P \cup Q$ can be split into the sum of I_P and I_Q and a remainder term $\Delta(P,Q)$ for which we have:

$$
\begin{aligned}
\Delta(P,Q) &= \sum_{j=1}^{p}\left\{n_P(\overline{x}_{P,j} - \overline{x}_{P\cup Q,j})^2 + n_Q(\overline{x}_{Q,j} - \overline{x}_{P\cup Q,j})^2\right\} \\
&= \sum_{j=1}^{p}\left\{n_P\left(\frac{n_Q\overline{x}_{P,j} - n_Q\overline{x}_{Q,j}}{n_P+n_Q}\right)^2 + n_Q\left(\frac{n_P\overline{x}_{P,j} - n_P\overline{x}_{Q,j}}{n_P+n_Q}\right)^2\right\} \\
&= \frac{n_P n_Q}{n_P+n_Q}\sum_{j=1}^{p}(\overline{x}_{P,j} - \overline{x}_{Q,j})^2 = \frac{n_P n_Q}{n_P+n_Q} d^2(P,Q).
\end{aligned}
$$

The change of inertia $\Delta(P,Q)$ resulting from the joining of the groups P and Q can be considered as a distance of the clusters P and Q. In order to implement the Ward method numerically, we have to derive a formula for the intercluster distance between cluster R and the newly created cluster $P \cup Q$.

Applying the result of Exercise 11.1, we can write:

$$
\begin{aligned}
\Delta(R, P\cup Q) &= \frac{n_R(n_P+n_Q)}{n_R+n_P+n_Q} d^2(R, P\cup Q) \\
&= \frac{n_R(n_P+n_Q)}{n_R+n_P+n_Q}\left\{\frac{n_P}{n_P+n_Q}\, d^2(R,P) + \frac{n_Q}{n_P+n_Q}\, d^2(R,Q)\right. \\
&\qquad \left. - \frac{n_P n_Q}{(n_P+n_Q)^2}\, d^2(P,Q)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n_R+n_P+n_Q}\left\{n_R n_P\, d^2(R,P) + n_R n_Q\, d^2(R,Q)\right. \\
&\qquad \left. - \frac{n_R n_P n_Q}{n_P+n_Q}\, d^2(P,Q)\right\} \\
&= \frac{n_R+n_P}{n_R+n_P+n_Q}\, \Delta(R,P) + \frac{n_R+n_Q}{n_R+n_P+n_Q}\, \Delta(R,Q) \\
&\qquad - \frac{n_R}{n_R+n_P+n_Q}\, \Delta(P,Q).
\end{aligned}
$$

The ability to express $\Delta(R, P\cup Q)$ using the distances $\Delta(R,P)$, $\Delta(R,Q)$, and $\Delta(P,Q)$ greatly simplifies the computer implementation of the Ward algorithm.

EXERCISE 11.3. *Prove that in the Ward method, the inertia $I_R = n_R\, tr(\mathcal{S}_R)$, where $\mathcal{S}_R$ denotes the empirical covariance matrix of the observations contained in group R.*

The inertia is defined as:

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

Assuming that $d(x_i, \overline{x}_R)$ is the usual Euclidean distance between the ith observation $x_i = (x_{i1}, \dots, x_{ip})^\top$ and the sample mean within group R, $\overline{x}_R = (x_{R1}, \dots, x_{Rp})^\top$, we have:

$$
\begin{aligned}
I_R &= \sum_{i=1}^{n_R} d^2(x_i, \overline{x}_R) = \sum_{i=1}^{n_R}\sum_{j=1}^{p}(x_{ij} - \overline{x}_{Rj})^2 \\
&= n_R\sum_{j=1}^{p}\frac{1}{n_R}\sum_{i=1}^{n_R}(x_{ij} - \overline{x}_{Rj})^2 = n_R\sum_{j=1}^{p} s_{X_jX_j} = n_R \operatorname{tr} \mathcal{S}_R.
\end{aligned}
$$

EXERCISE 11.4. *Explain the differences between various proximity measures by means of the 8 points example given in Härdle & Simar (2003, example 11.5).*

The eight points from Example 11.5 in Härdle & Simar (2003) are plotted in Figure 11.1. Selected distances between some of the points are marked by lines. Different proximity measures assign different values to these interpoint distances. It is clear that the choice of the proximity measure can influence the behavior of the clustering algorithm.

In Figure 11.2, we plot the dendrograms obtained for the eight points example using two different simple distances. In both dendrograms, we can see how

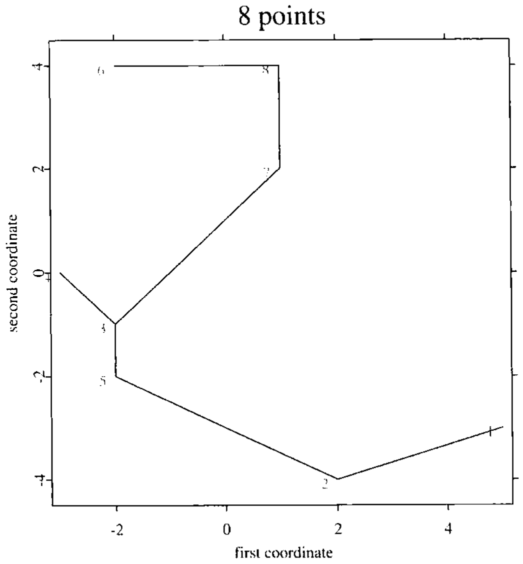

Fig. 11.1. 8 points example using single linkage. Q SMSclus8pd

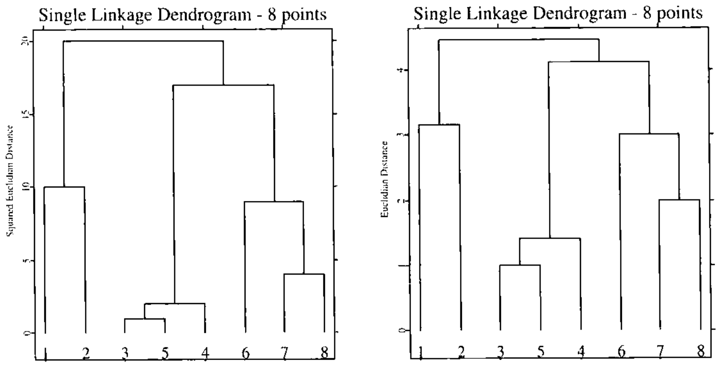

Fig. 11.2. Single linkage using squared Euclidean and Euclidean distance. Q SMSclus8pd

the n points were consecutively joined into only one cluster. The intercluster distances are given on the vertical axis. In both plots in Figure 11.2 we can see that in the first step of the algorithm, the points 3 and 5 were combined. Both the Euclidean and squared Euclidean distance between these points is equal to 1, see also Figure 11.1.

The distance in the right plot of Figure 11.2 is equal to the square root of the distance in the left plot. Thanks to the single linkage algorithm which defines the intercluster distance as the distance between closest points, we obtain exactly the same clustering in both plots. The only difference is the change of scale on the vertical axis.

The last step in cluster analysis is the choice of a number of cluster. For example, three clusters in the 8 points example can be obtained by cutting the dendrogram given in Figure 11.2 at a specified level. In this case, we would obtain clusters $\{1,2\}$, $\{3,4,5\}$, and $\{6,7,8\}$.

EXERCISE 11.5. *Repeat the 8 point example (Exercise 11.4) using the complete linkage and the Ward algorithm. Explain the difference to single linkage.*

The dendrograms obtained by complete linkage and Ward method are plotted on the right hand side in Figures 11.3 and 11.4. The left plots contain the original points with lines describing the successive joining of the clusters.

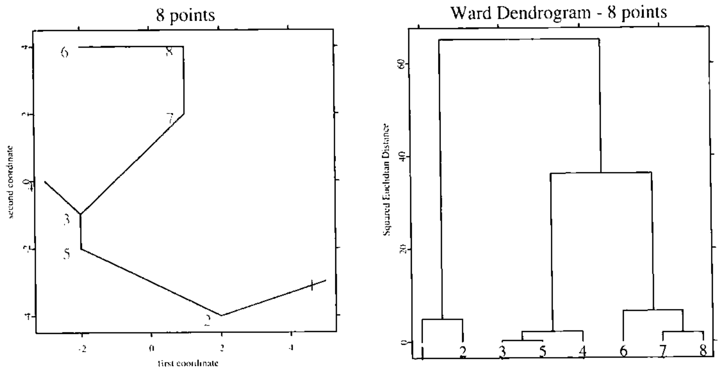

Fig. 11.3. Ward algorithm. Q SMSclus8p

The lines plotted in Figure 11.4 demonstrate how the intercluster distances are calculated in the complete linkage. For example, the line connecting points

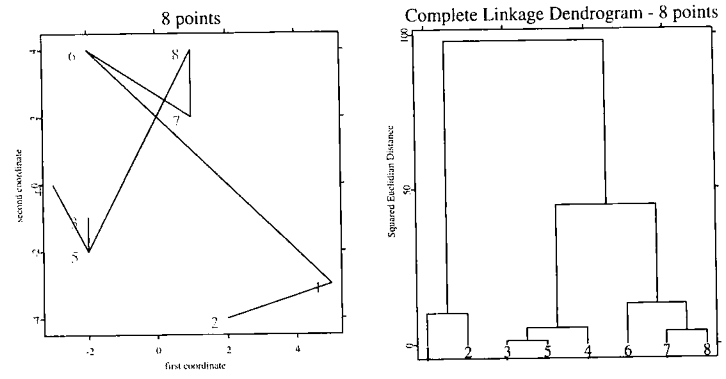

Fig. 11.4. Complete linkage. SMSclus8p

5 and 8 gives the distance between the clusters consisting of points {3,4,5} and {6,7,8}. In the single linkage method used in Exercise 11.4, the distance between these clusters would be given by the distance of the closest points, i.e., by the distance of points 3 and 7.

