7 Principal Components and Factor Analysis

7.1 Principal Components

a **Goal.** Relationships between two variables can be graphically well captured in a meaningful way. For three variables this is also possible, but for four it becomes harder. It is therefore desirable to reduce a problem that concerns many variables to one with two or three dimensions.

In the NIR spectra example in the introduction (1.2.g) it was made plausible that the essentials of the chemical process might play out in a few dimensions. In many other cases, there exists no such theory, but we still can try **to use the given variables to form two or three new ones that capture the "essence" of the data.**

This task of **dimension reduction** can be well illustrated if we initially discuss how we transform a two dimensional problem into a one dimensional one – which is not really necessary.

Data and Model. This and the following section can be studied even if the chapter about models has not yet been worked through, simply ignoring the paragraphs entitled "Model Version".

b **Two Variables.** First, we want to transform two dimensional data into one dimensional in a way that makes sense. Fig. 7.1.b illustrates the idea of **principal component analysis**:

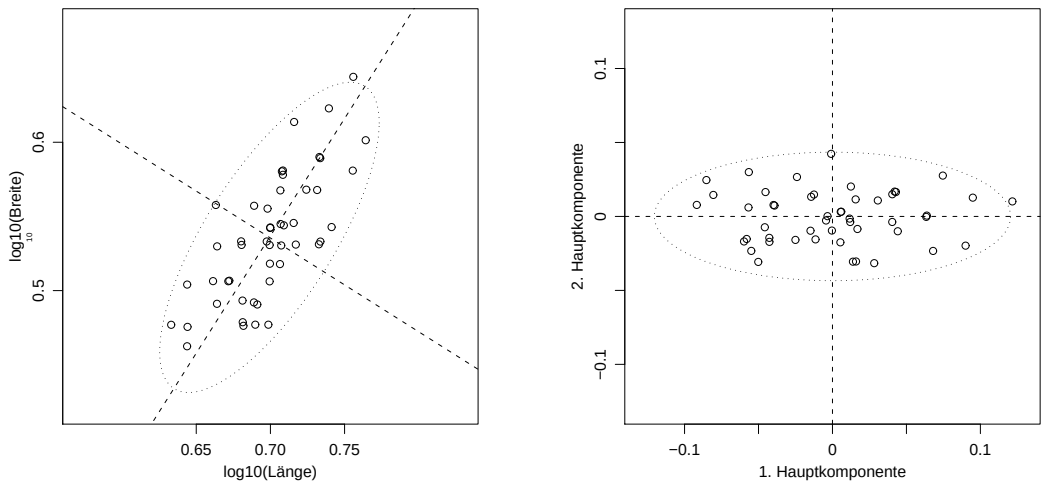

Figure 7.1.b: Principal components in the iris flowers example

The "scatter" in the data should be reflected as well as possible by a projection of the data points onto an appropriate direction. More precisely, we do this so that the variance of the projected points will be as large as possible!

How this direction is determined is given below, for the more general case of more possibly than two variables. In the figure, the direction perpendicular to the direction of maximal scatter is also shown. We can show that this is the direction for which the projections of the points have the smallest variance.

Model Version. We have discussed how the "scatter" of the data can be shown by ellipses that represent the "contours" of equal density of the corresponding bivariate normal distribution. The direction of the longer principal axis of the ellipse maximizes the variance of the projection of the random vector $\underline{X}$. The shorter principal axis of the ellipse, which is perpendicular to the longer, gives the direction with the smallest variance. (* This means that the longer principal axis coincides with the line determined by *orthogonal regression*.)

c **The Principal Axes Transformation.** The two directions that were found can be interpreted as a **new coordinate system**. They are called **factors** – as a loan from the more general method of **factor analysis**, see below (7.4.a). Unfortunately, this use of the word is in conflict with the idea of a factor in analysis of variance!

