Kapitel 4

Hauptkomponentenanalyse

4.1 Einführung

Die Hauptkomponentenanalyse ist eine variablenorientierte Methode, die, wie die Faktorenanalyse auch, versucht, die Originalvariablen durch eine kleinere Anzahl „dahinter liegender" Variablen zu ersetzen. Die Hauptkomponentenanalyse besteht darin, eine orthogonale Transformation der ursprünglichen Variablen in eine neue Menge unkorrelierter Variablen, die Hauptkomponenten (Englisch: *principal components*) genannt werden. Die Hauptkomponenten werden nacheinander in absteigender Bedeutung konstruiert. Die Hauptkomponenten sind Linearkombinationen der ursprünglichen Variablen. Man hofft, dass wenige der ersten Variablen für den größten Teil der Variation in den Originaldaten verantwortlich sind, so dass die effektive Dimension der Daten reduziert werden kann. Die erste Hauptkomponente wird so konstruiert, dass sie für den größten Teil der Variation verantwortlich ist. Wenn die ersten Hauptkomponenten den größten Teil der Variation erklären, kann die effektive Dimension des Problems verringert werden. Wenn einige der Urspungsvariablen hochkorreliert sind, sagen sie im wesentlichen dasselbe aus und es kann sein, dass es nahezu lineare Einschränkungen in den Variablen gibt. Man hofft dabei, dass sich die ersten Hauptkomponenten gut interpretieren lassen und man so die Daten besser verstehen kann. Eine grafische Darstellung der beiden ersten Hauptkomponenten kann z.B. hilfreich sein, um Cluster in den Daten zu finden. Da bei der Hauptkomponentenanalyse eine Menge korrelierter Variablen in eine Menge unkorrelierter Variablen transformiert wird, ist es naheliegend, vorher zu schauen, ob die Originalvariablen schon von vornherein nahezu unkorreliert sind, denn dann wäre eine Hauptkomponentenanalyse überflüssig. Sie würde nur Variablen finden, die ganz nah an den ursprünglichen Variablen sind, jedoch nach absteigender Varianz geordnet.

Die Hauptkomponentenanalyse ist ein mathematisches Verfahren, bei dem kein statistisches Modell zur Erklärung der Fehlerstruktur verlangt wird. Insbesondere werden keine Annahmen über die Verteilungen der Ursprungsvariablen gemacht. Dennoch ist die Interpretation der Hauptkomponenten einfacher, wenn eine multivariate Normalverteilung vorausgesetzt werden kann.

4.2 Herleitung der Hauptkomponenten

Nehmen wir an, dass $\boldsymbol{Y}^t = (Y_1, \ldots, Y_m)$ ein m-dimensionaler Zufallsvektor mit Erwartungswertvektor $\boldsymbol{\mu}$ und Kovarianzmatrix Σ ist. Unsere Aufgabe ist es, neue Variablen $Z_1, Z_2, \ldots, Z_m$ zu finden, die unkorreliert sind und deren Varianzen mit wachsendem Index $j = 1, \ldots, m$

fallen. Jedes Z_j ist eine Linearkombination der Y, so dass

$$Z_j = a_{1j}Y_1 + a_{2j}Y_2 + \ldots + a_{mj}Y_m = \boldsymbol{a}_j^t \boldsymbol{Y} \ , \tag{4.1}$$

wobei $\boldsymbol{a}_j^t = (a_{1j}, a_{2j}, \ldots, a_{mj})$ ein Vektor von Konstanten ist. Um eine willkürliche Skalierung zu vermeiden, wird der Vektor $\boldsymbol{a}_j$ so normiert, dass $\boldsymbol{a}_j^t \boldsymbol{a}_j = \sum_{k=1}^{m} a_{kj}^2 = 1$. Diese Normierung bewirkt, dass die gesamte Transformation der Daten orthogonal wird, d.h. Abstände im m-dimensionalen Raum bleiben erhalten.

Die erste Hauptkomponente Z_1 wird gefunden, indem man $\boldsymbol{a}_1$ so bestimmt, dass Z_1 die größtmögliche Varianz hat, d.h. wir wählen $\boldsymbol{a}_1$ so, dass $\text{Var}(\boldsymbol{a}_1^t \boldsymbol{Y})$ maximal wird unter der Nebenbedingung $\boldsymbol{a}_1^t \boldsymbol{a}_1 = 1$. Nun gilt nach Gleichung 2.3

$$\text{Var}(Z_1) = \text{Var}(\boldsymbol{a}_1^t \boldsymbol{Y}) = \boldsymbol{a}_1^t \Sigma \boldsymbol{a}_1 \tag{4.2}$$

Damit ist $\boldsymbol{a}_1^t \Sigma \boldsymbol{a}_1$ die zu maximierende Zielfunktion. Wir haben eine Funktion von m Variablen unter einer Nebenbedingung zu maximieren. Dazu benutzen wir die Methode der Lagrange Multiplikatoren: Sei $f(y_1, \ldots, y_m)$ eine unter der Nebenbedingung $g(y_1, \ldots, y_m) = c$ zu maximierende differenzierbare Funktion. Wir definieren die Lagrange-Funktion $L(\boldsymbol{y}, \lambda)$ durch

$$L(\boldsymbol{y}, \lambda) = f(\boldsymbol{y}) - \lambda[g(\boldsymbol{y}) - c]$$

Dabei ist λ der sogenannte Lagrange Multiplikator. Es ist klar, dass der Term in der eckigen Klammer wegen der Nebenbedingung Null ist. Ist $\boldsymbol{y}_0$ ein Extrempunkt unter der gegebenen Nebenbedingung, so hat die Lagrangefunktion dort einen stationären Punkt, d.h. es gilt

$$\frac{\partial f}{\partial y_i}(\boldsymbol{y}_0) - \lambda \frac{\partial g}{\partial y_i}(\boldsymbol{y}_0) = 0 \quad i = 1, \ldots, m \tag{4.3}$$

Diese m Gleichungen und die Nebenbedigung reichen aus, um die Koordinaten der stationären Punkte und den Wert von λ zu bestimmen. Es ist dann weiter zu schauen, ob ein stationärer Punkt ein Maximum, Minimum oder ein Sattelpunkt ist. Gleichung 4.3 kann dann geschrieben werden:

$$\frac{\partial L}{\partial \boldsymbol{y}} = \boldsymbol{0}$$

Dabei bedeutet $\dfrac{\partial L}{\partial \boldsymbol{y}}$ einen Spaltenvektor mit m Komponenten, bestehend aus den partiellen Ableitungen $\dfrac{\partial L}{\partial y_i}$. Wir wenden jetzt die Lagrange-Multiplikatorregel auf unser Problem an und erhalten:

$$L(\boldsymbol{a}_1) = \boldsymbol{a}_1^t \Sigma \boldsymbol{a}_1 - \lambda(\boldsymbol{a}_1^t \boldsymbol{a}_1 - 1)$$

Da

$$\frac{\partial(\boldsymbol{a}_1^t \Sigma \boldsymbol{a}_1)}{\partial \boldsymbol{a}_1} = 2\Sigma \boldsymbol{a}_1$$

folgt

$$\frac{\partial L}{\partial \boldsymbol{a}_1} = 2\Sigma \boldsymbol{a}_1 - 2\lambda \boldsymbol{a}_1$$

Wenn wir dies gleich Null setzen, erhalten wir:

$$(\Sigma - \lambda I)\boldsymbol{a}_1 = \boldsymbol{0} \tag{4.4}$$

Wir haben die $m \times m$-Einheitsmatrix I, d.h.

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

in die obige Gleichung eingefügt, so dass in den Klammern insgesamt eine $m \times m$-Matrix steht.

Nun hat das Gleichungssystem 4.4 genau dann eine nichttriviale, d.h. vom Nullvektor verschiedene Lösung, wenn die Matrix $(\Sigma - \lambda I)$ singulär ist (gleichbedeutend Rang$(\Sigma - \lambda I) < m$). Die Matrix ist genau dann singulär, wenn ihre Determinante Null ist, d.h. wir müssen λ so wählen, dass

$$\det(\Sigma - \lambda I) = 0$$

Das bedeutet: es gibt nur dann eine von Null verschiedene Lösung des Gleichungssystems 4.4, wenn λ ein Eigenwert der Matrix Σ ist.

Wir schieben kurz einige Bemerkungen zu Eigenwerten und Eigenvektoren ein. Sei Σ eine $m \times m$-Matrix. Die Eigenwerte (charakteristischen Wurzeln) sind die Lösungen der Gleichung

$$\det(\Sigma - \lambda I) = 0 \tag{4.5}$$

Diese Gleichung ist ein Polynom der Ordnung m in λ. Die Eigenwerte werden mit $\lambda_1, \lambda_2, \dots, \lambda_m$ bezeichnet.

Wir betrachten die Matrix

$$\Sigma = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

Dann gilt

$$\det(\Sigma - \lambda I) = \det \begin{pmatrix} 1-\lambda & 1/2 \\ 1/2 & 1-\lambda \end{pmatrix} = (1-\lambda)^2 - 1/4 = \lambda^2 - 2\lambda + 3/4$$

Diese Gleichung hat die beiden Lösungen $\lambda_{1,2} = 1 \pm \sqrt{1 - 3/4}$, d.h. $\lambda_1 = 3/2$ und $\lambda_2 = 1/2$.

Zu jedem Eigenwert λ_i gehört ein Vektor $\boldsymbol{c}_i$, der Eigenvektor genannt wird, für den gilt:

$$\Sigma \boldsymbol{c}_i = \lambda_i \boldsymbol{c}_i \tag{4.6}$$

In unserem Beispiel ist also für $\lambda_1 = 3/2$ das Gleichungssystem $(\Sigma - 3/2I)\boldsymbol{c} = \boldsymbol{0}$ zu lösen. Da

$$\Sigma - 3/2I = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} - \begin{pmatrix} 3/2 & 0 \\ 0 & 3/2 \end{pmatrix} = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$$

ist das zu lösende Gleichungssystem

$$\begin{aligned} -0.5c_{11} + 0.5c_{12} &= 0 \\ 0.5c_{11} - 0.5c_{12} &= 0 \end{aligned}$$

Das bedeutet $c_{11} = c_{12}$, d.h jeder Vektor $\boldsymbol{c}_1^t = (c_{11}, c_{11})$ ist eine Lösung.