Comparing the dendrograms in Figures 11.2–11.4, we see that, in this example, the three clustering algorithms arrive to the same result. The only difference lies in the scale on the vertical axis. Both the Ward algorithm in Figure 11.3 and the complete linkage in Figure 11.4 strongly suggest that the choice of three clusters might be appropriate in this case. The intercluster distances between the same three clusters are relatively smaller if single linkage is used.

In practice, the Ward algorithm usually provides the best interpretable results since it tends to create "homogeneous" clusters. On the contrary, the single linkage algorithm often finds chains of observations which do not have any other clear structure.

EXERCISE 11.6. *Perform a cluster analysis for 20 randomly selected Swiss bank notes in Table A.2.*

Recall that the data set contains 200 6-dimensional observations. The first 100 observations correspond to genuine and the other half to counterfeit bank notes. Here, we use only a subsample of size 20 so that the resulting dendrogram in Figure 11.5 is still readable. On the left plot in Figure 11.5 we plot the first two principal components for the data set. From Chapter 9 we know that this is, in some sense, the best two-dimensional representation of the data set. One can observe that the plot consists of two point clouds: on the left hand side, we have the genuine bank notes with numbers smaller than 100 and, on the right hand side, we observe point cloud of the counterfeit bank notes. The observation 161 is a bit separated from both these groups.

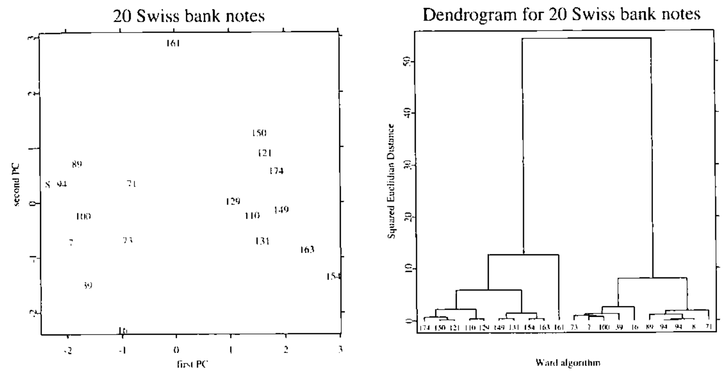

Fig. 11.5. Cluster analysis of 20 Swiss bank notes using Ward algorithm and squared Euclidean distance. SMSclusbank

The dendrogram, resulting from the Ward algorithm using the squared Euclidean distance, is plotted on the right hand side of Figure 11.5. If the dendrogram is cut to two clusters, we obtain exactly the genuine and counterfeit bank notes. The outlying observation 161 was correctly put into the counterfeit cluster but the dendrogram shows that the distance from the other counterfeit bank notes is largest from all (counterfeit) observations.

The dendrograms obtained by the single and complete linkage clustering algorithms are given in Exercise 11.7.

EXERCISE 11.7. *Repeat the cluster analysis of the bank notes example in Exercise 11.6 with single and complete linkage clustering algorithms.*

The dendrograms for both the single and complete linkage are plotted in Figures 11.6 and 11.7. The complete linkage plotted in Figure 11.6 provides better result since it correctly puts the observation 161 into the counterfeit group. However, comparing the complete linkage and the dendrogram obtained by the Ward algorithm in Figure 11.5, the Ward distance seems to be more appropriate in this case.

The single linkage dendrogram in Figure 11.7 shows the chain building tendency of this method. The observations are usually added one by one and the result of this method often consists of two clusters: one containing almost all

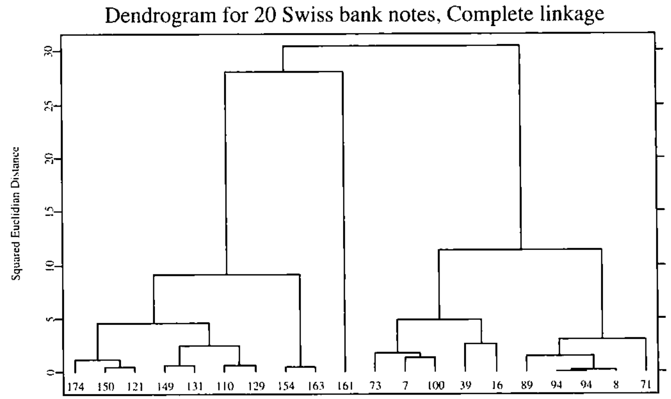

Fig. 11.6. Cluster analysis of 20 Swiss bank notes using squared Euclidean distance with complete linkage. SMSclusbank2

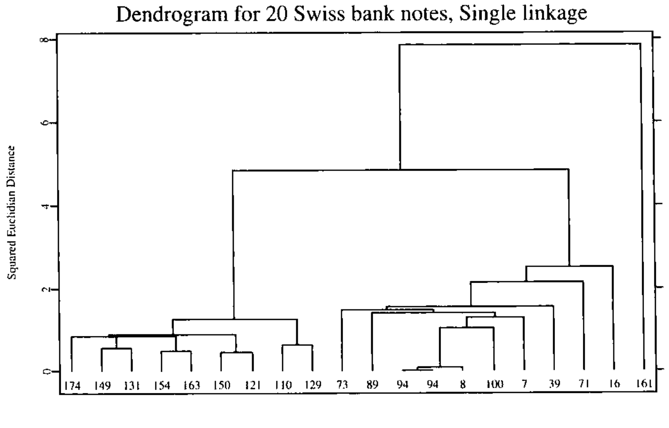

Fig. 11.7. Cluster analysis of 20 Swiss bank notes using squared Euclidean distance with single linkage. SMSclusbank2

observations and the other one or two outliers. This is exactly what happened in Figure 11.7, where the outlying observation 161 was put into a cluster by itself.

EXERCISE 11.8. *Repeat the cluster analysis of the bank notes example in Exercise 11.6 using the L_1 distance.*

The Euclidean distance is just a special case of the L_r-norms, $r \geq 1$,

$$d_{ij} = ||x_i - x_j||_r = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right\}^{1/r}, \tag{11.1}$$

where x_{ik} denotes the value of the kth variable measured on the ith individual.

Apart of the usual Euclidean distance (L_2-norm), the L_1-norm is the most popular member of this family. The L_1 distance has very simple interpretation since from (11.1) it is easy to see that the L_1 distance is just the sum of the absolute values of the differences observed in each variable. The L_1 metric is useful whenever we want to assign less weight to the outlying observations.

In the previous exercises, it appeared that the Ward method leads to nice and interpretable results. Hence, we apply the Ward method with L_1 distance to obtain the dendrogram plotted in Figure 11.8. The same analysis with

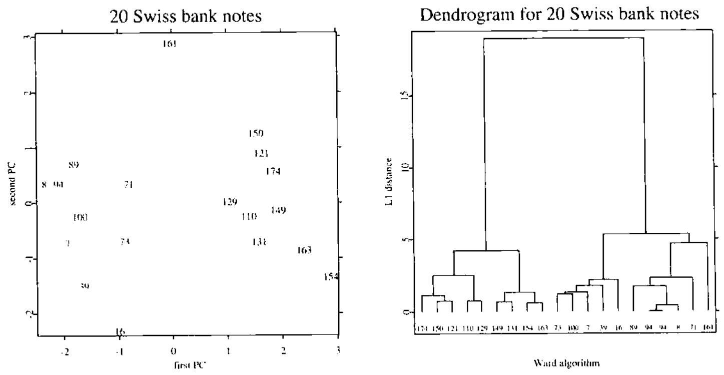

Fig. 11.8. Cluster analysis of 20 Swiss bank notes using Ward algorithm and L_1 distance. SMSclusbank3

the squared Euclidean distance was carried out in Exercise 11.6. Instead of the squared Euclidean distance, we have now selected the L_1 distance which should assign less weight to outlying observations.

The overall shape of the dendrogram plotted in Figure 11.8 looks very similar to the dendrogram given in Figure 11.5. Again, the bank notes are clearly split into two groups. However, in Figure 11.5, the counterfeit observation 161 lies in one cluster with the genuine bank notes.

EXERCISE 11.9. *Analyze the U.S. companies data set in Table A.17 using the Ward algorithm and L_1 distance.*

The six dimensional data set contains the information on the assets, sales, market value, profits, cash flow and number of employees of 79 U.S. companies. The companies are classified according to their type: Communication, Energy, Finance, Hi-Tech, Manufacturing, Medical, Other, Retail, and Transportation.

In Figure 11.9, we plot the first two principal components for a rescaled version of the data set. The rescaling is in this case necessary since otherwise we observe most of the points concentrated in the lower left corner with the two largest companies (IBM and General Electric) dominating the plot. The transformation was used only for plotting in Figures 11.9 and 11.11 and the cluster analysis was performed using the L_1 distances calculated from the original data set.

The transformation which is used on all columns of the data set for plotting is

$$f(x) = \log[x - \min(x) + \{\max(x) - \min(x)\}/200].$$

In this case, the choice of the transformation is quite arbitrary. The only purpose is to plot the observations on a scale that allows us to distinguish different companies in Figures 11.9 and 11.11.

Short inspection of the data set given in Table A.17 reveals that the units of measurements for different variables are not comparable. For example, it would not make much sense to assume that a unit change in the number of employees has the same significance as a unit change in sales or market value. Hence, the cluster analysis is performed on the standardized data set where all variables were divided by their estimated standard deviation.