The conversion of the variables $X^{(j)}$ from the original coordinate systems into principal component coordinates is done for the data by $\underline{z}_i = \widehat{\boldsymbol{B}}(\underline{x}_i - \overline{\underline{x}})$ with a matrix $\widehat{\boldsymbol{B}}$, which is determined from the covariance matrix $\widehat{\boldsymbol{\Sigma}}$. For the whole data matrix, this becomes

$$\boldsymbol{z} = \boldsymbol{x}_c \widehat{\boldsymbol{B}}^T \ .$$

Model Version. In the model, the random vector is determined in the new coordinate system via the linear transformation

$$\underline{Z} = \boldsymbol{B}(\underline{X} - \underline{\mu})$$

where $\boldsymbol{B}$ depends on $\boldsymbol{\Sigma}$.

The matrices $\widehat{\boldsymbol{B}}$ and $\boldsymbol{B}$ should be orthogonal so that the distances between the points remain the same and, in this sense, "relationships between the observations" are not altered. We will state below how they are calculated.

d **Standardized Random Variables.** The matrix $\boldsymbol{z}$ here is not a standardized data matrix like in 2.6.m. The columns of $\boldsymbol{z}$ do have mean value 0 and are, as we shall see, uncorrelated, but they have different variances. To get a standardized dataset, we must still divide the columns by their standard deviations.

Model Version. The random vector $\underline{Z}$ is not standardized like in 3.2.k. The components $Z^{(1)}$ and $Z^{(2)}$ do have expected value 0 and are uncorrelated, but they have different variances. To get a standardized random vector, we must divide the individual $Z^{(j)}$ by their standard deviations.

Earlier, we introduced standardized datasets (random vectors) without using the principal component analysis. We determined the transformation matrix from the Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}$ ($\boldsymbol{\Sigma}$) (2.6.m). This concretely shows that there are various possible ways to get standardized multidimensional data.

e **More than 2 Variables.** We require that $Z^{(1)}$ have a variance as large as possible, which is determined by the first principal component or the first row of $\widehat{\boldsymbol{B}}$ ($\boldsymbol{B}$). For $m=2$ dimensional data, the rotation of the coordinate system is therefore clear.

For $m>2$ the question remains of which linear transformation should form the second principal component. The obvious requirement is again that the new variable $Z^{(2)}$ should be scattered as much as possible – under the additional condition that the overall transformation matrix $\widehat{\boldsymbol{B}}$ (or $\boldsymbol{B}$) should end up orthogonal. We can proceed with such conditions until the determination of the last variable $Z^{(m)}$ allows no more freedom of choice.

f **Calculation of the Transformation Matrix.** Linear algebra provides the solution of this ambitious-sounding problem under the keyword **eigenvalue problem**. The matrix $\widehat{\boldsymbol{B}}$ depends only on the covariance matrix $\widehat{\boldsymbol{\Sigma}}$. It includes (as the rows $\underline{\widehat{b}}_k$) the so-called **eigenvectors** of $\widehat{\boldsymbol{\Sigma}}$. (The order is actually arbitrary for the eigenvalue problem and is normally determined by the convention that the associated **eigenvalues** $\widehat{\lambda}_k$ should be sorted by size, so that $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_m$. This order makes sure that we get the solution of the problem of the principal components as stated above.)
If this matrix $\widehat{\boldsymbol{B}}$ is used for the linear transformation $\underline{z}_i = \widehat{\boldsymbol{B}}(\underline{x}_i - \overline{\underline{x}})$, then

$$\widehat{\mathrm{var}}\langle \underline{Z} \rangle = \widehat{\boldsymbol{B}}\,\widehat{\boldsymbol{\Sigma}}\,\widehat{\boldsymbol{B}}^T = \widehat{\boldsymbol{D}} = \begin{bmatrix} \widehat{\lambda}_1 & 0 & \dots & 0 \\ 0 & \widehat{\lambda}_2 & \dots & 0 \\ \cdot & \vdots & \ddots & \\ 0 & 0 & \dots & \widehat{\lambda}_m \end{bmatrix} .$$

The empirical variances of the new coordinates $z_i^{(k)}$ are therefore equal to the eigenvalues of the matrix $\widehat{\boldsymbol{\Sigma}}$, and the $\underline{z}^{(k)}$ are uncorrelated.