Für $\lambda_2 = 1/2$ ist das Gleichungssystem $(\Sigma - 1/2I)\boldsymbol{c} = \mathbf{0}$ zu lösen.

Da

$$\Sigma - 1/2I = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} - \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

ist das zu lösende Gleichungssystem

$$\begin{aligned} 0.5c_{21} + 0.5c_{22} &= 0 \\ 0.5c_{21} + 0.5c_{22} &= 0 \end{aligned}$$

Das bedeutet $c_{21} = -c_{22}$, d.h jeder Vektor $\boldsymbol{c}_2^t = (c_{21}, -c_{21})$ ist eine Lösung.

Die Eigenvektoren sind nur bis auf einen konstanten Faktor eindeutig bestimmt. Daher werden sie gewöhnlich so normiert, dass $\boldsymbol{c}_i^t \boldsymbol{c}_i = 1$ gilt. In unserem Beispiel wären also $\boldsymbol{c}_1^t = (1/\sqrt{2}, 1/\sqrt{2})$ und $\boldsymbol{c}_2^t = (1/\sqrt{2}, -1/\sqrt{2})$ normierte Lösungen. Wenn es gleiche Eigenwerte gibt, können die Eigenvektoren so gewählt werden, dass sie orthonormiert sind (orthogonal und normiert). Wir geben hier einige Eigenschaften von Eigenwerten und Eigenvektoren.

a) $\sum_{i=1}^{m} \lambda_i = \text{Spur}(\Sigma)$ (Die Spur einer Matrix ist die Summe der Elemente in der Diagonalen.)

b) $\prod_{i=1}^{m} \lambda_i = \det(\Sigma)$

c) Wenn Σ eine reelle symmetrische Matrix ist, dann sind die Eigenwerte und Eigenvektoren reell.

d) Wenn die Matrix Σ positiv definit ist, dann sind alle Eigenwerte strikt positiv.

e) Wenn Σ positiv semidefinit ist mit $\text{Rang}(\Sigma) = p < m$, dann hat Σ p positive Eigenwerte und m-p Eigenwerte sind gleich Null.

f) Die zu zwei ungleichen Eigenwerten gehörenden normalisierten Eigenvektoren sind orthogonal und damit orthonormiert.

g) Wenn wir eine $m \times m$-Matrix C bilden, die in der i-ten Spalte den normierten Eigenvektor $\boldsymbol{c}_i$ enthält, dann gilt $C^t C = I$ und

$$C^t \Sigma C = \Lambda \tag{4.7}$$

Dabei ist Λ eine Diagonalmatrix mit den Elementen $\lambda_1, \lambda_2, \ldots, \lambda_m$ in der Diagonalen. Man nennt dies die kanonische Darstellung von Σ. Die Matrix C transformiert die quadratische Form von Σ in eine reduzierte quadratische Form, die nur quadratische Terme enthält. Setzen wir $\boldsymbol{x} = C\boldsymbol{y}$, so gilt

$$\begin{aligned} \boldsymbol{x}^t \Sigma \boldsymbol{x} &= \boldsymbol{y}^t C^t \Sigma C \boldsymbol{y} \\ &= \boldsymbol{y}^t \Lambda \boldsymbol{y} \\ &= \lambda_1 y_1^2 + \ldots \lambda_p y_p^2 \end{aligned}$$

Dabei ist $p = \text{Rang}(\Sigma)$. Aus Gleichung 4.7 folgt

$$\Sigma = C\Lambda C^t = \lambda_1 \boldsymbol{c}_1 \boldsymbol{c}_1^t + \ldots + \lambda_p \boldsymbol{c}_p \boldsymbol{c}_p^t \tag{4.8}$$

Kehren wir nun zu unserem Problem der Bestimmung der ersten Hauptkomponente zurück. Wir wollten

$$\text{Var}(Z_1) = \text{Var}(\boldsymbol{a}_1^t \boldsymbol{Y}) = \boldsymbol{a}_1^t \Sigma \boldsymbol{a}_1$$

maximieren. Dazu müssen wir das Gleichungsystem 4.4

$$(\Sigma - \lambda I)\boldsymbol{a}_1 = \boldsymbol{0}$$

lösen. Damit überhaupt eine Lösung exisiert, muss λ ein Eigenwert der Matrix Σ sein. Die Matrix Σ wird i.a. m Eigenwerte haben, die alle nichtnegativ sein müssen, da Σ positiv semidefinit ist. Die Eigenwerte seien $\lambda_1, \lambda_2, \ldots, \lambda_m$. Wir nehmen an, dass sie alle verschieden sind, so dass $\lambda_1 > \lambda_2 > \ldots > \lambda_m \geq 0$. Welchen Eigenwert sollen wir zur Bestimmung der ersten Hauptkomponente nehmen. Nun gilt mit Gleichung 4.4 (und unter Beachtung, dass die Lösung $\boldsymbol{a}_1$ normiert sein sollte, d.h. $\boldsymbol{a}_1^t \boldsymbol{a}_1 = 1$):

$$\text{Var}(\boldsymbol{a}_1^t \boldsymbol{Y}) = \boldsymbol{a}_1^t \Sigma \boldsymbol{a}_1 = \boldsymbol{a}_1^t \lambda I \boldsymbol{a}_1 = \lambda \boldsymbol{a}_1^t I \boldsymbol{a}_1 = \lambda \boldsymbol{a}_1^t \boldsymbol{a}_1 = \lambda \tag{4.9}$$

Da wir diese Varianz maximieren wollen, wählen wir den größten Eigenwert, d.h. λ_1. Nun folgt mit Gleichung 4.4, dass $\boldsymbol{a}_1$ der zu λ_1 gehörige Eigenvektor sein muss.

Jetzt wollen wir die zweite Hauptkomponente bestimmen, d.h. $Z_2 = \boldsymbol{a}_2^t \boldsymbol{Y}$. Zusätzlich zur Normierungsbedingung $\boldsymbol{a}_2^t \boldsymbol{a}_2 = 1$ haben wir als weitere Bedingung, dass Z_2 und Z_1 unkorreliert sein sollen. Nun gilt

$$\text{Cov}(Z_2, Z_1) = \text{Cov}(\boldsymbol{a}_2^t \boldsymbol{Y}, \boldsymbol{a}_1^t \boldsymbol{Y}) = E[\boldsymbol{a}_2^t (\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^t \boldsymbol{a}_1] = \boldsymbol{a}_2^t \Sigma \boldsymbol{a}_1 \tag{4.10}$$

Dies soll Null sein. Da jedoch $\Sigma \boldsymbol{a}_1 = \lambda_1 \boldsymbol{a}_1$, muss $\boldsymbol{a}_2^t \boldsymbol{a}_1 = 0$ sein, d.h, $\boldsymbol{a}_2$ und $\boldsymbol{a}_1$ müssen orthogonal sein. Wir müssen jetzt also die Varianz von Z_2, d.h.

$$\boldsymbol{a}_2^t \Sigma \boldsymbol{a}_2 \quad \text{unter den zwei Nebenbedingungen} \quad \boldsymbol{a}_2^t \boldsymbol{a}_2 = 1 \text{ und } \boldsymbol{a}_2^t \boldsymbol{a}_1 = 0$$

maximieren. Dazu wählen wir zwei Lagrangemultiplikatoren, λ und δ und betrachten dann die Funktion

$$L(\boldsymbol{a}_2) = \boldsymbol{a}_2^t \Sigma \boldsymbol{a}_2 - \lambda(\boldsymbol{a}_2^t \boldsymbol{a}_2 - 1) - \delta \boldsymbol{a}_2^t \boldsymbol{a}_1$$

In den stationären Punkten muss gelten:

$$\frac{\partial L}{\partial \boldsymbol{a}_2} = 2(\Sigma - \lambda I)\boldsymbol{a}_2 - \delta \boldsymbol{a}_1 = \boldsymbol{0} \tag{4.11}$$

Wir multiplizieren dies von links mit $\boldsymbol{a}_1^t$ und erhalten

$$2\boldsymbol{a}_1^t \Sigma \boldsymbol{a}_2 - 2\lambda \boldsymbol{a}_1^t \boldsymbol{a}_2 - \delta \boldsymbol{a}_1^t \boldsymbol{a}_1 = 2\boldsymbol{a}_1^t \Sigma \boldsymbol{a}_2 - \delta = \boldsymbol{0} \ ,$$

da $\boldsymbol{a}_1^t \boldsymbol{a}_2 = 0$ und $\boldsymbol{a}_1^t \boldsymbol{a}_1 = 1$.

Aber aus Gleichung 4.10 folgt, dass $\boldsymbol{a}_2^t \Sigma \boldsymbol{a}_1 = \boldsymbol{a}_1^t \Sigma \boldsymbol{a}_2 = 0$ sein muss (das war die Kovarianz von Z_2 und Z_1). Deshalb muss $\delta = 0$ sein und damit wird aus Gleichung 4.11

$$(\Sigma - \lambda I)\boldsymbol{a}_2 = 0$$

Diese Gleichung ist analog zu Gleichung 4.4 für die erste Hauptkomponente. Mit ähnlichen Überlegungen folgern wir jetzt, dass wir für λ den zweitgrößten Eigenwert von Σ wählen und für $\boldsymbol{a}_2$ den zugehörigen Eigenvektor.