In Figure 11.10, we display the dendrogram obtained by running the Ward algorithm on the L_1 distances calculated from the standardized U.S. companies data set. From the graphics, it looks reasonable to split the data set into 3 or 5 clusters. In Figure 11.10, we give also the first two letter of the type of the company. It is interesting that in Figure 11.10, the same types of company are often close to each other. See, for example, the large groups of financial or energy companies. However, if we choose lower number of cluster, these groups are mixed with other types of companies.

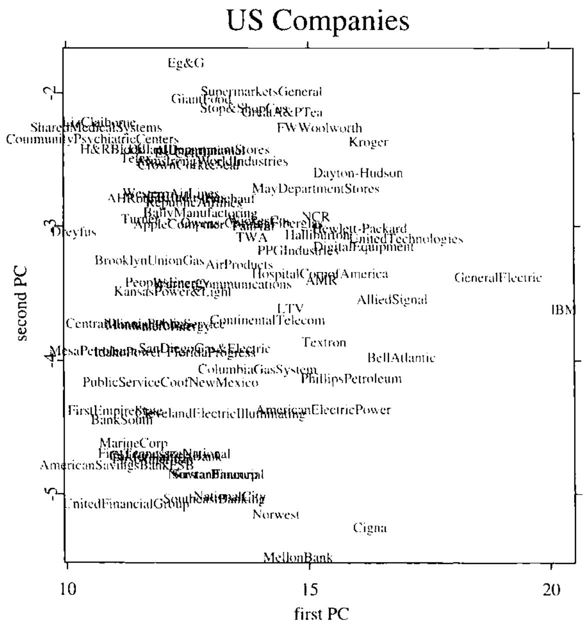

Fig. 11.9. Plot of the first two principal components for the rescaled U.S. companies data set. SMScluscomp

The resulting five clusters are plotted in Figure 11.11 where different plotting symbols were used for each cluster. The type of each company is also specified by the first two letters. Two hi-tech companies form a cluster by themselves: IBM and General Electric. In the upper part of Figure 11.11, we can observe a large group of retail companies. Unfortunately, the Ward algorithm puts this group into two different clusters. The same could be said for the group of financial companies visible in the lower left part of Figure 11.11.

The cluster analysis could be summarized in the following way: the clusters seem to split the data set mainly in the direction of the first principal component which seems to be related mainly to the size of the company. Hence, the clustering algorithm does not recover the type of company which seems to be better explained by the (less important) second principal component.

An improvement in clustering might be achieved also by transforming the data set before calculating the distance matrix used in the clustering algorithm. One possible transformation might be the logarithmic transformation

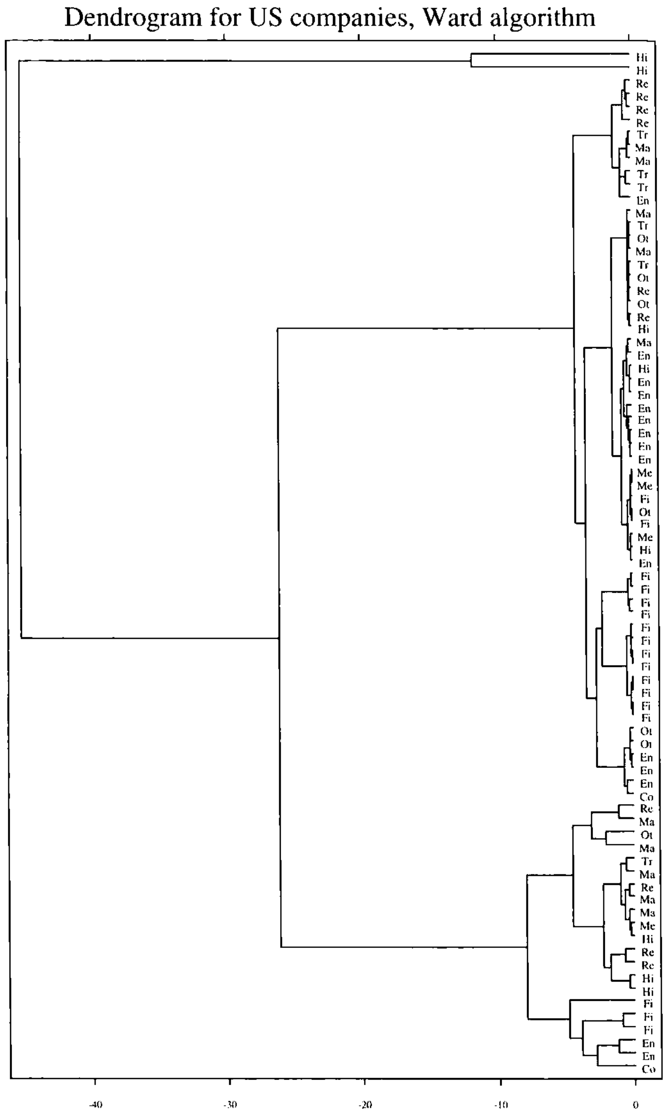

Fig. 11.10. Dendrogram for U.S. companies using Ward algorithm and L_1 distance. SMScluscomp

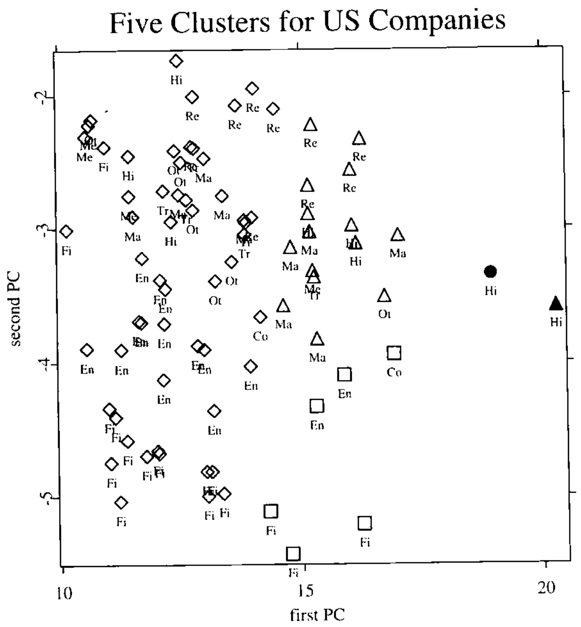

Fig. 11.11. Plot of the first two principal components for the rescaled U.S. companies data set with five clusters denoted by different symbols. SMScluscomp

used for plotting in Figures 11.9 and 11.11 or possibly another transformation correcting for the effect of the size of the company.

EXERCISE 11.10. *Analyze the U.S. crime data set in Table A.18 with the Ward algorithm. Use the χ^2-metric measuring differences between rows of a contingency table and compare the results to the usual L_2-norm on standardized variables.*

The U.S. crime data set contains the reported number of 7 types of crimes in the 50 USA states. The entries in this data set can be interpreted as counts and the data set as a (50 × 7) contingency table.

In a given contingency table, the ith row can be interpreted as the conditional frequency distribution $\frac{x_{ik}}{x_{i\bullet}}$, $k = 1, \ldots, p$, where $x_{i\bullet} = \sum_{j=1}^{p} x_{ij}$. The distance between the ith and jth row can be defined as a χ^2 distance between the respective frequency distributions:

$$d^2(i,j) = \sum_{k=1}^{p} \frac{1}{\left(\frac{x_{\bullet k}}{x_{\bullet\bullet}}\right)} \left(\frac{x_{ik}}{x_{i\bullet}} - \frac{x_{jk}}{x_{j\bullet}} \right)^2 ,$$

see, e.g., Härdle & Simar (2003, section 11.2).

χ^2 Distance

The χ^2 distances between the rows (observations) in the U.S. crime data set are used to construct the distance matrix. The dendrogram plotted in Figure 11.12 was obtained by the Ward method. Each observation displayed in the dendrogram in Figure 11.12 is marked by the abbreviation of the state and by the region number (1=Northeast, 2=Midwest, 3=South, 4=West).

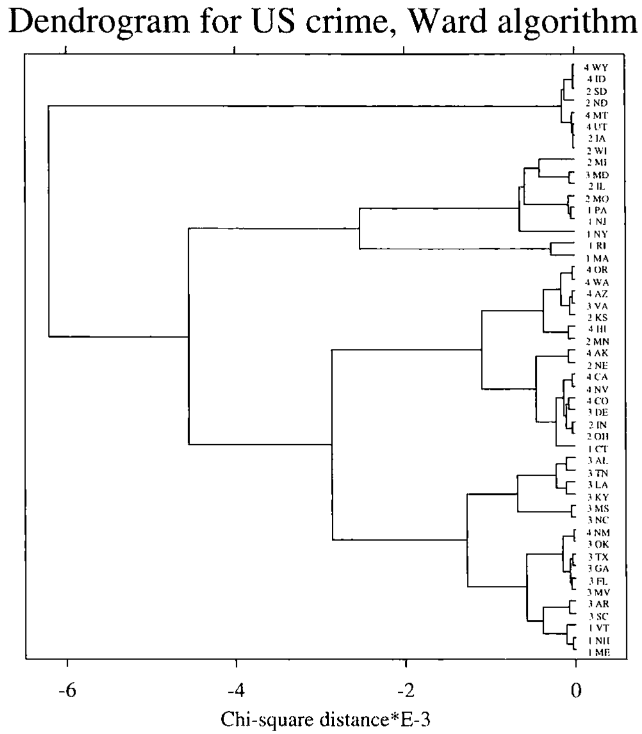

Fig. 11.12. Cluster analysis of U.S. crime data set using Ward algorithm and χ^2 distance. Q `SMScluscrimechi2`