Model Version. The exact same relationship holds for the random vectors $\underline{X}$ and $\underline{Z}$, if we leave out the "hats" $\widehat{\ }$ in the equations.

g* We would like to sketch the relationship between the identified statistical problem and the eigenvalue problem of linear algebra for the model version. In 7.1.b we looked for an orthogonal transformation that brings the ellipse into principal axis position. This means that the covariance matrix of $\underline{Z}$ should be diagonal. There therefore seek an orthogonal $\boldsymbol{B}$, so that the previous equation is true. If we multiply this equation with $\boldsymbol{B}^T$ from the left, we get $\boldsymbol{\Sigma}\,\boldsymbol{B}^T = \boldsymbol{B}^T\boldsymbol{D}$. If we write the columns of this matrix equation individually, this becomes $\boldsymbol{\Sigma}\underline{b}_k = \lambda_k \underline{b}_k$. This is the equation that in linear algebra is called the **eigenvalue problem**. How the solution is concretely calculated will not be discussed here. (There are maximally m different solutions for λ_k, but always at least m different $\underline{b}_k$. If there are m different λ_k, the $\underline{b}_k$ are unique, otherwise it is more complicated.)

h The goal of the exercise was **dimension reduction**. A two dimensional data set can be projected onto the first principal axis. Similarly, in a higher dimensional space, we can limit ourselves to the first p principal components which, together, reflect a maximal portion of the "scatter in the data". Formulated more precisely, we transform the data into principal component coordinates and leave out (set to zero) those that correspond to the $m-p$ smallest eigenvalues λ_j.

i **Model.** This simple idea can also be formulated as a model that appears similar to a linear regression model. We solve 7.1.c for $\boldsymbol{x}_c$ and get, since $\widehat{\boldsymbol{B}}^{-1} = \widehat{\boldsymbol{B}}^T$,

$$\boldsymbol{x}_c = \boldsymbol{z}\,\boldsymbol{B} \ .$$

If we now, on the right side, take "seriously" only the first p principal components – i.e. the first p columns of $\boldsymbol{z}$ – and denote them with $\boldsymbol{S} = \boldsymbol{z}^{[1:p]}$, we get as the "essential part" of $\boldsymbol{x}_c$ the matrix $\widehat{\boldsymbol{x}}_c$,

$$\boldsymbol{x}_c \approx \widehat{\boldsymbol{x}}_c = \boldsymbol{S}\,\boldsymbol{C}^T \ ,$$

where $\boldsymbol{C}^T = \boldsymbol{B}_{[1:p]}$ contains the first p rows of $\boldsymbol{B}$. We therefore break up the above equation into an "essential part" and a remainder $\boldsymbol{r}$,

$$\boldsymbol{x}_c = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{r} \ .$$

This looks similar to a regression model in matrix form, $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{E}$, and even more similar to the multivariate form of this, $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{E}$, see 6.1.b. It suggest the division of the data into a "systematic structure" and deviations that can be interpreted as random.

However, we won't do ths here; the idea will just serve as a connection and starting point for later. **Principal component analysis is essentially a method of descriptive statistics.** To use the above equation as a model, we must at least have assumptions about the distribution of the deviations $\boldsymbol{r}$. We'll do that later and thereby advance to **factor analysis**.

j ▷ **I**n the **NIR spectra example** the absorptions for $m = 700$ wavelengths in the infrared region between 1100 and 2500 nm are the variables, which have been measured for $n = 121$ time points in the course of a chemical reaction. Fig. 7.1.j shows the course in the space of the first five principal components via a scatter plot matrix. In the first four principal components we see a "curve" that corresponds to the time course of the reaction and can be interpreted as "the essential structure" of the data. ◁

k **Choice of Dimension.** How large a p should be chosen is dependent on the purpose of the analysis. In a scatter plot matrix of the principal components $\boldsymbol{Z}$ we begin looking at the upper left and stop if no more interesting structure is recognizable in the individual diagrams. In the example, the fifth principal component shows the typical image of purely random scatter.