Mit den gleichen Argumenten findet man heraus, dass die j-te Hauptkomponente der Eigenvektor ist, der zum j-ten Eigenwert (der Größe nach geordnet) von Σ gehört. Wenn einige der Eigenwerte identisch sind, ist die Bestimmung der Eigenvektoren nicht eindeutig. Dann kommt es nur darauf an, dass die Eigenvektoren, die zu mehrfachen Eigenwerten gehören, orthogonal gewählt werden. Wir bezeichnen die $m \times m$-Matrix, in deren Spalten die Eigenvektoren stehen, mit A:

$$A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]$$

Den Vektor der Hauptkomponenten bezeichnen wir mit $\boldsymbol{Z}$. Dann gilt:

$$\boldsymbol{Z} = A^t \boldsymbol{Y} \tag{4.12}$$

Die $m \times m$-Kovarianzmatrix von $\boldsymbol{Z}$ bezeichnen wir mit Λ. Da die Varianzen der Hauptkomponenten gleich den Eigenwerten sind und die Hauptkomponenten unkorreliert sind, ist sie gegeben durch:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & \lambda_m \end{pmatrix} \tag{4.13}$$

Wegen Gleichung 2.5 gilt auch $\text{Var}(\boldsymbol{Z}) = \text{Var}(A^t \boldsymbol{Y}) = A^t \Sigma A$, so dass

$$\Lambda = A^t \Sigma A \tag{4.14}$$

Diese Gleichung stellt also eine Beziehung her zwischen der Kovarianzmatrix von $\boldsymbol{Y}$ und $\boldsymbol{Z}$. Da A eine orthogonale Matrix ist mit $AA^t = I$, kann Gleichung 4.14 auch in der Form

$$\Sigma = A \Lambda A^t \tag{4.15}$$

geschrieben werden.

Die Eigenwerte können als Varianzen der entsprechenden Hauptkomponenten interpretiert werden. Nun ist die Summe dieser Varianzen

$$\sum_{i=1}^{m} \text{Var}(Z_i) = \sum_{i=1}^{m} \lambda_i = \text{Spur}(\Lambda)$$

Nun ist aber:

$$\text{Spur}(\Lambda) = \text{Spur}(A^t \Sigma A) = \text{Spur}(\Sigma A A^t) = \text{Spur}(\Sigma) = \sum_{i=1}^{m} \text{Var}(Y_i)$$

Dabei wurde benutzt, dass die $\text{Spur}(BC) = \text{Spur}(CB)$, vorausgesetzt, dass BC eine quadratische Matrix ist.

Wir haben jetzt das wichtige Resultat, dass die Summe der Varianzen der ursprünglichen Variablen und die Summe der Varianzen der Hauptkomponenten identisch sind. So kann man sagen, dass z.B. die i-te Hauptkomponente den Anteil $\lambda_i / \sum_{j=1}^{m} \lambda_j$ an der Totalvariation der ursprünglichen Variablen hat. Oder man sagt, dass die ersten p Hauptkomponenten $(\sum_{j=1}^{p} \lambda_j / \sum_{j=1}^{m} \lambda_j) \cdot 100\%$ der Totalvariation erklären.

Beispiel: Wir betrachten den Datensatz `teil01.frame`, der aus allen vollständigen Datensätzen der drei Variablen `Groesse, Schuh, Gewicht` zum Fragebogen in Statistik I besteht. Wir berechnen zunächst die Kovarianzmatrix:

```
cov(teil01.frame)
            Groesse      Schuh    Gewicht
 Groesse   92.34441  24.387788  103.94037
 Schuh     24.38779   8.922498   32.98380
 Gewicht  103.94037  32.983795  195.32201
```

Mit der Funktion `eigen` berechnen wir dann die Eigenwerte und Eigenvektoren:

```
eigen(cov(teil01.frame))
 $values
[1] 266.322975 28.256195 2.009739


$vectors
            Gewicht       Schuh     Groesse
 Groesse  0.5226376  -0.8312910   0.1892224
 Schuh    0.1569001  -0.1243717  -0.9797520
 Gewicht  0.8379930   0.5417443   0.0654283
```

In der Ausgabe erhalten wir zunächst unter `$values` die der Größe nach geordneten Eigenwerte der Kovarianzmatrix. In unserer Notation haben wir also auf zwei Nachkommastellen gerundet:

$$\lambda_1 = 266.32 \quad \lambda_2 = 28.26 \quad \lambda_3 = 2.01$$

Die unter `$vectors` stehende Matrix enthält in den Spalten die Eigenvektoren, d.h. in unserer Notation ist auf vier Nachkommastellen gerundet:

$$\begin{aligned} \boldsymbol{a}_1^t &= (0.5226, 0.1569, 0.8380) \\ \boldsymbol{a}_2^t &= (-0.8313, -0.1244, 0.5417) \\ \boldsymbol{a}_3^t &= (0.1892, -0.9798, 0.0654) \end{aligned}$$

Die Hauptkomponenten sind dann also

HK1 = + 0.5226 Groesse + 0.1569 Schuh + 0.8380 Gewicht
HK2 = - 0.8313 Groesse - 0.1244 Schuh + 0.5417 Gewicht
HK3 = + 0.1892 Groesse - 0.9798 Schuh + 0.0654 Gewicht

Wir können die Hauptkomponenten in **R** (abgesehen von speziellen Funktionen) auf die folgende Weise berechnen:

```
Eigenvektoren<-eigen(cov(teil01.frame))$vectors

HK1<-as.matrix(teil01.frame)%*%Eigenvektoren[,1]
HK2<-as.matrix(teil01.frame)%*%Eigenvektoren[,2]
HK3<-as.matrix(teil01.frame)%*%Eigenvektoren[,3]
```

Etwas einfacher geht es mit dem folgenden Befehl:

```
as.matrix(teil01.frame)%*%Eigenvektoren
```

Wir wollen jetzt die Varianzen und Kovarianzen der Hauptkomponenten zur Kontrolle berechnen. Wenn wir richtig gerechnet haben, müssten die Varianzen gleich den Eigenwerten sein und die Kovarianzen müssten Null sein, da die Hauptkomponenten unkorreliert sind.

```
cov(cbind(HK1,HK2,HK3))
        [,1]          [,2]          [,3]
[1,] 2.663230e+02 3.356920e-14 -1.334241e-15
[2,] 3.356920e-14 2.825619e+01 1.217791e-14
[3,] -1.334241e-15 1.217791e-14 2.009739e+00
```

Bis auf Rechenungenauigkeiten stimmen unsere Resultate. Wir wollen jetzt die durch die einzelnen Hauptkomponenten erklärten Anteile der Varianz bestimmen. Wir schreiben zunächst die Eigenwerte in den Vektor `Eigenwerte`:

```
Eigenwerte<-eigen(cov(teil01.frame))$values
Eigenwerte
[1] 266.322975 28.256195 2.009739
```

Wir lassen uns jetzt die gerundeten Anteile an der Totalvariation ausdrucken:

```
print(round(Eigenwerte/sum(Eigenwerte)*100,digits=2))
[1] 89.80 9.53 0.68
```

Die erste Hauptkomponente erkärt also 89.80% der Totalvariation, die zweite 9.53%, die dritte 0.68%. Die kumulierten Anteile sind:

```
print(round(cumsum(Eigenwerte)/sum(Eigenwerte)*100,digits=2))
[1] 89.80 99.32 100.00
```

Die ersten beiden Hauptkomponenten erkären also zusammen 99.32% der Totalvariation.

Schätzung der Hauptkomponenten: Wir haben in dem vorigen Beispiel eine geschätzte Kovarianzmatrix verwendet, obwohl wir die Hauptkomponenten für eine gegebene Kovarianzmatrix Σ hergeleitet hatten. Im allgemeinen wird Σ nicht bekannt sein und wir werden Σ durch S schätzen. Wir hatten S in Gleichung 3.2 definiert. Die Herleitung der Hauptkomponenten erfolgt genauso wie bisher. Die Hauptkomponenten werden mit den Eigenvektoren von S berechnet. Streng genommen müssten wir die Eigenwerte von S mit $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$ und die zugehörigen Eigenvektoren mit $\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{a}}_2, \ldots, \hat{\boldsymbol{a}}_m$. Dabei wurde angenommen, dass die Eigenwerte der Größe nach geordnet sind. Da die geschätzte Kovarianzmatrix positiv semidefinit ist, sind alle Eigenwerte nichtnegativ und sind gleich den geschätzten Varianzen der entsprechenden Hauptkomponenten.

Verwendung der Korrelationsmatrix anstelle der Kovarianzmatrix: Häufig wird anstelle der Kovarianzmatrix Σ die Korrelationsmatrix P zur Berechnung der Hauptkomponenten verwendet. Das bedeutet, dass alle Zufallsvariablen standardisiert worden sind und die Varianz für alle Zufallsvariablen gleich 1 ist, d.h. alle Variablen haben das gleiche Gewicht. Die Eigenwerte und Eigenvektoren von Σ und P sind nicht identisch. In der Korrelationsmatrix sind alle Diagonalelemente 1. Damit ist die Summe der Diagonalelemente oder die Summe der Varianzen der standardisierten Variablen gleich m. Damit ist die Summe der Eigenwerte von P ebenfalls gleich m. Der Beitrag der j-ten Hauptkomponente zur Totalvariation ist somit λ_j/m.

Wir verwenden das frühere Beispiel, um die Hauptkomponenten aus der geschätzten Korrelationsmatrix herzuleiten.

```
cor(teil01.frame)
           Groesse      Schuh    Gewicht
Groesse  1.0000000  0.8496184  0.7739330
Schuh    0.8496184  1.0000000  0.7900996
Gewicht  0.7739330  0.7900996  1.0000000
```

Wir berechnen die Eigenwerte und Eigenvektoren mit der Funktion `eigen`:

```
eigen(cor(teil01.frame))
$values
[1] 2.6095464 0.2410677 0.1493859

$vectors
           Gewicht       Schuh      Groesse
Groesse  0.5809216  -0.4715234   0.66347255
Schuh    0.5846446  -0.3254109  -0.74316787
Gewicht  0.5663222   0.8196179   0.08663539
```

Damit sind die Eigenwerte:

$$\lambda_1 = 2.61 \quad \lambda_2 = 0.24 \quad \lambda_3 = 0.15$$

Man beachte, dass die Summe der Eigenwerte 3 ist.