The dendrogram suggests that it would be reasonable to split the data set into 5 or 7 clusters. Let us try to consider 5 clusters and let us define cluster one as ME, NH, VT, MV, NC, SC, GA, FL, KY, TN, AL, MS, AR, LA, OK, TX, and NM. Cluster 2 consists of CT, OH, IN, MN, NE, KS, DE, VA, CO, AZ, NV, WA, OR, CA, AK, and HI, cluster 3 contains MA and RI, cluster 4 NY, NJ, PA, IL, MI, MO, and MD. Cluster 5 is WI, IA, ND, SD, MT, ID, WY, UT. In Table 11.1, we give the average relative frequencies within the five clusters. The information given in Table 11.1 allows us to describe the differences between the clusters. It seems that larceny is very "popular" mainly in cluster 5 consisting mainly of only from West and Midwest states (region code 4). Auto theft is relatively more spread out in cluster 3 consisting only from Massachusetts and Richmond. Cluster 4 (NY, NJ, ...) contains more robberies. Cluster 1, consisting mainly of southern states (region code 3), slightly overrepresents rape and burglaries.

	murder	rape	robbery	assault	burglary	larceny	auto theft
1	0.00	0.01	0.02	0.06	0.30	0.52	0.09
2	0.00	0.00	0.02	0.03	0.26	0.57	0.11
3	0.00	0.00	0.02	0.03	0.27	0.46	0.22
4	0.00	0.00	0.06	0.04	0.27	0.49	0.13
5	0.00	0.00	0.01	0.02	0.21	0.70	0.06

Table 11.1. The average relative frequencies for U.S. crimes within the 5 clusters obtained with χ^2 distance. Q `SMScluscrimechi2`

Euclidean Distance

The results of the Ward algorithm performed on the Euclidean distances between standardized observations are summarized in Figure 11.13 and Table 11.2. Here, we have chosen to consider four clusters.

The first cluster contains the states: ME, NH, VT, PA, WI, IA, ND, SD, NE, MV, MT, ID, and WY. The second cluster is MA, RI, CT, NJ, OH, IN, MN, KS, UT, WA, OR, and HI. The third cluster consists of VA, NC, SC, GA, KY, TN, AL, MS, AR, and OK. The fourth cluster contains NY, IL, MI, MO, DE, MD, FL, LA, TX, CO, NM, AZ, NV, CA, and AK. From the regional point of view, it is interesting to notice that the third cluster contains only southern states.

Table 11.2 allows us to describe the differences between clusters. Cluster 1 contains the states with low criminality since the average of the standardized

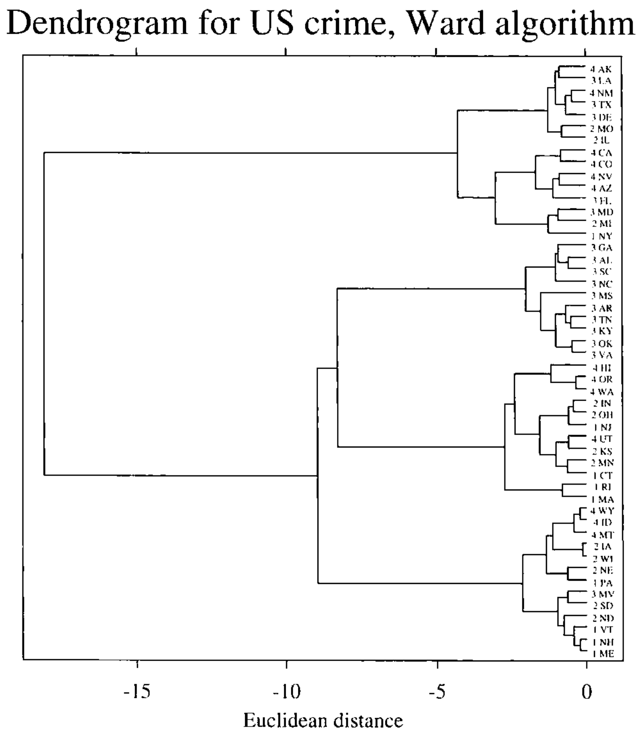

Fig. 11.13. Cluster analysis of U.S. crime data set using Ward algorithm and Euclidean distance. SMScluscrime

	murder	rape	robbery	assault	burglary	larceny	auto theft
1	−0.96	−0.91	−0.80	−1.03	−1.07	−0.70	−0.97
2	−0.72	−0.37	−0.06	−0.63	0.37	0.40	0.62
3	1.06	−0.14	−0.43	0.55	−0.40	−0.82	−0.66
4	0.70	1.18	1.03	1.03	0.91	0.83	0.78

Table 11.2. The averages of the standardized U.S. crime data set within the 3 clusters obtained with Euclidean distance. SMScluscrime

number of all crimes is negative. On the other side, cluster 4 contains the states with high criminality rate. Cluster 2 corresponds to states with a tendency towards burglary, larceny, and auto theft. The souther cluster 3 has large rates of murder and assault.

Comparison

We have seen that each distance leads to another view at the data set. The χ^2 distance compares relative frequencies whereas the Euclidean distance compares the absolute values of the number of each crime. The choice of the method depends in practice mainly on the point of view of the investigator.

EXERCISE 11.11. *Perform the cluster analysis of the U.S. health data set in Table A.19.*

The description of the U.S. health data set is given in Appendix A.19. Basically, it contains the number of deaths in 50 U.S. states classified according to 7 causes of death. We are interested in the numbers of deaths and hence we have decided to perform the analysis using Euclidean analysis on the original data set. The resulting dendrogram is plotted in Figure 11.14.

Cluster 1 contains ME, MA, RI, NY, NJ, PA, IA, MO, SD, NE, MV, FL, and AR. Cluster 2 consists of VT, CT, OH, IN, IL, MI, WI, KS, DE, KY, TN, AL, MS, and OK. Cluster 3 is NH, MN, ND, MD, VA, NC, SC, GA, LA, TX, MT, ID, AZ, NV, WA, OR, and CA and the last cluster 4 consists of WY, CO, NM, UT, AK, and HI. Cluster 4 contains only western states (region code 4). The other three clusters are regionally less homogeneous.

	acc	card	canc	pul	pneu	diab	liv
1	39.56	484.70	210.73	29.35	23.87	16.95	11.78
2	42.48	432.56	189.33	26.41	20.69	16.29	9.99
3	45.55	365.65	168.25	26.16	20.54	13.52	10.48
4	55.37	225.58	111.68	21.37	17.13	10.58	9.38

Table 11.3. The averages of the U.S. health data set within the 4 clusters. SMSclushealth

The differences between clusters are summarized in Table 11.3. It seems that most of the differences are due to the number of deaths due to cancer and cardiovascular problems, i.e., to the most common causes of deaths.

In Figure 11.15, we plot the first two principal components. The observations belonging to the four different clusters are plotted using different text size. Obviously, the cluster separated the observations according to their position

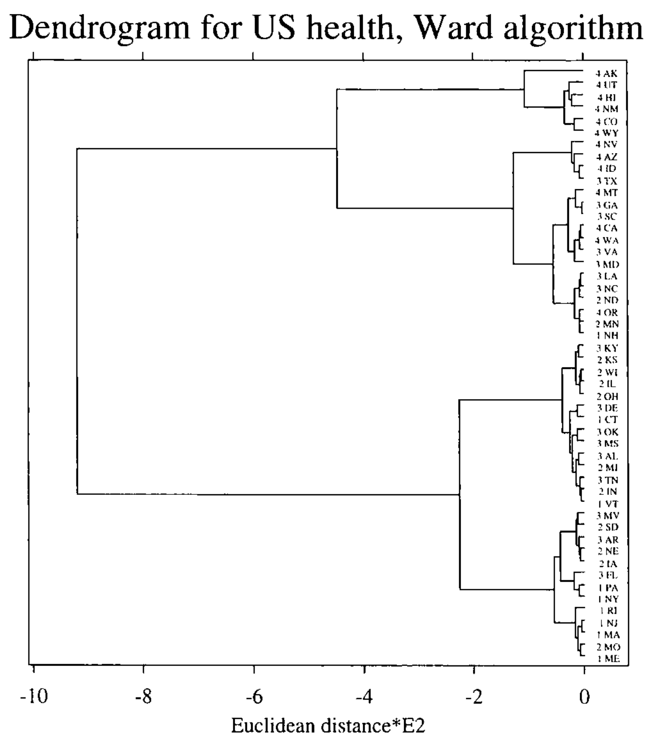

Fig. 11.14. Cluster analysis of U.S. health data set using Ward algorithm and Euclidean distance. `SMSclushealth`

on the horizontal axis of the plot, i.e., according to the value of the first principal component, see also the principal component analysis in Exercise 9.9.