The variances of the principal components $\mathrm{var}\langle Z^{(k)}\rangle$ are a useful tool for choosing the dimension. We can graphically show the few numbers in a bar chart; the figure is called a **scree plot**.

In the example (Fig. 7.1.k) we see that the first dimension already captures the majority of the variability of the data. The log representation gives the idea that an additional three or four components also are meaningful. Later, the "curve" clearly becomes flatter and we suspect that these components contain no more important information. We like to find such a "knee" in the scree plot. Then, the number where the "joint" occurs corresponds to the first variable that can be *neglected*, and p is thus one less.

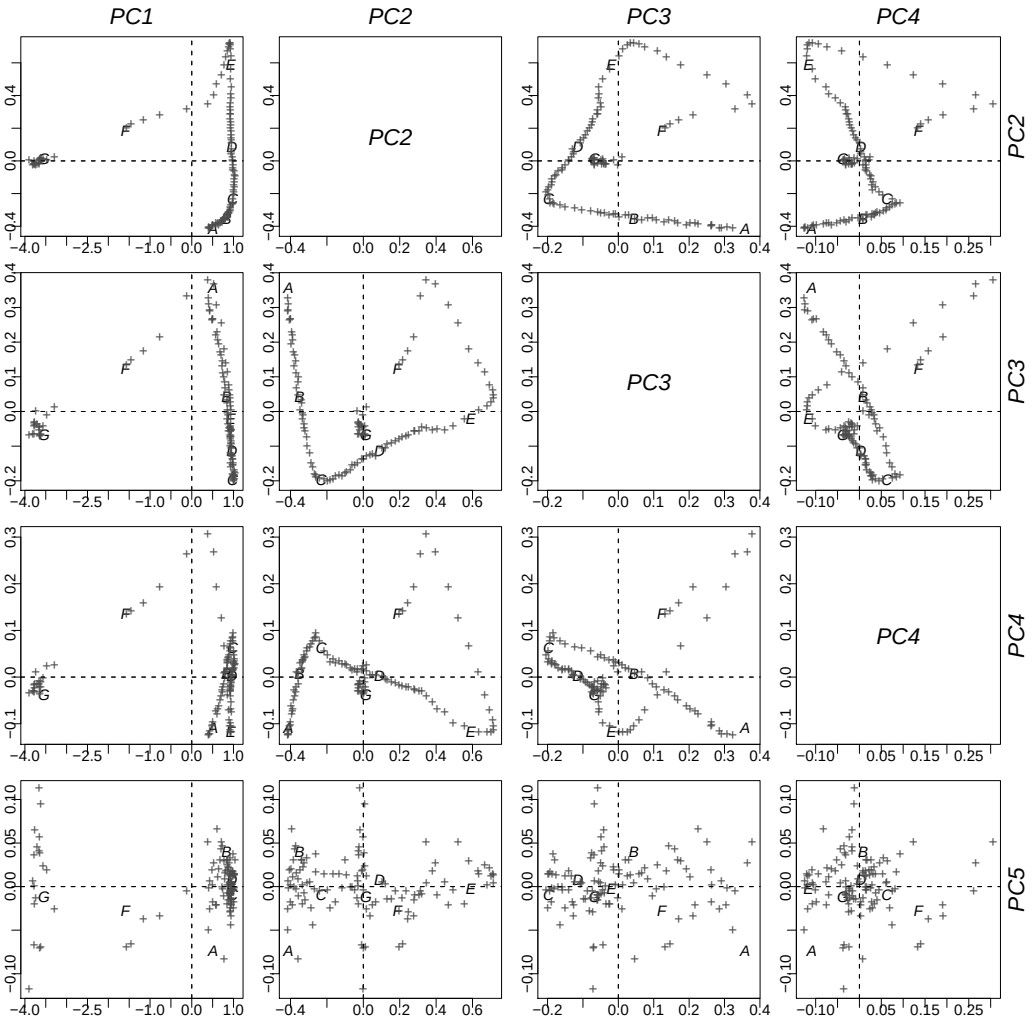

Figure 7.1.j: Scatter plot matrix of the first five principal components in the NIR spectra example