Die Eigenvektoren sind jetzt:

$$\begin{aligned} \boldsymbol{a}_1^t &= (0.5809, 0.5846, 0.5663) \\ \boldsymbol{a}_2^t &= (-0.4715, -0.3254, 0.8196) \\ \boldsymbol{a}_3^t &= (0.6635, -0.7432, 0.0866) \end{aligned}$$

Zur Berechnung der Hauptkomponenten müssten alle Variablen durch ihre Standardabweichung dividiert werden. Die Standardabweichungen erhalten wir durch den folgenden Befehl:

```
sqrt(diag(cov(teil01.frame)))
Groesse   Schuh   Gewicht
9.609600 2.987055 13.975765
```

Dabei wird zunächst mit `cov` die Kovarianzmatrix berechnet, dann zieht der Befehl `diag` die Diagonale, also die Varianzen aus der Kovarianzmatrix heraus. Mit `sqrt` werden dann die Quadratwurzeln aus den Varianzen gezogen, d.h. die Standardabweichungen berechnet.

Wir speichern die Standardabweichungen in dem Vektor `stand`:

```
stand<-sqrt(diag(cov(teil01.frame)))
```

Jetzt bilden wir eine Diagonalmatrix, die in der Diagonalen die Inversen der Standardabweichungen enthält.

$$K = \begin{pmatrix} 1/s_1 & 0 & \dots & 0 \\ 0 & 1/s_2 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1/s_m \end{pmatrix}$$

Die standardisierte Datenmatrix erhalte ich dann durch $X^* = XK$. In **R** erreichen ich dies durch den Befehl:

```
teil01stand.frame<-as.matrix(teil01.frame)%*%diag(1/stand)
```

Dabei bildet `diag(1/stand)` eine Diagonalmatrix mit den Werten `1/stand` in der Diagonalen. Wir können uns überzeugen, dass die Variablen jetzt auf Varianz 1 normiert sind, indem wir von diesen Werten die Kovarianzmatrix ausrechnen:

```
cov(as.matrix(teil01.frame)%*%diag(1/stand))
1.0000000 0.8496184 0.7739330
0.8496184 1.0000000 0.7900996
0.7739330 0.7900996 1.0000000
```

Die Hauptkomponenten können mit dem Befehl

`teil01stand.frame%*%eigen(cor(teil01.frame))$vectors` berechnet werden.

Die Anteile in Prozent der einzelnen Hauptkomponenten an der Totalvariation berechnen wir mit dem Befehl:

```
eigen(cor(teil01.frame))$values*100/3
86.984879 8.035589 4.979531
```

Die kumierten Anteile sind:

```
cumsum(eigen(cor(teil01.frame))$values*100)/3
86.98488 95.02047 100.00000
```

Die erste Hauptkomponente erklärt also 87%, die zweite 8%, die dritte 5% der Totalvariation. Die beiden ersten Hauptkomponenten erklären zusammen 95% der Variation.

Um einen Eindruck zu vermitteln, was die letzte Ausage bedeutet, drucken wir einmal die ersten 20 Werte der Hauptkomonenten aus. Achten Sie dabei auf die unterschiedliche Variation.

```
round((teil01stand.frame%*%eigen(cor(teil01.frame))
    $vectors)[1:20,], digits=2)
  HK1   HK2   HK3
21.64 -8.51 1.83
22.32 -9.37 1.40
19.12 -9.22 1.96
18.82 -8.85 2.01
20.42 -8.68 2.30
23.97 -9.29 2.21
22.12 -9.18 2.50
22.28 -9.16 1.27
22.63 -9.89 1.86
19.91 -9.19 3.02
```

```
22.11 -9.85 2.34
19.22 -8.54 2.19
24.42 -9.07 2.91
21.53 -9.53 3.09
19.70 -9.19 1.95
22.14 -9.01 2.44
23.35 -9.02 2.95
19.56 -9.13 2.27
20.88 -8.82 2.27
23.78 -9.46 2.57
```

4.3 Weiteres zur Hauptkomponentenanalyse

Mittelwertkorrektur: Bisher haben wir die Hauptkomponentenanalyse auf die Ursprungsdaten angewendet. Das bedeutet auch, dass die zugrunde liegenden Zufallsvariablen $Y_1, Y_2, \ldots, Y_m$ einen von Null verschiedenen Erwartungswertvektor haben. Damit haben auch die Hauptkomponenten eine von Null verschiedene Erwartung. Die Beziehung zwischen den Hauptkomponenten $\boldsymbol{Z}$ und den Originalvariablen $\boldsymbol{Y}$ war in Gleichung 4.12 gegeben:

$$\boldsymbol{Z} = A^t \boldsymbol{Y}$$

Es ist zweckmäßig den Erwartungswertvektor abzuziehen, d.h.

$$\boldsymbol{Z} = A^t(\boldsymbol{Y} - \boldsymbol{\mu})$$

Die Daten werden also zunächst verschoben, bevor die Transformation in die Hauptkomponenten durchgeführt wird. Bekanntlich bleibt die Kovarianzmatrix unverändert, wenn die Daten nur verschoben werden, d.h. die Eigenwerte und Eigenvektoren ändern sich nicht.

Wir müssen unsere Datenmatrix so verändern:

$$(X - \mathbf{1}\hat{\boldsymbol{\mu}}^t)$$

Dabei ist $\mathbf{1}$ ein Spaltenvektor mit n Einsen.

In **R** erreichen wir das auf die folgende Weise. Wir schätzen zunächst den Erwartungswertvektor durch:

```
mitte<-apply(teil01.frame,2,mean)
```

Um in jeder Zeile den Mittelwertvektor abziehen zu können, bilden wir den Eins-Vektor:

```
eins<-rep(1,nrow(teil01.frame))
```

Jetzt ziehen wir von jeder Beobachtung den Mittelwert aller Beobachtungen zu derselben Variablen ab:

```
teil01.frame-eins%*%t(mitte)
```

Die Hauptkomponenten nach Mittelwertkorrektur erhalten wir dann so:

```
(as.matrix(teil01.frame)-eins%*%t(mitte))%*%Eigenvektoren
```

```
round((as.matrix(teil01.frame)-eins%*%t(mitte))%*%Eigenvektoren,
digits=2)[1:10,]
```

```
    Gewicht  Schuh  Groesse
 1     2.55   8.89    -0.90
 2     1.45   1.65    -3.23
 3   -26.77   2.53    -0.71
 4   -26.81   6.23    -0.17
 5    -8.70   5.80     0.75
 6    21.49  -1.99    -0.57
 7     5.26  -1.05     0.79
 8     2.08   4.40    -3.47
 9     2.59  -5.80    -2.10
10   -14.47  -2.16     2.87
```

Die Beschriftung über den Spalten ist bei dieser Berechnungsmethode nicht ganz korrekt. Es werden die Namen aus dem Data-Frame `teil01.frame` bei allen Rechenoperationen beibehalten.

Hauptkomponentenanalyse mit R-Funktion `prcomp`

Mit dem Befehl

```
prcomp(teil01.frame)
```

erhalten wir die folgende Ausgabe:

```
Standard deviations:
16.319405 5.315656 1.417653

Rotation:
                PC1         PC2         PC3
 Groesse  0.5226376   0.8312910   0.1892224
 Schuh    0.1569001   0.1243717  -0.9797520
 Gewicht  0.8379930  -0.5417443   0.0654283
```

Vergleichen wir diese Ausgabe mit der Ausgabe der Eigenwerte und Eigenvektoren (siehe S. 38), so stellen wir fest, dass hier unter `Rotation` die Eigenvektoren stehen. Unter `Standard deviations` stehen die Standardabweichungen der Hauptkomponenten. Da die Varianzen der Hauptkomponenten gleich den Eigenwerten sind, sind die Standardabweichungen gerade die Quadratwurzeln aus den Eigenwerten.

In der Hilfe zu `prcomp` erfahren wir die weiteren Argumente dieser Funktion:

```
prcomp(x,retx=TRUE,center=TRUE,scale.=FALSE,tol=NULL)
```

Das Argument `center=True` bedeutet, dass standardmäßig die Daten am Mittelwert zentriert werden, d.h. die Daten werden nach Null verschoben (ohne Änderung der Kovarianzstruktur). Es geschieht also genau das, was wir gerade besprochen haben (siehe S. 42). Setzt man das Argument `scale.=TRUE` so werden die Daten so skaliert, dass sie danach Varianz 1 haben, d.h. es wird mit der Korrelationsmatrix gerechnet. Falls das Argument `retx=TRUE` (`retx` steht für *return x* werden die rotierten Daten, d.h, die Hauptkomponenten ausgegeben. Dennoch kriegt man sie nicht ohne weiteres zu sehen. Die Ergebnisse der Berechnungen werden nämlich in eine Liste geschrieben. Diese Liste hat in diesem Fall die Komponenten `sdev, rotation, x`. Die Namen dieser Komponenten erfahren sie in der Hilfe unter `Value`. Nicht alle Komponenten der Liste werden in der Standardausgabe angezeigt. Am besten speichern Sie diese Liste in einem **R**-Objekt, z.B.:

```
aus<-prcomp(teil01.frame)
```

Sie erhalten dann keine Ausgabe auf dem Bildschirm. Die einzelnen Komponenten der Liste erhalten Sie, indem Sie hinter `aus` nach einem Dollarzeichen den Namen der Komponente angeben:

```
aus$sdev
16.319405 5.315656 1.417653
```

Es werden die Standardabweichungen der Hauptkomponenten, also die Wurzeln aus den Eigenwerten angegeben.