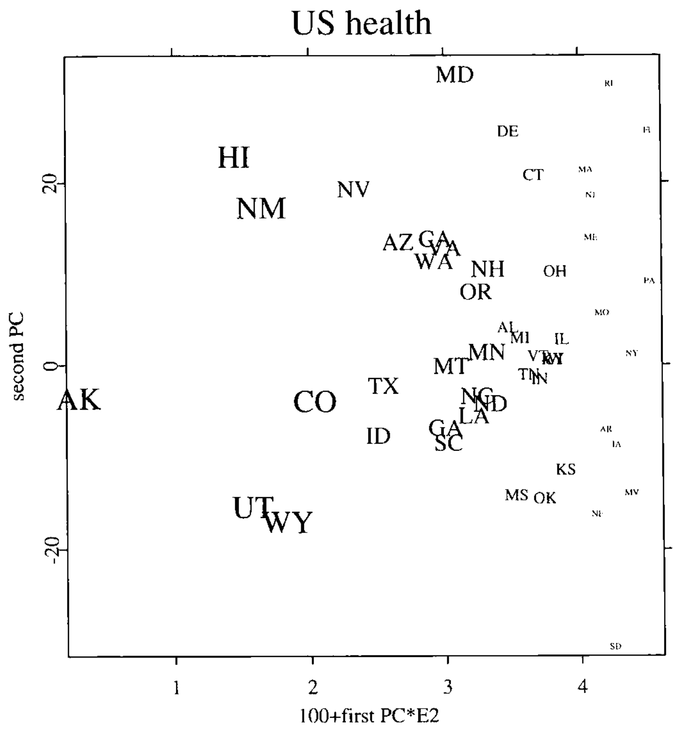

Fig. 11.15. Plot of the first two principal components of the U.S. health data. The size of the symbols is given by the clustering using Ward algorithm and Euclidean distance. `SMSclushealth`

12

Discriminant Analysis

> ... if a gentleman walks into my rooms smelling of iodoform, with a black mark of nitrate of silver upon his right fore-finger, and a bulge on the side of his top-hat to show where he has secreted his stethoscope, I must be dull indeed, if I do not pronounce him to be an active member of the medical profession.
> Sherlock Holmes in "A Scandal in Bohemia"

Discriminant analysis is used in situations where the clusters are known a priori. The aim of discriminant analysis is to classify an observation, or several observations, into these known groups. For instance, in credit scoring, a bank knows from past experience that there are good customers (who repay their loan without any problems) and bad customers (who have had difficulties repaying their loans). When a new customer asks for a loan, the bank has to decide whether or not to give the loan. The information of the bank is given in two data sets: multivariate observations on the two categories of customers (including age, salary, marital status, the amount of the loan, and the like).

The discrimination rule has to classify the customer into one of the two existing groups, and the discriminant analysis should evaluate the risk of a possible misclassification. Many other examples are described herein. We present ML discrimination and Fisher's linear discrimination function.

In the mathematical formulation of the problem, we try to allocate an observation to one of the populations $\Pi_j, j = 1, 2, ..., J$. A discriminant rule is a separation of the sample space (in general $\mathbb{R}^p$) into disjoint sets R_j such that if a new observation falls into the region R_j, it is identified as a member of population Π_j.

The quality of a discriminant rule can be judged on the basis of the error of misclassification.

If the probability density functions in the populations Π_j are known, we may easily derive a discriminant rule based on the maximum likelihood approach.

Maximum Likelihood Discriminant Rule

Let us assume that each population Π_j, $j = 1, \dots, J$, can be described by a probability density function (pdf) $f_j(x)$.

The maximum likelihood discriminant rule (ML rule) allocates the new observation x to the population Π_k, maximizing the likelihood $L_k(x) = f_k(x) = \max_{i=1,\dots,J} f_i(x)$.

Formally, the sets R_j, $j = 1, \dots, J$, given by the ML discriminant rule are:

$$R_j = \{x : f_j(x) \geq f_i(x) \text{ for } i = 1, \dots, J\}.$$

In practice, the sets R_j are constructed from estimates of the unknown densities. If the densities are assumed to have a known shape, i.e., normal distribution, it suffices to estimate the unknown parameters; see Exercise 12.1.

Bayes Discriminant Rule

The quality of the ML discriminant rule may be improved if some prior information about the probability of the populations is known. Let π_j denote the prior probability of class j. Note that $\sum_{j=1}^{J} \pi_j = 1$.

The Bayes discriminant rule allocates x to the population Π_k that gives the largest value of $\pi_i f_i(x)$, $\pi_k f_k(x) = \max_{i=1,\dots,J} \pi_i f_i(x)$. The Bayes discriminant rule can be formally defined by:

$$R_j = \{x : \pi_j f_j(x) \geq \pi_i f_i(x) \text{ for } i = 1, \dots, J\}.$$

The Bayes rule is identical to the ML discriminant rule if $\pi_j = 1/J$.

Fisher's Linear Discrimination Function

The classical Fisher's linear discriminant rule is based on the maximization of the ratio of the between to the within variance of a projection $a^\top x$.

Suppose we have samples $\mathcal{X}_j$, $j = 1, \dots, J$, from J populations. Let $\mathcal{Y} = \mathcal{X}a$ and $\mathcal{Y}_j = \mathcal{X}_j a$ denote linear combinations of observations. The within-group sum of squares is given by

$$\sum_{j=1}^{J} \mathcal{Y}_j^\top \mathcal{H}_j \mathcal{Y}_j = \sum_{j=1}^{J} a^\top \mathcal{X}_j^\top \mathcal{H}_j \mathcal{X}_j a = a^\top \mathcal{W} a, \tag{12.1}$$

where $\mathcal{H}_j$ denotes the $(n_j \times n_j)$ centering matrix. The between-group sum of squares is

$$\sum_{j=1}^{J} n_j (\overline{y}_j - \overline{y})^2 = \sum_{j=1}^{J} n_j \{a^\top (\overline{x}_j - \overline{x})\}^2 = a^\top \mathcal{B} a, \tag{12.2}$$

where $\overline{y}_j$ and $\overline{x}_j$ denote the means of $\mathcal{Y}_j$ and $\mathcal{X}_j$ and $\overline{y}$ and $\overline{x}$ denote the sample means of $\mathcal{Y}$ and $\mathcal{X}$.

Fisher noticed that the vector a that maximizes $a^\top \mathcal{B} a / a^\top \mathcal{W} a$ is the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ that corresponds to the largest eigenvalue.

Finally, observation x is classified into group j, which is closest to the projected $a^\top x$,

$$R_j = \{x : |a^\top (x - \overline{x}_j)| \leq |a^\top (x - \overline{x}_i)| \text{ for } i = 1, \dots, J\}.$$

EXERCISE 12.1. *Derive the ML discriminant rule if $\Pi_j = N_p(\mu_j, \Sigma)$, $j = 1, \dots, J$. Discuss the special case $J = 2$.*

Let us assume that the variance matrix Σ is positive definite. The likelihood of observation x in each of the populations Π_j, $j = 1, \dots, J$ is

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

According to the ML rule, we allocate x to the population Π_j with the largest likelihood. Omitting the constant $|2\pi\Sigma|^{-1/2}$ and taking logarithms, the maximization problem may be equivalently solved by minimizing

$$\begin{aligned} \delta^2(x, \mu_j) &= (x - \mu_j)^\top \Sigma^{-1} (x - \mu_j) \\ &= \{\Sigma^{-1/2}(x - \mu_j)\}^\top \Sigma^{-1/2}(x - \mu_j). \end{aligned}$$

Clearly, $\delta^2(x, \mu_j)$ is the square of the Mahalanobis distance between x and μ_j, see also Exercise 9.7 for the discussion of the Mahalanobis transformation.

Hence, in case of normal distribution with common covariance matrix, the ML rule allocates x to the closest group in the Mahalanobis sense.

For $J = 2$, the observation x is allocated to Π_1 if

$$(x - \mu_1)^\top \Sigma^{-1} (x - \mu_1) \leq (x - \mu_2)^\top \Sigma^{-1} (x - \mu_2).$$

Rearranging terms leads to

$$\begin{aligned} 0 &\geq -2\mu_1^\top \Sigma^{-1} x + 2\mu_2^\top \Sigma^{-1} x + \mu_1^\top \Sigma^{-1} \mu_1 - \mu_2^\top \Sigma^{-1} \mu_2 \\ 0 &\geq 2(\mu_2 - \mu_1)^\top \Sigma^{-1} x + (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 + \mu_2) \\ 0 &\leq (\mu_1 - \mu_2)^\top \Sigma^{-1} \{x - \frac{1}{2}(\mu_1 + \mu_2)\} \\ 0 &\leq \alpha^\top (x - \mu), \end{aligned}$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

It follows that in case of two multinormal populations, the discriminant rule can be written as:

$$R_1 = \{x : \alpha^\top (x - \mu) \geq 0\}.$$

EXERCISE 12.2. *Apply the rule from Exercise 12.1 for $J = 2$ and $p = 1$ and modify it for unequal variances.*

For two univariate normally distributed populations $\Pi_1 = N(\mu_1, \sigma)$ and $\Pi_2 = N(\mu_2, \sigma)$, the ML rule can be written as

$$\begin{aligned} R_1 &= \left\{ x : (\mu_1 - \mu_2)\left(x - \frac{\mu_1 + \mu_2}{2}\right) \geq 0 \right\} \\ R_1 &= \left\{ x : \operatorname{sign}(\mu_1 - \mu_2)\left(x - \frac{\mu_1 + \mu_2}{2}\right) \geq 0 \right\} \\ R_1 &= \left\{ x : \operatorname{sign}(\mu_1 - \mu_2) x \geq \operatorname{sign}(\mu_1 - \mu_2)\frac{\mu_1 + \mu_2}{2} \right\}. \end{aligned}$$

Assuming that $\mu_1 < \mu_2$, we obtain

$$R_1 = \left\{ x : x \leq \frac{\mu_1 + \mu_2}{2} \right\},$$

i.e., we classify x to R_1 if it is closer to μ_1 than to μ_2.

Assuming that the two normal populations have different variances, $\Pi_1 = N(\mu_1, \sigma_1^2)$ and $\Pi_2 : N(\mu_2, \sigma_2^2)$, we allocate x to R_1 if $L_1(x) > L_2(x)$, where the likelihood is:

$$L_i(x) = (2\pi\sigma_i^2)^{-1/2} \exp\left\{ -\frac{1}{2}\left(\frac{x - \mu_i}{\sigma_i}\right)^2 \right\}.$$

$L_1(x) \geq L_2(x)$ is equivalent to $L_1(x)/L_2(x) \geq 1$ and we obtain

$$\begin{aligned} \frac{\sigma_2}{\sigma_1} \exp\left\{ -\frac{1}{2}\left[\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - \left(\frac{x - \mu_2}{\sigma_2}\right)^2\right] \right\} &\geq 1 \\ \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2}\left[\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - \left(\frac{x - \mu_2}{\sigma_2}\right)^2\right] &\geq 0 \\ \frac{1}{2}\left[\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - \left(\frac{x - \mu_2}{\sigma_2}\right)^2\right] &\leq \log\frac{\sigma_2}{\sigma_1} \\ x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) &\leq 2\log\frac{\sigma_2}{\sigma_1}. \end{aligned}$$

If $\sigma_1 = \sigma_2$, most of the terms in the above formula disappear and the result simplifies to the discriminant rule obtained in Exercise 12.1.