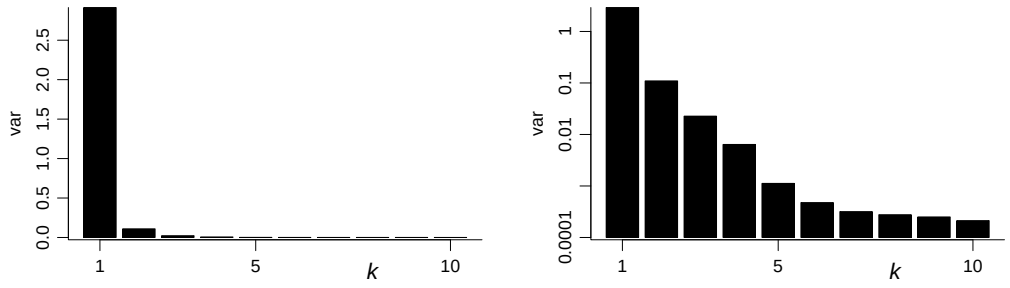

Figure 7.1.k: Variances of the Principal Components

7.2 The Biplot

a Principal component analysis provides the approximation $\boldsymbol{x}_c \approx \widehat{\boldsymbol{x}}_c = \boldsymbol{S}\,\boldsymbol{C}^T$, where $\boldsymbol{S}$ and $\boldsymbol{C}$ have only p columns. If the approximation for $p = 2$ is good, then the scatter plot of the first two principal components represents "the essentials". If we take care that the units on the two axes are the same, then the distances between the points are approximately equal to the distances of the observation vectors, and therefore meaningfully represent dissimilarities between observations.

b In the same figure we now also draw the rows $\underline{C}_j$ of the matrix $\boldsymbol{C}$ as arrows that characterize the variables (Fig. 7.2.b). (To get an informative picture, the scale for plotting them is adjusted suitably. It is given at the top and right margin.)