Mit `aus$rotation` werden die Eigenvektoren ausgegeben, während wir mit `aus$x` die Hauptkomponenten erhalten. Wir geben hier nur die ersten fünf gerundeten Werte wieder.

```
      PC1   PC2    PC3
1    2.55  8.89  -0.90
2    1.45  1.65  -3.23
3  -26.77  2.53  -0.71
4  -26.81  6.23  -0.17
5   -8.70  5.80   0.75
```

Durch einen Vergleich mit der Tabelle auf Seite 43 können Sie sich überzeugen, dass dies die Hauptkomponenten nach Mittelwertkorrektur sind. Das Argument `center` ist standardmäßig auf `TRUE` gesetzt.

Hauptkomponentenanalyse mit R-Funktion `princomp`

In der Hilfe zu `princomp` sehen wir unter

```
Usage:
    princomp(x, cor = FALSE, scores = TRUE, covmat = NULL,
       subset = rep(TRUE, nrow(as.matrix(x))))
```

Die Argumente dieser Funktion sind also:

- `x:` Datenmatrix
- `cor:` Falls `cor=TRUE` wird die Korrelationsmatrix verwendet
- `scores:` Falls `scores=TRUE` werden die Hauptkomponenten in die Ausgabeliste geschrieben
- `covmat:` Anstelle der Datenmatrix kann auch eine mit anderen **R**-Funktionen berechnete Kovarianzmatrix eingegeben werden.
- `subset:` Es kann eine Teilmenge der Variablen der Datenmatrix ausgewählt werden.

Die Standardausgabe der Funktion `princomp` ist die folgende:

```
princomp(teil01.frame)
Call:
princomp(x = teil01.frame)
Standard deviations:
   Comp.1   Comp.2   Comp.3
16.283260 5.303882 1.414513
3 variables and 226 observations.
```

(Vergleichen Sie die Ergebnisse mit denen der Funktion `prcomp`, so fallen geringfügige Abweichungen in den Standardabweichungen auf, obwohl mit exakt demselben Datensatz gearbeitet wurde. Berechnet man die Eigenwerte der Kovarianzmatrix von `teil01.frame` und zieht daraus die Quadratwurzeln, so ergeben sich dieselben Werte wie mit `prcomp`. Es ist zu vermuten, dass `princomp` intern mit gerundeten Werten rechnet.)

Wie bei `prcomp` werden hier die weiteren Ergebnisse in eine Liste geschrieben, die in diesem Fall die folgenden Elemente hat. In der Hilfe erfahren wir:

`'princomp' returns a list with class "princomp" containing the following components:`

`sdev: the standard deviations of the principal components.` *Standardabweichungen, Quadratwurzeln aus den Eigenwerten*

`loadings: the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors).` *Matrix der Eigenvektoren*

`center: the means that were subtracted.` *Mittelwerte der Variablen*

`scale: the scalings applied to each variable.` *Normierungsfaktor, falls z.B.* `cor=T`

`n.obs: the number of observations.` *n Anzahl der Beobachtungen*

`scores: if 'scores = TRUE', the scores of the supplied data on the principal components.` *Hauptkomponenten nach Mittelwertkorrektur*

`call: the matched call.` *Funktionsaufruf*

Die Rücktransformation:

Wir haben auf S. 42 gesehen, dass zwischen den Hauptkomponenten und den Originalvariablen nach Mittelwertkorrektur der Zusammenhang

$$\boldsymbol{Z} = A^t(\boldsymbol{Y} - \boldsymbol{\mu}) \tag{4.16}$$

bestand. Bei vorliegenden Daten ist der Erwartungswertvektor $\boldsymbol{\mu}$ durch den Mittelwertvektor in der Stichprobe zu ersetzen. Wir geben jetzt die inverse Transformation an, die aus den Hauptkomponenten wieder die ursprünglichen Variablen berechnet.

$$\boldsymbol{Y} = A\boldsymbol{Z} + \boldsymbol{\mu} \tag{4.17}$$

Kleine Eigenwerte oder Eigenwerte gleich Null: Wenn die Originalvariablen linear abhängig sind, werden einige Eigenwerte der Kovarianzmatrix Σ Null sein. Die Dimension des Raumes, der die Beobachtungen enthält ist gleich dem Rang von Σ und dieser ist $m - k$, wenn k die Anzahl der Eigenwerte ist, die Null sind. Das bedeutet: es gibt k unabhängige lineare Beziehungen zwischen den Variablen.

Die Existenz von exakten linearen Abhängigkeiten ist selten. Es ist wichtiger, beinahe lineare Beziehungen zwischen den Variablen aufzudecken. Wenn der kleinste Eigenwert λ_m beinahe Null ist, dann wird die m-te Hauptkomponente $\boldsymbol{a}_m^t$ beinahe konstant sein, denn ihre Varianz ist λ_m. Damit ist die Dimension des Datenraumes beinahe kleiner als m. Wenn man die letzten Eigenwerte als klein betrachtet und damit die Dimension auf p beschränkt, so kann der Anteil der ersten p Hauptkomponenten an der Totalvariation als Gütekriterium dieser Maßnahme betrachtet werden. Dieser Anteil war $\sum_{i=1}^{p} \lambda_i / \sum_{i=1}^{m} \lambda_i$. Die zu kleinen Eigenwerten

gehörenden Hauptkomponenten sind Variablen, die für alle Merkmalsträger fast denselben Wert haben. Wir demonstrieren, dass an einem Beispiel aus Rinne (2000, S. 21). Sie finden die Daten als **R**-Objekt im Internet. Bei den Daten handelt es sich um die vier sogenannten Konvergenzkriterien, an denen die Fitness eines EU-Mitgliedstaates für den Eintritt in die EWU (Europäische Währungsunion) gemessen wurde. Der Datensatz heißt in **R**

```
EWU.frame
    Staat   X1   X2    X3     X4  Beitritt
 1      B  1.4  5.7   2.1  122.2      1957
 2     DK  1.9  6.2  -0.7   65.1      1973
 3      D  1.4  5.6   2.7   61.3      1957
 4    FIN  1.3  5.9   0.9   55.8      1995
 5      F  1.2  5.5   3.0   58.0      1957
 6     GR  5.2  9.8   4.0  108.0      1981
 7    IRL  1.2  6.2  -0.9   66.3      1973
 8      I  1.8  6.7   2.7  121.6      1957
 9      L  1.4  5.6  -1.7    6.7      1957
10     NL  1.8  5.5   1.4   72.7      1957
11      A  1.1  5.6   2.5   66.1      1995
12      P  1.8  6.2   2.5   62.0      1986
13      S  1.9  6.5   0.8   76.6      1995
14      E  1.8  6.3   2.6   68.8      1986
15     GB  1.8  7.0   1.9   53.4      1973
```

Die Bedeutung der Variablen ist

X1 Inflationsrate 1997 in %

X2 langfristiger Zinssatz 1997 in %

X3 öffentliche Neuverschuldung 1997 in % des BIP 1997

X4 öffentlicher Schuldenstand 1997 in % des BIP 1997

Wir berechnen die Eigenwerte der Kovarianzmatrix:

```
EWU.eigen<-eigen(cov(EWU.frame[,2:5]))$values
round(EWU.eigen,digits=2)
836.17 2.15 1.38 0.06
```

Insbesondere der letzte Eigenwert ist sehr klein.

Wir schauen uns die Hauptkomponenten an:

```
round(princomp(EWU.frame[,2:5])$scores,digits=2)
```

```
    Comp.1  Comp.2  Comp.3  Comp.4
 1   51.19    1.97    0.59    0.14
 2   -5.94    1.49   -1.48    0.18
 3   -9.65   -0.68    1.37    0.13
 4  -15.20    0.37    0.09   -0.11
 5  -12.94   -0.89    1.75    0.04
 6   37.18   -3.68   -2.26    0.15
 7   -4.76    2.00   -1.28   -0.33
 8   50.63    0.84    0.23   -0.25
 9  -64.35    0.60   -1.25    0.17
10    1.71    0.58    0.31    0.53
11   -4.86   -0.21    1.36   -0.09
12   -8.94   -0.97    0.70    0.03
13    5.60    0.65   -0.72   -0.04
14   -2.14   -0.84    0.68   -0.04
15  -17.54   -1.23   -0.10   -0.51
```

Wir sehen, dass die letzte Komponente kaum variiert. Die Standardabweichungen sind:

```
round(princomp(EWU.frame[,2:5])$sdev,digits=2)
Comp.1 Comp.2 Comp.3 Comp.4
27.94   1.42   1.13   0.24
```

Die kumulierten Anteile der Hauptkomponenten an der Gesamtvariation sind:

```
round(cumsum(EWU.eigen)*100/sum(EWU.eigen),digits=2)
99.57 99.83 99.99 100.00
```

Hier würde man wohl nur die erste Hauptkomponente verwenden, da sie bereits 99.57% der Variation erklärt.

Man beachte aber, dass dies anders aussieht, wenn man die Korrelationsmatrix zur Berechnung der Hauptkomponenten heranzieht, siehe Rinne (2000, S. 107).

```
ei<-eigen(cor(EWU.frame[,2:5]))$values
round(ei,digits=4)
2.5639 0.9337 0.4444 0.0580
round(cumsum(ei)/sum(ei)*100,digits=2)
64.10 87.44 98.55 100.00
```

Jetzt erklären die beiden ersten Hauptkomponenten 87.44%, die drei ersten 98.55% der Variation.