EXERCISE 12.3. *Calculate the ML discrimination rule based on observations of a one-dimensional variable with an exponential distribution.*

The pdf of the exponential distribution $Exp(\lambda)$ is:

$$f(x) = \lambda \exp\{-\lambda x\} \text{ for } x > 0.$$

Comparing the likelihoods for two populations $\Pi_1 = Exp(\lambda_1)$ and $\Pi_2 = Exp(\lambda_2)$, we allocate the observation x into population Π_1 if

$$\begin{aligned} L_1(x) &\geq L_2(x) \\ L_1(x)/L_2(x) &\geq 1 \\ \frac{\lambda_1}{\lambda_2} \exp\{-x(\lambda_1 - \lambda_2)\} &\geq 1 \\ \log\frac{\lambda_1}{\lambda_2} - x(\lambda_1 - \lambda_2) &\geq 0 \\ x(\lambda_1 - \lambda_2) &\leq \log\frac{\lambda_1}{\lambda_2}. \end{aligned}$$

Assuming that $\lambda_1 < \lambda_2$, we obtain the discriminant rule:

$$R_1 = \left\{ x : x \geq \frac{\log\lambda_1 - \log\lambda_2}{\lambda_1 - \lambda_2} \right\}.$$

The observation x is classified into population Π_1 if it is greater than the constant $(\log\lambda_1 - \log\lambda_2)/(\lambda_1 - \lambda_2)$.

EXERCISE 12.4. *Calculate the ML discrimination rule based on observations of a two-dimensional random vector, where the first component has an exponential distribution and the other has an alternative distribution. What is the difference between the discrimination rule obtained in this exercise and the Bayes discrimination rule?*

Let us assume that the two populations, $\Pi_1 = \{Exp(\lambda_1), Alt(p_1)\}^\top$ and $\Pi_2 = \{Exp(\lambda_2), Alt(p_2)\}^\top$, are characterized by the exponential distribution with parameter λ_j and the alternative distribution with parameter p_j, $j = 1, 2$. The corresponding likelihood can be written as:

$$L_j(x_1, x_2) = \lambda_j \exp(-\lambda_j x_1)\{p_j x_2 + (1 - p_j)(1 - x_2)\}.$$

Since x_2 has the alternative distribution, it can have only two possible outcomes.

Assuming that $x_2 = 1$, we allocate the observation $(x_1, x_2)^\top$ to Π_1 if $L_1(x_1, 1) \geq L_2(x_1, 1)$, i.e.,

$$\begin{aligned} L_1(x_1,1)/L_2(x_1,1) &\geq 1 \\ \frac{\lambda_1 p_1}{\lambda_2 p_2} \exp\{-x_1(\lambda_1-\lambda_2)\} &\geq 1 \\ \log\frac{\lambda_1 p_1}{\lambda_2 p_2} - x_1(\lambda_1-\lambda_2) &\geq 0 \\ x_1(\lambda_1-\lambda_2) &\leq \log\frac{\lambda_1 p_1}{\lambda_2 p_2} \end{aligned}$$

Similarly, if $x_2 = 0$, we allocate the observation $(x_1, x_2)^\top$ to Π_1 if

$$x_1(\lambda_1-\lambda_2) \leq \log\frac{\lambda_1(1-p_1)}{\lambda_2(1-p_2)}.$$

Combining both cases and assuming that $\lambda_1 < \lambda_2$, the discriminant rule R_1 can be written as:

$$\left\{\begin{pmatrix}x_1\\x_2\end{pmatrix} : x_1 \geq \frac{\lambda_1\{x_2p_1+(1-x_2)(1-p_1)\}-\lambda_2\{x_2p_2+(1-x_2)(1-p_2)\}}{\lambda_1-\lambda_2}\right\}.$$

If the prior probabilities of $\Pi_1 = Exp(\lambda_1)$ and $\Pi_2 = Exp(\lambda_2)$ are π_1 and $\pi_2 = 1-\pi_1$, respectively, the Bayes rule can be derived by comparing $\pi_i L_i(x)$, $i = 1, 2$, exactly as in Exercise 12.3:

$$R_1 = \left\{x : x \geq \frac{\log \pi_1\lambda_1 - \log \pi_2\lambda_2}{\lambda_1-\lambda_2}\right\}.$$

Now, it is easy to see that the conditional discriminant rule obtained for the two dimensional random vector under the condition $x_2 = 1$ is equivalent to the Bayes discriminant rule for exponential distribution with $\pi_1 = p_1/(p_1+p_2)$. Similarly, the conditional discriminant rule if $x_2 = 0$ is a Bayes discriminant rule with $\pi_1 = (1-p_1)/(2-p_1-p_2)$.

EXERCISE 12.5. *Apply the Bayes rule to the car data in Table A.4 in order to discriminate between U.S., Japanese, and European cars. Consider only the variable milage (miles per gallon) and take the relative frequencies as prior probabilities.*

The three regions of origins in the data set are denoted by numbers 1, 2, and 3 standing for U.S., Japanese, and European cars, respectively. Based on the 74 observations given in Table A.4, we will construct a discriminant rule that would allow us to classify a new (75th) car with unknown origin.

Let us start with the maximum likelihood discriminant rule. Usually, the ML rule is based upon assumptions of normality. However, plots of the observed milage suggest that the normality is violated. Hence, instead of mileage measured in miles per gallon, we analyze fuel efficiency measured in liters per 100 kilometers. The averages in U.S., Japan, and Europe are: $\overline{x}_1 = 12.5207$, $\overline{x}_2 = 9.4577$, $\overline{x}_3 = 10.7712$. On average, Japanese cars (group 2) are more fuel efficient than European and U.S. cars.

The ML discriminant rule is calculated according to the description given in Exercise 12.1. In Figure 12.1, we plot the three point clouds corresponding to

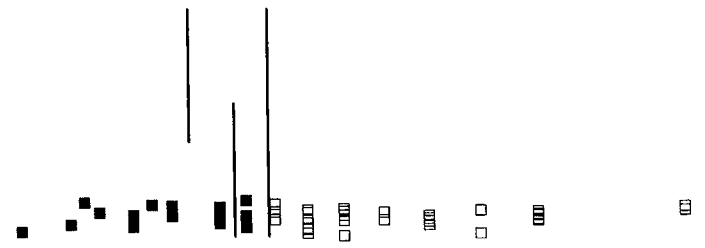

Fig. 12.1. Discrimination of the three regions according to "liters per 100km" with the ML discriminant rule. Q SMSdisccar

the three regions and, as vertical lines, we show also the points that separate the discriminant rules R_1, R_2, and R_3. The lowest point cloud (squares) in Figure 12.1 contains U.S. cars, the middle point (circles) cloud the Japanese, and the top point cloud (triangles) the European cars. The correctly classified cars are denoted by empty symbols whereas the filled symbols denote misclassified cars. The counts are given in Table 12.1.

	R_1: U.S.	R_2: JPN	R_3: EUR
Group 1 (U.S.)	33	11	8
Group 2 (Japanese)	2	7	2
Group 3 (European)	3	5	3

Table 12.1. The true region of origins and the region suggested by the ML discriminant rule based on fuel efficiency. The number of correct classifications for each region is given on the diagonal of the table.

The apparent error rate (APER), defined as the percentage of misclassified observations is $(11+8+2+2+3+5)/79 = 41.89\%$. It seems that the rule is not particularly good since we have less than 60% chance of correct classification. Moreover, this estimate is based on the observations which were used to construct the discriminant rule and it might be way too optimistic.

Let us now consider the Bayes rule which is based on the comparison of the likelihoods weighted by the prior probabilities of the groups. More formally, we allocate the new observation x to the population Π_j maximizing

$$\pi_j L_j(x) = \pi_j f_j(x) = \pi_j |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu_j)^\top \Sigma^{-1}(x-\mu_j)\right\},$$

where π_j, $j = 1, \dots, J$ are the prior probabilities of the respective populations.

Similarly as in Exercise 12.1, this problem is equivalent to minimizing

$$\begin{aligned}\delta^2(x, \mu_j, \pi_j) &= (x-\mu_j)^\top \Sigma^{-1}(x-\mu_j) - \log \pi_j \\ &= \{\Sigma^{-1/2}(x-\mu_j)\}^\top \Sigma^{-1/2}(x-\mu_j) - \log \pi_j.\end{aligned}$$

For $J = 2$, the observation x is allocated to Π_1 if

$$(x-\mu_1)^\top \Sigma^{-1}(x-\mu_1) - \log \pi_1 \le (x-\mu_2)^\top \Sigma^{-1}(x-\mu_2) - \log \pi_2.$$

Rearranging terms leads to

$$\begin{aligned}\log \pi_1 - \log \pi_2 &\ge -2\mu_1^\top \Sigma^{-1} x + 2\mu_2^\top \Sigma^{-1} x + \mu_1^\top \Sigma^{-1}\mu_1 - \mu_2^\top \Sigma^{-1}\mu_2 \\ \log \pi_1 - \log \pi_2 &\ge 2(\mu_2 - \mu_1)^\top \Sigma^{-1} x + (\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 + \mu_2) \\ \log \pi_2 - \log \pi_1 &\le (\mu_1 - \mu_2)^\top \Sigma^{-1}\{x - \frac{1}{2}(\mu_1 + \mu_2)\} \\ \log \frac{\pi_2}{\pi_1} &\le \alpha^\top (x - \mu),\end{aligned}$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$. Hence, the Bayes discriminant rule can be written as:

$$R_1 = \left\{x : \alpha^\top (x - \mu) \ge \log \frac{\pi_2}{\pi_1}\right\}.$$

In our car data example, we use the relative frequencies observed in the data set, $\pi_1 = 0.7027$, $\pi_2 = 0.1486$, $\pi_3 = 0.1486$, as the prior probabilities.