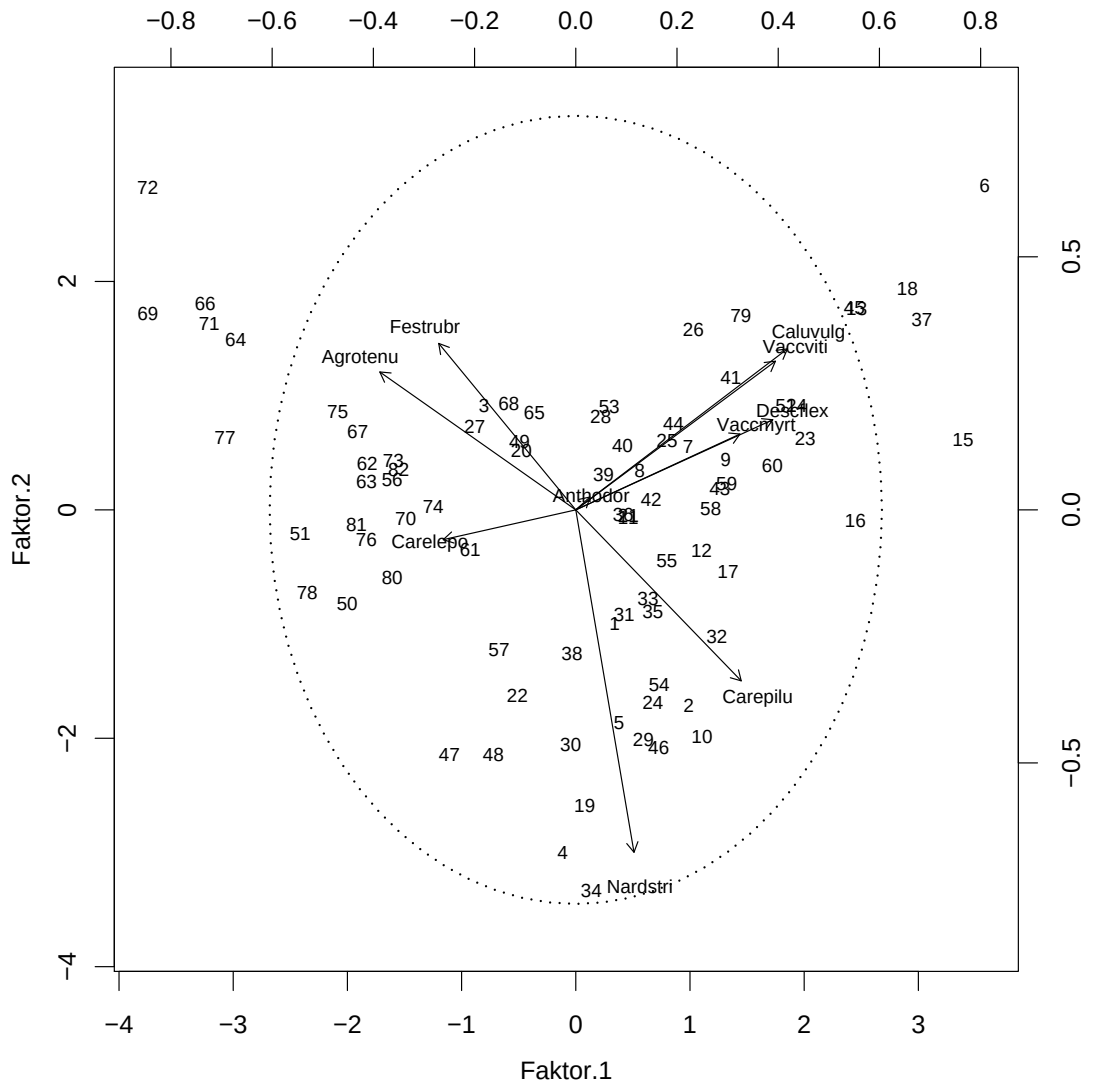

Figure 7.2.b: Biplot in the vegetation study example (RMP variant). The variables were previously standardized.

The mentioned approximation, written for the value of the jth variable for the ith observation is

$$(x_c)_i^{(j)} \approx \underline{S}_i^T \underline{C}_j \ .$$

In order to discuss the additional possible interpretations, we assume that $m = 2$. Then the equation holds exactly. For higher m – only then is the principal component analysis worthwhile – the interpretation is still approximately correct.

For $m = 2$, $\boldsymbol{C}$ is the entire matrix $\widehat{\boldsymbol{B}}^{-1}$ of the principal component analysis, and this is orthogonal. The vector $\underline{C}_j$ is the jth row of $\boldsymbol{B}^{-1}$, and its length is therefore equal to 1. In this case, $\underline{S}_i^T \underline{C}_j$ equals the projection of $\underline{S}_i$ on $\underline{C}_j$. It therefore follows:

If we project a point i onto an arrow j, the distance of the projected point to the zero point approximately reflects the (centered) value of the jth variable for the ith observation.

▷ If we project the point with the label 38 (below the center) in the example onto the arrow "Nardstri", we get 1.19 units. The standardized number of Nardus stricta for this observation amounts to 1.13. For observation 47, we get 1.93 instead of 2.00. The variable Carelepo (Carex leporina) is not as well represented. For observation 42 we get -0.70 instead of -0.36, observation 47 amounts to 1.54 instead of 0.58. ◁

c For $m = 2$ it can be calculated that $\widehat{\text{var}}\langle X^{(j)} \rangle = \sum_k \widehat{\lambda}_k \big(C_k^{(j)}\big)^2$. If we have standardized the variables $X^{(j)}$ to variance 1, then the arrow tips must lie on an ellipse, whose half axes measure $1/\sqrt{\widehat{\lambda}_k}$, $(k = 1, 2)$ and point in the direction the plot's axes.

In general it holds: How precisely the variables are represented in the figure can be seen, in the case of standardized X variables, via the comparison of the arrow lengths with the stated ellipse.