Orthogonalität: Wir haben bereits an früherer Stelle (S. 33) gesagt, dass bei der Transformation in Hauptkomponenten die Abstände im m-dimensionalen Raum erhalten bleiben. Wir wollen dies jetzt zeigen. Sei A die Matrix der Eigenvektoren und X sei eine mittelwertbereinigte Datenmatrix, dann erhalten wir die Hauptkomponentenmatrix durch

$$Z = XA$$

und es gilt dann

$$ZZ^t = XAA^tX^t = XX^t$$

Die Matrizen ZZ^t und XX^t enthalten in der i-ten Zeile und j-ten Spalte die Skalarprodukte der Merkmale für die Merkmalsträger i und j. Insbesondere stehen in der Diagonalen die Summen der Quadrate der Merkmale für die einzelnen Merkmalsträger, d.h. die quadrierten

Normen der Beobachtungsvektoren für die Merkmalsträger. Diese verändern sich also nicht bei der Transformation auf die Hauptkomponenten.

Komponentenladungen: Wir hatten die Eigenvektoren so normiert, dass $\boldsymbol{a}_j^t \boldsymbol{a}_j = 1$. Eine andere Möglichkeit ist es, anstelle $\boldsymbol{a}_j$ den Vektor $\boldsymbol{a}_j^* = \lambda_j^{1/2} \boldsymbol{a}_j$ zu betrachten. Die Summe der Quadrate dieser Vektoren ist dann nicht 1, sondern gleich dem Eigenwert λ_j, da $\boldsymbol{a}_j^{*t} \boldsymbol{a}_j^* = \lambda_j \boldsymbol{a}_j^t \boldsymbol{a}_j = \lambda_j$.

Setzen wir jetzt $C = [\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \ldots, \boldsymbol{a}_m^*]$, so gilt $C = A\Lambda^{1/2}$ (zur Erinnerung. Die Matrix A enthielt in den Spalten die Eigenvektoren.) Mit Gleichung 4.15 ($\Sigma = A\Lambda A^t$) folgt dann $\Sigma = CC^t$. Die Matrix C ist jetzt so, dass die Koeffizienten der wichtigeren Komponenten i.a. größer sind als die der unbedeutenderen. Die so skalierten Vektoren haben zwei direkte Interpretationen. Die erste ist, dass man sie in Analogie zu den Faktorladungen der Faktorenanalyse als Komponentenladungen bezeichnen kann (jedoch werden in den **R**-Programmen die Koeffizienten der Hauptkomponenten als Ladungen bezeichnet, z.B. in `princomp`). Dazu setze man $\boldsymbol{Z}^* = \Lambda^{-1/2} \boldsymbol{Z}$, d.h. die Hauptkomponenten werden so skaliert, dass sie Varianz 1 haben. Dann haben wir unter der Annahme, dass der Erwartungswertvektor Null ist, für die inverse Transformation $\boldsymbol{Y} = A\boldsymbol{Z}$ die neue Form

$$\boldsymbol{Y} = A\Lambda^{1/2} \boldsymbol{Z}^* = C\boldsymbol{Z}^* \; .$$

In der Faktorenanalyse nimmt man ein Modell $\boldsymbol{Y} = \Lambda \boldsymbol{f} + \boldsymbol{e}$ an. Dabei ist $\boldsymbol{f}^t = [f_1, f_2, \ldots, f_p]$ der Vektor der Faktoren und Λ ist eine $m \times p$-Matrix der Faktorladungen. (Man beachte, dass diese Matrix Λ im Modell der Faktorenanalyse nicht mit der Diagonalmatrix Λ der Eigenwerte zu verwechseln ist!) Der Vektor $\boldsymbol{e}^t = [e_1, e_2, \ldots, e_m]$ enthält die Fehler, die die Restvariation der einzelnen Variablen beschreiben. In Analogie zu den Faktorladungen spricht man dann bei der Matrix C von Komponentenladungen. Die auf Varianz 1 standardisierten Hauptkomponenten werden auch als Hauptfaktoren bezeichnet.

Die zweite Interpretation von C ergibt sich, wenn man die Korrelationsmatrix zur Berechnung der Hauptkomponenten verwendet. In dem Fall ist $P = CC^t$. Dann gilt mit der inversen Transformation $\boldsymbol{Y} = A\boldsymbol{Z}$:

$$\text{cov}(Z_j, Y_i) = \text{cov}\left(Z_j, \sum_{k=1}^{m} a_{ik} Z_k\right) = a_{ij} \text{Var}(Z_j) = a_{ij} \lambda_j$$

Dabei wurde die Unkorreliertheit der Hauptkomponenten benutzt und ferner, dass $\text{Var}(Z_j) = \lambda_j$ gilt. Bei Verwendung der Korrelationsmatrix geht man von standardisierten Variablen Y_i aus, d.h. $\text{Var}(Y_i) = 1$. Damit ist dann

$$\text{Korr}(Z_j, Y_i) = \lambda_j a_{ij} / \lambda_j^{1/2} = a_{ij} \lambda_j^{1/2} \tag{4.18}$$

Damit ist die Matrix der Korrelationskoeffizienten zwischen $\boldsymbol{Z}$ und $\boldsymbol{Y}$ gegeben durch:

$$\text{Korr}(\boldsymbol{Z}, \boldsymbol{Y}) = A\Lambda^{1/2} = C \tag{4.19}$$

Wenn also C aus der Korrelationsmatrix P berechnet wurde, messen die Elemente von C die Korrelationen zwischen den Hauptkomponenten und den (standardisierten) Originalvariablen.

Ein weiterer interessanter Zusammenhang ergibt sich, wenn man in Gleichung 4.18 die Summe über i der quadrierten Korrelationskoeffizienten bildet und dabei beachtet, dass die Eigenvektoren normiert sind:

$$\sum_{i=1}^{m} (\text{Korr}(Z_j, Y_i))^2 = \lambda_j \sum_{i=1}^{m} a_{ij}^2 = \lambda_j \tag{4.20}$$

Das bedeutet: Die Varianz der j-ten Hauptkomponente ist gleich der Summe der quadrierten Korrelationskoeffizienten zwischen dieser Hauptkomponente und allen m Originalvariablen.

Als Beispiel verwenden wir wieder unseren bekannten Datensatz `teil01.frame`. Wir standardisieren unsere Daten so, dass sie Mittelwert Null und Varianz 1 haben.

```
Xstand<-(as.matrix(teil01.frame)-eins%*%t(mitte))%*%diag(1/stand)
```

Wir berechnen die Korrelationsmatrix. Vergleichen Sie dieses Ergebnis mit dem früheren (S. 40).

```
round(cor(Xstand),digits=4)
1.0000 0.8496 0.7739
0.8496 1.0000 0.7901
0.7739 0.7901 1.0000
```

Wir schreiben die Eigenwerte in eine Diagonalmatrix Lambda und die Eigenvektoren in eine Matrix A (siehe wieder S. 40).

```
Lambda<-diag(eigen(cor(Xstand))$values)
round(Lambda,digits=4)
2.6095 0.0000 0.0000
0.0000 0.2411 0.0000
0.0000 0.0000 0.1494
```

```
A<-eigen(cor(Xstand))$vectors
round(A,digits=4)
0.5809 -0.4715  0.6635
0.5846 -0.3254 -0.7432
0.5663  0.8196  0.0866
```

Wir berechnen die Komponentenladungen, bezeichnen Sie aber mit CL statt C, da es in **R** eine Funktion mit dem Namen C gibt.

```
CL<-A%*%Lambda^(1/2)
round(CL,digits=4)
0.9384 -0.2315  0.2564
0.9444 -0.1598 -0.2872
0.9148  0.4024  0.0335
```

Wir berechnen die Hauptkomponenten und bezeichnen die Matrix mit Z.

```
Z<-Xstand%*%A
round(Z[1:5,],digits=4)

1 -0.0615  0.6900 -0.4322
2  0.6253 -0.1754 -0.8644
3 -2.5835 -0.0256 -0.3011
```

```
4 -2.8796  0.3478 -0.2470
5 -1.2758  0.5189  0.0388
```

Wir berechnen die Korrelationen zwischen den standardisierten Variablen und den Hauptkomponenten:

```
round(cor(Xstand,Z),digits=4)
       Z1      Z2      Z3
Y1 0.9384 -0.2315  0.2564
Y2 0.9444 -0.1598 -0.2872
Y3 0.9148  0.4024  0.0335
```

Diese Korrelationen stimmen also mit den Komponentenladungen (gegeben in der Matrix CL) überein. Alle Variablen korrelieren also hoch mit der ersten Hauptkomponente.

Wir berechnen jetzt für jede Spalte der Matrix CL die Summe der Quadrate, also die Summe der quadrierten Korrelationskoeffizienten der Hauptkomponenten mit allen Originalvariablen. Dann müssten sich nach Gleichung 4.20 die Eigenwerte, d.h. die Varianzen der Hauptkomponenten ergeben.

```
round(sum(CL[,1]^2),digits=4)
2.6095
round(sum(CL[,2]^2),digits=4)
0.2411
round(sum(CL[,3]^2),digits=4)
0.1494
```

Struktur außerhalb der Diagonalen: Bei Chatfield und Collins (1991, S. 62) wird gezeigt, dass die Eigenvektoren und damit die Hauptkomponenten für $m = 2$ bei Verwendung von standardisierten Variablen nicht vom Korrelationskoeffizienten abhängen. Später zeigen sie (S. 67), dass sich die Eigenvektoren nicht ändern, wenn man alle Elemente außerhalb der Diagonalen mit einem konstanten Faktor multipliziert. Dabei ändern sich nur die Eigenwerte und damit der Anteil der Varianz, der durch die Hauptkomponenten erklärt wird.