The resulting discriminant rule is graphically displayed in Figure 12.2. Notice that with these weights, it is impossible to classify any new observation as a European car.

The same results are given in Table 12.2. The apparent error rate is equal to 28.38%. Obviously, the Bayes discriminant rule leads to better results since it give large weights to U.S. cars which constitute more than 60% of the entire data set.

EXERCISE 12.6. *Derive simple expressions for matrices $\mathcal{W}$ and $\mathcal{B}$ and the Fisher discriminant rule in the setup of the Swiss bank notes data set given in Table A.2.*

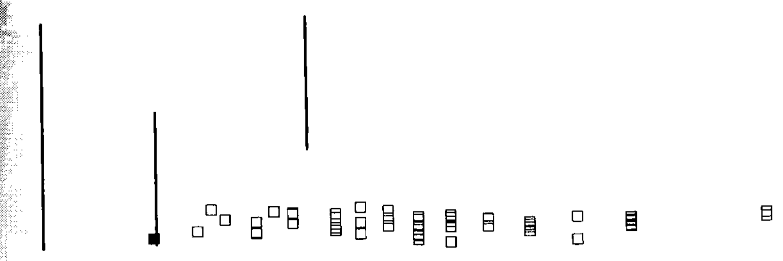

Fig. 12.2. Discrimination of the three regions according to "liters per 100km" using the Bayes rule. SMSdiscbaycar

	R_1: U.S.	R_2: JPN	R_3: EUR
Group 1 (U.S.)	51	1	0
Group 2 (Japanese)	9	2	0
Group 3 (European)	10	1	0

Table 12.2. The true region of origins and the region suggested by the Bayes discriminant rule based on fuel efficiency. The number of correct classifications for each region is given on the diagonal of the table.

The Swiss bank notes data set, $\mathcal{X}$, contains six measurements taken on 100 genuine and 100 counterfeit bank notes. Let us denote the measurements taken on genuine and counterfeit by $\mathcal{X}_g$ and $\mathcal{X}_f$, respectively. The corresponding linear combinations are $\mathcal{Y} = \mathcal{X}a$, $\mathcal{Y}_g = \mathcal{X}_g a$, and $\mathcal{Y}_f = \mathcal{X}_f a$.

The within-group sum of squares (12.1) satisfies the relation

$$\mathcal{Y}_f^\top \mathcal{H}_f \mathcal{Y}_f + \mathcal{Y}_g^\top \mathcal{H}_g \mathcal{Y}_g = a^\top \mathcal{W} a,$$

where $\mathcal{H}_f$ and $\mathcal{H}_g$ denote the appropriate centering matrices of dimensions $n_f = n_g = 100$. Observe that

$$a^\top \mathcal{W} a = a^\top (\mathcal{X}_f^\top \mathcal{H}_f \mathcal{X}_f + \mathcal{X}_g^\top \mathcal{H}_g \mathcal{X}_f) a$$

and, hence, the matrix $\mathcal{W}$ can be written as:

$$\begin{aligned}\mathcal{W} &= \mathcal{X}_f^\top \mathcal{H}_f \mathcal{X}_f + \mathcal{X}_g^\top \mathcal{H}_g \mathcal{X}_g = \mathcal{H}_f \mathcal{X}_f^\top \mathcal{H}_f \mathcal{X}_f + \mathcal{H}_g \mathcal{X}_g^\top \mathcal{H}_g \mathcal{X}_g \\ &= n_f \mathcal{S}_f + n_g \mathcal{S}_g = 100(\mathcal{S}_f + \mathcal{S}_g),\end{aligned}$$

where $\mathcal{S}_g$ and $\mathcal{S}_f$ denote the empirical covariances w.r.t. the genuine and counterfeit bank notes.

For the between-group sum of squares (12.2) we have

$$a^\top \mathcal{B} a = n_f(\overline{y}_f - \overline{y})^2 + n_g(\overline{y}_g - \overline{y})^2,$$

where $\overline{y}$, $\overline{y}_f$, and $\overline{y}_g$ denote respectively the sample means of $\mathcal{Y}$, $\mathcal{Y}_f$, and $\mathcal{Y}_g$. It follows that

$$a^\top \mathcal{B} a = a^\top \{n_f(\overline{x}_f - \overline{x})(\overline{x}_f - \overline{x})^\top + n_g(\overline{x}_g - \overline{x})(\overline{x}_g - \overline{x})^\top\} a,$$

where $\overline{x}$, $\overline{x}_f$, and $\overline{x}_g$ denote respectively the column vectors of sample means of $\mathcal{X}$, $\mathcal{X}_f$, and $\mathcal{X}_g$. Hence, we obtain

$$\begin{aligned}
\mathcal{B} &= n_f(\overline{x}_f - \overline{x})(\overline{x}_f - \overline{x})^\top + n_g(\overline{x}_g - \overline{x})(\overline{x}_g - \overline{x})^\top \\
&= 100\{(\overline{x}_f - \overline{x})(\overline{x}_f - \overline{x})^\top + (\overline{x}_g - \overline{x})(\overline{x}_g - \overline{x})^\top\} \\
&= 100\left\{\left(\overline{x}_f - \frac{\overline{x}_f + \overline{x}_g}{2}\right)\left(\overline{x}_f - \frac{\overline{x}_f + \overline{x}_g}{2}\right)^\top\right. \\
&\quad \left. + \left(\overline{x}_g - \frac{\overline{x}_f + \overline{x}_g}{2}\right)\left(\overline{x}_g - \frac{\overline{x}_f + \overline{x}_g}{2}\right)^\top\right\} \\
&= 25(\overline{x}_f - \overline{x}_g)(\overline{x}_f - \overline{x}_g)^\top.
\end{aligned}$$

The vector a maximizing the ratio $a^\top \mathcal{B} a / a^\top \mathcal{W} a$ can be calculated as the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ corresponding to the largest eigenvalue, see Härdle & Simar (2003, theorem 12.4).

For the Swiss bank notes, it is easy to see that the matrix $\mathcal{W}^{-1}\mathcal{B}$ can have at most one nonzero eigenvalue since $\operatorname{rank} \mathcal{B} \leq 1$. The nonzero eigenvalue λ_1 can be calculated as:

$$\begin{aligned}
\lambda_1 = \sum_{j=1}^{p} \lambda_j &= \operatorname{tr} \mathcal{W}^{-1}\mathcal{B} = \operatorname{tr} \mathcal{W}^{-1} 25(\overline{x}_f - \overline{x}_g)(\overline{x}_f - \overline{x}_g)^\top \\
&= 25 \operatorname{tr}(\overline{x}_f - \overline{x}_g)^\top \mathcal{W}^{-1}(\overline{x}_f - \overline{x}_g) = 25(\overline{x}_f - \overline{x}_g)^\top \mathcal{W}^{-1}(\overline{x}_f - \overline{x}_g).
\end{aligned}$$

From the equation:

$$\mathcal{W}^{-1}\mathcal{B}\mathcal{W}^{-1}(\overline{x}_f - \overline{x}_g) = 25(\overline{x}_f - \overline{x}_g)^\top \mathcal{W}^{-1}(\overline{x}_f - \overline{x}_g)\mathcal{W}^{-1}(\overline{x}_f - \overline{x}_g)$$

it follows that the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ corresponding to the largest eigenvalue is $a = \mathcal{W}^{-1}(\overline{x}_f - \overline{x}_g)$. Assuming that $\overline{y}_f > \overline{y}_g$, the corresponding discriminant rule can be formally written as:

$$R_f = \{x : (\overline{x}_f - \overline{x}_g)^\top \mathcal{W}^{-1}(x - \overline{x}) \geq 0\}.$$

EXERCISE 12.7. *Compute Fisher's linear discrimination function for the 20 bank notes from Exercise 11.6. Apply it to the entire bank data set. How many observations are misclassified?*

Applying the formulas derived in the previous Exercise 12.6 with $n_f = n_g = 10$, using the randomly chosen observations with indices 7, 8, 16, 39, 71, 73, 89, 94, 94, 100, 110, 121, 129, 131, 149, 150, 154, 161, 163, and 174, we obtain $\overline{x}_g = (214.72, 129.79, 129.64, 8.00, 10.18, 141.48)^\top$, $\overline{x}_f = (214.85, 130.13, 130.13, 10.33, 11.31, 139.53)^\top$, and

$$\mathcal{W} = \begin{pmatrix}
3.36 & 0.40 & 0.90 & -3.32 & -0.00 & 0.38 \\
0.40 & 1.49 & 0.95 & 0.41 & -0.52 & 0.91 \\
0.90 & 0.95 & 1.91 & 2.43 & -1.38 & 1.31 \\
-3.32 & 0.41 & 2.43 & 18.02 & -10.17 & 2.86 \\
-0.00 & -0.52 & -1.38 & -10.17 & 11.46 & -2.39 \\
0.38 & 0.91 & 1.31 & 2.86 & -2.39 & 3.66
\end{pmatrix}.$$

The eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ corresponding to the largest eigenvalue can then be calculated as

$$(\overline{x}_f - \overline{x}_g)^\top \mathcal{W}^{-1} = (-1.56, -1.19, 1.38, -1.21, -0.88, 0.87)^\top.$$

The new observation x will be allocated as a counterfeit bank note if $a^\top(x - \overline{x}) \geq 0$. Calculating the Fisher linear discriminant rule for all observations in the Swiss bank notes data set, we obtain altogether six genuine bank notes classified as counterfeit. None of the counterfeit bank notes is classified as genuine. Hence, the estimated error rate is $6/200 = 3\%$. This estimate might be too optimistic since some of the bank notes used for the construction were used also for the evaluation of the rule.