Unfortunately, the ellipse is not shown by the usual programs, and therefore the arrows, taken by themselves, are not easy to interpret.

d Finally, if we imagine the figure compressed in the vertical direction so that the ellipse becomes a circle, then the angle between the "compressed" arrows (approximately) reflects the correlation.

▷ In the example, according to the plot, the species Caluvulg, Vaccviti, Vaccmyrt and Descflex form a group of highly correlated variables. However, if we calculate the correlations, we note that the correlations within this group vary from 0.25 to 0.53, while in the plot, the angles measure about 0 and 12^o, which would correspond to correlations of 1 and 0.97! At least in this example, the angles are not very interpretable. ◁

e **What the Biplot Shows.** In summary, the points $\underline{S}_i$ and arrows $\underline{C}_j$ show– exactly, if $m = 2$ and approximately for $m > 2$ – the following:

(a) The points show the first two principal components.

(b) If the variables were standardized to variance 1, the arrow length/ellipse relationship corresponds to the relationship of represented/total variance (=1) of the variable $X^{(j)}$.

(c) The cosine of the angle between two arrows shows the correlation of the corresponding variables.

(d) The projection of the point i in the direction of an arrow $\underline{C}_j$ reflects the (centered) original observations $x_i^{(j)} - \bar{x}^{(j)}$.

(e) The distances between points $\|\underline{S}_h - \underline{S}_i\|$ correspond to the distances between the observations $\|\underline{x}_h - \underline{x}_i\|$.

f There are variants of the biplot. A common variant represents the arrows so that they conform to the "compression" mentioned above. Then the angle between arrows is directly related to correlation. So that the projection of points on the arrow directions can still be interpreted in the same way, the points must be "back transformed", or stretched in the vertical direction. This means that, instead of the principal components $\underline{Z}_i$ we use the standardized principal components for drawing the points. The distances between points are then no longer equal to the distances of the observation vectors for $m = 2$.

This variant of the biplot goes by the name "column metric preserving (CMP)", while the previous has the name "row metric preserving (RMP)".

7.S S-Functions

a A principal component analysis is carried out with the function

```
> princomp(x, cor = FALSE, scores = TRUE)
```

Argument `x`: Data as matrix or as data frame.

Argument `cor`: If `TRUE`, the analysis is carried out for standardized data – or on the basis of the correlation matrix – otherwise with the covariance matrix.
Notes: The standardization is done with an unusual variant of the standard deviation: The sum of squares is divided by n instead of $n-1$. This causes (usually small) discrepancies in the reported eigenvalues compared to other programs.

The result of `princomp`() can be displayed by `summary`(). With this, standard deviation is given for each principal component, along with the proportion and cumulative proportion of the total variance. If in `summary` we set the argument `loadings=` TRUE, then additionally the so-called loadings are given. We can also get these separately with the function `loadings`.

b **Graphic Output.** `plot(t.r)` displays a so-called scree plot, which is also obtained with the function `screeplot(t.r)`.

The goal of a principal component analysis is often the graphical representation of the first pair of components. The corresponding coordinates are found under `$scores`,

```
> t.r <- princomp(d.abst,cor=TRUE)
> plot(t.r)
> pairs(t.r$scores[,1:3],
  panel=function(x,y) text(x,y,row.names(d.abst)))
```

c There also exists an older function `prcomp` for principal component analysis. It calculates the principal components from the covariance matrix. If we want to use the correlation matrix, the data has to be standardized first. For this, the function `scale` can be called, which uses the usual version of the standard deviation. `prcomp` does not allow the use of formulas and therefore no longer fits with the newer concepts of the S language.

d **Factor Analysis.** The corresponding function is called `factanal`.

e **Biplot.** The R function for generating a biplot is also called thus: `biplot`. It is mostly applied to the results from `princomp` – or from `factanal`.

```
> biplot(princomp(iris[1:50,1:4],cor=TRUE))
```

The function mentioned in the script that draws the reference ellipse and which, in certain details, differs from `biplot`, is called `g.biplot` and is available on the website.