Unkorrelierte Variablen: Wir beginnen mit einem Beispiel. Zu unseren drei Variablen `Groesse, Schuh, Gewicht` fügen wir als vierte Variable `Woerter` hinzu und berechnen die Korrelationsmatrix.

```
round(cor(fragmet.frame[,c(2,3,4,7)],use="c"),digits=4)
          Groesse   Schuh  Gewicht  Woerter
 Groesse   1.0000  0.8525   0.7897   0.0248
 Schuh     0.8525  1.0000   0.8041   0.0154
 Gewicht   0.7897  0.8041   1.0000  -0.0166
 Woerter   0.0248  0.0154  -0.0166   1.0000
```

Diese Variable ist nahezu unkorreliert mit allen anderen. Wir berechnen noch die Kovarianzmatrix:

```
round(cov(fragmet.frame[,c(2,3,4,7)],use="c"),digits=2)
          Groesse  Schuh  Gewicht  Woerter
 Groesse    88.35  23.87   103.14    18.36
 Schuh      23.87   8.87    33.28     3.61
 Gewicht   103.14  33.28   193.07   -18.19
 Woerter    18.36   3.61   -18.19  6198.23
```

Jetzt bestimmen wir die Eigenwerte und Eigenvektoren der Kovarianzmatrix.

```
round(eigen(cov(fragmet.frame[,c(2,3,4,7)],use="c"))$values,digits=2)
6198.34 262.94 25.29 1.94
round(eigen(cov(fragmet.frame[,c(2,3,4,7)],use="c"))$vectors,digits=2)
0.00 -0.52  0.83  0.19
0.00 -0.16  0.12 -0.98
0.00 -0.84 -0.54  0.07
1.00  0.00  0.00  0.00
```

Der erste Eigenwert stimmt in diesem Fall nahezu mit der Varianz der Variablen `Woerter` überein. Der entprechende Eigenvektor, d.h. die Koeffizienten der ersten Hauptkomponente, enthält an der vierten Stelle eine 1, die übrigen Zahlen sind nahezu Null, d.h. die erste Hauptkomponente stimmt nahezu mit der Variablen `Woerter` überein.

Es gilt das allgemeine Resultat: wenn eine Variable (z.B. Y_i) unkorreliert ist mit allen anderen Variablen und die Varianz λ_i besitzt, so ist λ_i ein Eigenwert der Kovarianzmatrix und der entsprechende Eigenvektor hat an der i-ten Stelle eine 1 und sonst Nullen. Damit ist dann Y_i eine der Hauptkomponenten. Wenn alle Variablen untereinander unkorreliert sind, dann stimmen die Hauptkomponenten mit den Originalvariablen überein. Sie werden nur nach absteigender Varianz geordnet. In diesem Fall ist eine Hauptkomponentenanalyse wenig sinnvoll.

Als Beispiel dazu schauen wir uns die Korrelationsmatrix für die letzten vier Variablen aus der Datei `fragmet.frame` an:

```
round(cor(fragmet.frame[,5:8],use="c"),digits=4)
            UeGewicht  GroeBoe  Woerter   ZuZahl
 UeGewicht     1.0000  -0.0801   0.0456   0.0561
 GroeBoe      -0.0801   1.0000   0.0900  -0.0503
 Woerter       0.0456   0.0900   1.0000  -0.1126
 ZuZahl        0.0561  -0.0503  -0.1126   1.0000
```

Die Variablen sind nahezu unkorreliert. Wir berechnen die Kovarianzmatrix:

```
round(cov(fragmet.frame[,5:8],use="c"),digits=0)
            UeGewicht  GroeBoe  Woerter  ZuZahl
 UeGewicht      85339     -190     1054     456
 GroeBoe         -190       66       58     -11
 Woerter         1054       58     6266    -248
 ZuZahl           456      -11     -248     772
```

Die Eigenwerte und Eigenvektoren der Kovarianzmatrix sind:

```
round(eigen(cov(fragmet.frame[,5:8],use="c"))$values,digits=0)
85356 6264 758 65
round(eigen(cov(fragmet.frame[,5:8],use="c"))$vectors,digits=2)
             HK1    HK2    HK3    HK4
 UeGewicht  1.00  -0.01   0.01   0.00
 GroeBoe    0.00   0.01   0.01  -1.00
 Woerter    0.01   1.00  -0.05   0.01
 ZuZahl     0.01  -0.05  -1.00  -0.01
```

Die Eigenwerte entsprechen in etwa den Varianzen der Variablen in der Reihenfolge `UeGewicht, Woerter, ZuZahl, GroeBoe`. Die Eigenvektoren haben an der entsprechenden Stelle

eine 1 oder -1, die übrigen Werte sind nahezu Null. Die Hauptkomponenten stimmen also mit den ursprünglichen Variablen oder dem Negativen überein. Die Hauptkomponentenanalyse ist also nicht sinnvoll.

Das Problem der Skalierung: Wir haben bereits gesehen, dass wir andere Hauptkomponenten erhalten, wenn wir die Korrelationsmatrix anstelle der Kovarianzmatrix verwenden. Wieder andere Ergebnisse würden wir bei der Analyse der Kovarianzmatrix erhalten, wenn wir die Körpergröße in Meter und das Gewicht in Pfund messen würden. Die Hauptkomponenten hängen von der Skalierung ab. Eine Variable mit großer Varianz wird die erste Hauptkomponente der Kovarianzmatrix dominieren, egal wie die Korrelationsstruktur aussieht. Wenn alle Variablen so skaliert werden, dass sie Varianz 1 haben, gehen alle Variablen mit gleichem Gewicht in die Analyse ein und die Ergebnisse werden ganz anders sein. Man umgeht das Skalierungsproblem, indem man die Korrelationsmatrix untersucht. Sonst ist sie nur sinnvoll, wenn alle Variablen in etwa gleiche Varianz haben.

4.4 Auswahl der Hauptkomponenten

Wir setzen voraus, dass wir die Korrelationsmatrix P für die Hauptkomponentenanalyse verwenden. Mit A hatten wir die Matrix bezeichnet, in deren Spalten die Eigenvektoren stehen. Mit Λ hatten wir die Diagonalmatrix der Eigenwerte bezeichnet. Dann ist Λ die Kovarianzmatrix der Hauptkomponenten. Zwischen P und Λ bestehen dann die folgenden Zusammenhänge (siehe Gleichungen 4.7 und 4.13)

$$\Lambda = A^t P A \qquad \text{und} \qquad P = A \Lambda A^t \tag{4.21}$$

Die zweite Gleichung zeigt, dass durch die Hauptkomponenten nicht nur die Varianz der Originalvariablen, sondern auch die Korrelationsstruktur reproduziert wird.

Reduziert man die Anzahl der Hauptkomponenten von m auf m_1, indem man diejenigen weglässt, die zu den kleinsten Eigenwerten gehören, so wird die Gesamtvarianz nicht mehr vollständig wiedergegeben, da

$$m = \sum_{j=1}^{m} \text{Var}(Y_j) > \sum_{j=1}^{m_1} \text{Var}(Z_j)$$

Ähnlich lässt sich die Korrelationsmatrix zerlegen. Nach Gleichung 4.8 gilt

$$P = A \Lambda A^t = \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1^t + \ldots + \lambda_m \boldsymbol{a}_m \boldsymbol{a}_m^t \tag{4.22}$$

Man bezeichnet dies als Spektralzerlegung der Korrelationsmatrix P. Jeder Summand $P_j := \lambda_j \boldsymbol{a}_j \boldsymbol{a}_j^t$ stellt eine $m \times m$-Matrix dar und es gilt:

$$P = \sum_{j=1}^{m} P_j \tag{4.23}$$

Die j-te Hauptkomponente liefert den Beitrag P_j zur Korrelationsmatrix. Man kann sich also die Korrelationsmatrix als Überlagerung von m Schichten darstellen, wobei die j-te Schicht

die Matrix P_j ist. Wählt man nur die ersten m_1 Hauptkomponenten aus, so erhält man als Korrelationsmatrix des reduzierten Modells:

$$\tilde{P} := \sum_{j=1}^{m_1} P_j \tag{4.24}$$

$\tilde{P}$ gibt die Korrelationen zwischen den m Originalvariablen an, wenn diese nur durch m_1 Hauptkomponenten erklärt würden. Man bezeichnet dann

$$P^* := P - \tilde{P} \tag{4.25}$$

als Rest- oder Fehlermatrix. Im Zusamenhang mit den Komponentenladungen hatten wir anstelle der auf eins normierten Eigenvektoren die auf λ_j normierten Vektoren $\boldsymbol{a}_j^* = \lambda_j^{1/2}\boldsymbol{a}_j$ betrachtet. Dann gilt mit der Matrix $C = [\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \ldots, \boldsymbol{a}_m^*]$

$$C = A\Lambda^{1/2}$$

und

$$P = CC^t = \boldsymbol{a}_1^*\boldsymbol{a}_1^{*t} + \ldots + \boldsymbol{a}_m^*\boldsymbol{a}_m^{*t}$$

In der Diagonalen der Korrelationsmatrix P stehen die Varianzen der standardisierten Originalvariablen. Es gilt somit

$$1 = \sum_{j=1}^{m} (a_{ij}^*)^2 \tag{4.26}$$

Das bedeutet: Die Varianz der i-ten standardisierten Variablen ist die Summe der quadrierten Korrelationskoeffizienten dieser i-ten Variablen mit allen Hauptkomponenten (vergleiche mit Gleichung 4.20. Dort war die Varianz der j-ten Hauptkomponente gleich der Summe der quadrierten Korrelationskoeffizienten zwischen dieser Hauptkomponente und allen m Originalvariablen.)

Lässt man nun die letzten Hauptkomponenten weg, so wird i.a. $\sum_{j=1}^{m_1} (a_{ij}^*)^2 < 1$ sein, d.h. die Matrix $\tilde{P}$ enthält in der Diagonalen keine Einsen. Man nennt die Elemente $\tilde{\rho}_{ii}$ in der Diagonalen von $\tilde{P}$ die Kommunalität der i-ten standardisierten Originalvariablen. Die Kommunalität ist jener Teil der Varianz der standardisierten Originalvariablen, der durch die m_1 wichtigsten Hauptkomponenten reproduziert wird.