EXERCISE 12.8. *Derive a discriminant rule based on the ML method with $J = 2$ minimizing the expected cost misclassification considering the prior probability $\pi_1 = \frac{1}{3}$ and the expected cost of misclassification $C(2|1) = 2C(1|2)$.*

The expected cost of misclassification is given by $ECM = C(2|1)p_{21}\pi_1 + C(1|2)p_{12}\pi_2$, where p_{21} is the probability of wrong classification of observation coming from group 1 and p_{12} is the probability of wrong classification of observation coming from group 2.

Assuming that the populations Π_1 and Π_2 are characterized by the probability densities $f_1(.)$ and $f_2(.)$, we can derive the loss $L(R_1)$ as a function of the discriminant rule R_1:

$$\begin{aligned}
L(R_1) &= C(2|1)\pi_1 p_{21} + C(1|2)\pi_2 p_{12} \\
&= C(2|1)\pi_1 \int_{R_2} f_1(x)dx + C(1|2)\pi_2 \int_{R_1} f_2(x)dx \\
&= C(2|1)\pi_1 \int \{1 - \mathbf{I}(x \in R_1)\} f_1(x)dx + C(1|2)\pi_2 \int \mathbf{I}(x \in R_1) f_2(x)dx \\
&= C(2|1)\pi_1 + \int \mathbf{I}(x \in R_1)\{C(1|2)\pi_2 f_2(x) - C(2|1)\pi_1 f_1(x)\}dx.
\end{aligned}$$

The loss $L(R_1)$ is obviously minimized if R_1 is chosen so that $x \in R_1$ is equivalent to $C(1|2)\pi_2 f_2(x) - C(2|1)\pi_1 f_1(x) < 0$. Hence, the optimal discriminant rule is:

$$\begin{aligned} R_1 &= \{x : C(1|2)\pi_2 f_2(x) - C(2|1)\pi_1 f_1(x) < 0\} \\ &= \{x : C(2|1)\pi_1 f_1(x) > C(1|2)\pi_2 f_2(x)\} \\ &= \left\{x : \frac{f_1(x)}{f_2(x)} > \frac{C(1|2)\pi_2}{C(2|1)\pi_1}\right\}. \end{aligned}$$

Assuming that $\pi_1 = \frac{1}{3}$ and that the expected cost of misclassification $C(2|1) = 2C(1|2)$ leads $\pi_2 = 1 - \pi_1 = 2/3 = 2\pi_1$ and the resulting discriminant rule is:

$$R_1 = \left\{x : \frac{f_1(x)}{f_2(x)} > \frac{C(1|2)2\pi_1}{2C(1|2)\pi_1}\right\} = \left\{x : \frac{f_1(x)}{f_2(x)} > 1\right\} = \{x : f_1(x) > f_2(x)\},$$

i.e., we obtain the ML discriminant rule. SMSdisfbank

EXERCISE 12.9. *Explain the effect of changing π_1 or $C(1|2)$ on the relative location of the region $R_j, j = 1, 2$ in Exercise 12.8.*

In Exercise 12.8, we have derived the discriminant rule

$$R_1 = \left\{x : \frac{f_1(x)}{f_2(x)} > \frac{C(1|2)\pi_2}{C(2|1)\pi_1}\right\}.$$

Increasing the cost of misclassification $C(1|2)$ would increase the constant in the definition of R_1 and, hence, it would make the region R_1 smaller.

Increasing the prior probability π_1 of the population Π_1 would make the same constant smaller and the region R_1 would grow.

EXERCISE 12.10. *Prove that Fisher's linear discrimination function is identical to the ML rule for multivariate normal distributions with equal covariance matrices ($J = 2$).*

The ML rule in this situation has been derived in Exercise 12.1,

$$R_1^{ML} = \{x : \alpha^\top (x - \mu) \geq 0\},$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ and $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

Fisher's linear discrimination rule derived for $J = 2$ in Exercise 12.6 is:

$$R_1^F = \{x : (\overline{x}_1 - \overline{x}_1)^\top \mathcal{W}^{-1}(x - \overline{x}) \geq 0\}.$$

In the same exercise, we have also shown that $\mathcal{W} = n\mathcal{S}$, where $\mathcal{S}$ denotes the pooled covariance matrix and n the number of observations. Defining the empirical version of α as $\widehat{\alpha} = (\overline{x}_1 - \overline{x}_2)^\top \mathcal{S}^{-1}$, we can rewrite the Fisher's discriminant rule as:

$$R_1^F = \{x : \widehat{\alpha}^\top (x - \overline{x}) \geq 0\}.$$

Comparing this expression with the ML discriminant rule, we see that Fisher's rule R_1^F may be interpreted as the empirical version (estimate) of the ML discriminant rule R_1^{ML}.

EXERCISE 12.11. *Suppose that the observations come from three distinct populations, Π_1, Π_2, and Π_3, characterized by binomial distributions:*

$$\begin{aligned} \Pi_1 : X &\sim \text{Bi}(10, 0.2) \quad \textit{with the prior probability } \pi_1 = 0.5; \\ \Pi_2 : X &\sim \text{Bi}(10, 0.3) \quad \textit{with the prior probability } \pi_2 = 0.3; \\ \Pi_3 : X &\sim \text{Bi}(10, 0.5) \quad \textit{with the prior probability } \pi_3 = 0.2. \end{aligned}$$

Use the Bayes method to determine the discriminant rules R_1, R_2, and R_3.

The corresponding Bayes discriminant rules R_j for $j = 1, 2, 3$ are defined as:

$$R_j = \{x \in \{0, 1, \ldots, 9, 10\} : \pi_j f_j(x) \geq \pi_i f_i(x) \text{ for } i = 1, 2, 3\}.$$

x	$f_1(x)$	$f_2(x)$	$f_3(x)$	$\pi_1 f_1(x)$	$\pi_2 f_2(x)$	$\pi_3 f_3(x)$	$\pi_j f_j(x)$	j
0	0.107374	0.028248	0.000977	0.053687	0.008474	0.000195	0.053687	1
1	0.268435	0.121061	0.009766	0.134218	0.036318	0.001953	0.134218	1
2	0.301990	0.233474	0.043945	0.150995	0.070042	0.008789	0.150995	1
3	0.201327	0.266828	0.117188	0.100663	0.080048	0.023438	0.100663	1
4	0.088080	0.200121	0.205078	0.044040	0.060036	0.041016	0.060036	2
5	0.026424	0.102919	0.246094	0.013212	0.030876	0.049219	0.049219	3
6	0.005505	0.036757	0.205078	0.002753	0.011027	0.041016	0.041016	3
7	0.000786	0.009002	0.117188	0.000393	0.002701	0.023438	0.023438	3
8	0.000074	0.001447	0.043945	0.000037	0.000434	0.008789	0.008789	3
9	0.000004	0.000138	0.009766	0.000002	0.000041	0.001953	0.001953	3
10	0.000000	0.000006	0.000977	0.000000	0.000002	0.000195	0.000195	3

Table 12.3. The values of the likelihood and Bayesian likelihood for three binomial distributions.

The values of $\pi_i f_i(x)$, for $i = 1, \ldots, 3$ and $x = 0, \ldots, 10$ are given in Table 12.3 from which it directly follows that the discriminant rules are:

$$\begin{aligned} R_1 &= \{0, 1, 2, 3\}, \\ R_2 &= \{4\}, \\ R_3 &= \{5, 6, 7, 8, 9, 10\}. \end{aligned}$$

EXERCISE 12.12. *Use the Fisher's linear discrimination function on the WAIS data set (Table A.21) and evaluate the results by re-substitution to calculate the probabilities of misclassification.*

The WAIS data set contains results of four subtests of the Wechsler Adult Intelligence Scale for two categories of people. Group 2 contains 12 observations of those presenting a senile factor and group 1 contains 37 people serving as a control.

Applying the formulas derived in Exercise 12.6 and proceeding as in Exercise 12.7, we obtain the eigenvector

$$(\overline{x}_2 - \overline{x}_1)^\top \mathcal{W}^{-1} = (-0.0006, -0.0044, -0.0002, -0.0095)^\top .$$

Calculating the Fisher's discriminant rule from all observations leads to 4 misclassified observations in group 2 and 8 misclassified observations in group 1.

Hence, the apparent error rate (APER) is equal to $(4+8)/49 = 24.49\%$. The disadvantage of this measure of the quality of the discriminant rule is that it is based on the same observations that were used to construct the rule.

In order to obtain a more appropriate estimate of the misclassification probability, we may proceed in the following way:

1. Calculate the discrimination rule from all but one observation.
2. Allocate the omitted observation according to the rule from step 1.
3. Repeat steps 1 and 2 for all observations and count the number of correct and wrong classifications.

The estimate of the misclassification rate based on this procedure is called the actual error rate (AER).

Running the algorithm for the WAIS data set, we misclassify 4 observations in group 2 and 11 observations in group 1. The AER is $(4+11)/49 = 30.61\%$.

Hence, if a new patient arrives, he will be correctly classified with probability approximately 70%. Q SMSdisfwais