Jetzt stellt sich die Frage, wie soll m_1 gewählt werden. Dazu gibt es verschiedene Vorschläge:

a) Man stellt die Eigenwerte λ_j auf der Ordinate und den Index j auf der Abszisse dar, wie in Abbildung 4.1 zu sehen ist, die mit den folgenden **R**-Befehlen erzeugt wurde:

```
EWUcor.eigen<-eigen(cor(EWU.frame[,2:5]))$values
plot(EWUcor.eigen,type="o",ylab="Eigenwerte")
title(main="Graph zur Durchführung des Scree-Tests")
```

Weist der Graph einen deutlichen Knick auf, so kann man alle Eigenwerte rechts von diesem Knick als nur zufällig von Null verschieden auffassen. Diese werden dann nicht verwendet. Dieses Verfahren heißt **Scree-Test**. In **R** gibt es dazu eine Funktion, die in folgender Weise benutzt werden kann.

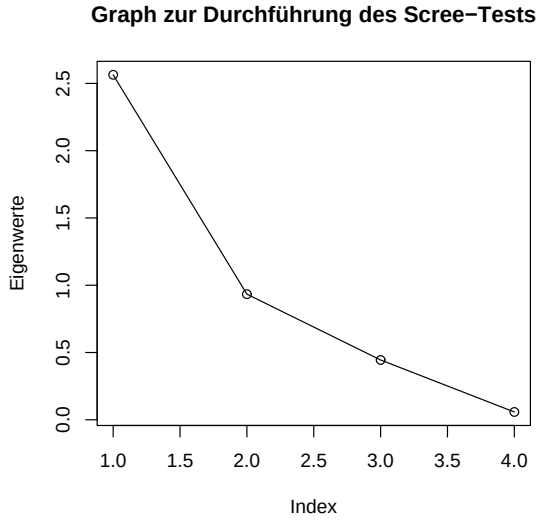

Abbildung 4.1: Grafik zum Scree-Test

```
pr.aus<-princomp(EWU.frame[,2:5],cor=T)
screeplot(pr.aus,type="l ")
```

Es geht auch mit `plot` statt `screeplot`. Mit dem Argument `type` können Sie zwischen `barplot` und `lines` wählen.

b) Man berechnet nur so viele Hauptkomponenten wie Eigenwerte $\lambda_j > 1$ sind, d.h. man nimmt nur diejenigen Hauptkomponenten, deren Varianz größer als die einer standardisierten Originalvariablen ist.

c) Man beginnt mit der ersten Hauptkomponente zum Eigenwert λ_1 und nimmt dann so lange die folgende Hauptkomponente dazu, bis die kumulierte Varianz der Hauptkomponenten einen vorgegebenen hohen Anteil der Gesamtvariation der standardisierten Originalvariablen überschritten hat, d.h. bei einem vorgegebenen Anteil von z.B. 0.90 sucht man das kleinste m_1, für das $\sum_{j=1}^{m_1} \lambda_j/m > 0.90$ ist.

d) Unter der Voraussetzung, dass die Daten eine multivariate Normalverteilung besitzen, kann ein Test von Bartlett durchgeführt werden. Es wird geprüft, ob sich die $m - m_1$ kleinsten Eigenwerte noch siginifikant unterscheiden. Kann die Nullhypothese

$$H_0 : \ \lambda_{m_1+1} = \ldots = \lambda_m$$

erstmals nicht mehr verworfen werden, so verwendet man nur die ersten m_1 Hauptkomponenten. Die Prüfgröße ist:

$$B := (n-1)\left[-\ln\left(\prod_{j=1}^{m}\lambda_j\right) + \ln\left(\prod_{j=1}^{m_1}\lambda_j\right) + (m-m_1)\ln\left(\frac{m-\sum_{j=1}^{m_1}\lambda_j}{m-m_1}\right)\right] \quad (4.27)$$

Die Prüfgröße ist mit kritischen Werten der χ^2-Verteilung mit $(m - m_1 + 2)(m - m_1 - 1)/2$ Freiheitsgraden zu vergleichen. Die Hypothese wird für große Werte von B abgelehnt.

Beispiel: Wir verwenden den Datensatz `EWU.frame` (siehe auch Rinne, 2000, S. 107). Nach dem Scree-Test (siehe Abbildung 4.1 würde man $m_1 = 2$ wählen, da der Graph bei j=2 einen deutlichen Knick zeigt. Die Eigenwerte waren:

```
round(EWUcor.eigen,digits=2)
2.56 0.93 0.44 0.06
```

Würde man also nach der Regel aus b) nur so viele Hauptkomponenten verwenden, wie es Eigenwerte größer als 1 gibt, hätte man $m_1 = 1$ zu wählen.

Würde man nach der Regel in c) so viele Hauptkomponenten verwenden wollen, bis sie einen Anteil von 90% der Gesamtvariation wiederspiegeln, müsste man $m_1 = 3$ wählen, da die kumulierten Anteile gegeben sind durch:

```
cumsum(round(EWUcor.eigen,digits=2))/4
0.6400 0.8725 0.9825 0.9975
```

Für Bartletts Test habe ich eine **R**-Funktion `Bartlett.fun` geschrieben, die von $m_1 = 1$ bis $m_1 = m - 2$ alle Prüfgrößen und die zugehörigen P-Werte berechnet. Außerdem werden die Freiheitsgrade ausgegeben. Als Argument ist die Datenmatrix einzugeben.

```
Bartlett.fun(EWU.frame[,2:5])
$m1
1
$B
21.2401
$PWert
7e-04
$FG
5

$m1
2
$B
12.5372
$PWert
0.0019
$FG
2
```

In diesem Fall wird der Test für $m_1 = 2$ noch verworfen, d.h. man würde $m_1 = 3$ verwenden.

Wir berechnen für dieses Beispiel die Matrix $C = A\Lambda^{1/2}$ der Komponentenladungen (siehe Gleichung 4.19), d.h. die Korrelationen der Ursprungsvariablen mit den Hauptkomponenten.

```
round(EWUcor.Vektor%*%diag(EWUcor.eigen)^(1/2),digits=4)

     HK1     HK2     HK3     HK4
X1 0.8904 -0.4209  0.0335  0.1700
X2 0.8896 -0.4235 -0.0183 -0.1700
X3 0.6936  0.5500  0.4651 -0.0096
X4 0.7062  0.5241 -0.4760  0.0092
```

Wir wollen für dieses Beispiel noch die Spektralzerlegung der Korrelationsmatrix berechnen.

Wir hatten nach Gleichung 4.22

$$R = A\Lambda A^t = \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1^t + \ldots + \lambda_m \boldsymbol{a}_m \boldsymbol{a}_m^t$$

Nach Gleichung 4.23 ist

$$R = \sum_{j=1}^{m} R_j$$

mit $R_j := \lambda_j \boldsymbol{a}_j \boldsymbol{a}_j^t$. Wir schreiben R statt P, da wir eine geschätzte Korrelationsmatrix haben.

Hier ist R gegeben durch:

```
round(cor(EWU.frame[2:5]),digits=4)
        X1      X2      X3      X4
 X1  1.0000  0.9408  0.4000  0.3938
 X2  0.9408  1.0000  0.3772  0.4134
 X3  0.4000  0.3772  1.0000  0.5565
 X4  0.3938  0.4134  0.5565  1.0000
```

Wir definieren in **R** eigene Symbole für die Eigenwerte und Eigenvektoren, um dann damit die Spektralzerlegung auszurechnen:

```
lambda1<-EWUcor.eigen[1]
lambda2<-EWUcor.eigen[2]
lambda3<-EWUcor.eigen[3]
lambda4<-EWUcor.eigen[4]
a1<-EWUcor.Vektor[,1]
a2<-EWUcor.Vektor[,2]
a3<-EWUcor.Vektor[,3]
a4<-EWUcor.Vektor[,4]
```

Nun erhalten wir R_1:

```
R1<-round(lambda1*a1%*%t(a1),digits=4)
R1
0.7928 0.7921 0.6176 0.6288
0.7921 0.7914 0.6170 0.6282
0.6176 0.6170 0.4811 0.4898
0.6288 0.6282 0.4898 0.4987
```

R_2:

```
R2<- round(lambda2*a2%*%t(a2),digits=4)
R2
 0.1772  0.1783 -0.2315 -0.2206
 0.1783  0.1794 -0.2329 -0.2219
-0.2315 -0.2329  0.3025  0.2882
-0.2206 -0.2219  0.2882  0.2746
```

R_3:

```
R3<- round(lambda3*a3%*%t(a3),digits=4)
R3
 0.0011 -0.0006  0.0156 -0.0160
-0.0006  0.0003 -0.0085  0.0087
```

```
 0.0156 -0.0085  0.2163 -0.2214
-0.0160  0.0087 -0.2214  0.2266
```

R_4:

```
R4<-round(lambda4*a4%*%t(a4),digits=4)
R4
 0.0289 -0.0289 -0.0016  0.0016
-0.0289  0.0289  0.0016 -0.0016
-0.0016  0.0016  0.0001 -0.0001
 0.0016 -0.0016 -0.0001  0.0001
```

Man beachte wie die Beiträge R_j zur Korrelationsmatrix mit wachsenden j abnehmen. Würde man $m_1 = 1$ wählen, hätte man als Korrelationsmatrix

$\tilde{R} = R_1$

```
R1
```
0.7928 0.7921 0.6176 0.6288
0.7921 **0.7914** 0.6170 0.6282
0.6176 0.6170 **0.4811** 0.4898
0.6288 0.6282 0.4898 **0.4987**

Die Kommunalitäten haben wir fett gedruckt. Die Fehlermatrix wäre dann:

$R^* = R - R_1 = R_2 + R_3 + R_4$

```
R2+R3+R4
 0.2072  0.1488 -0.2175 -0.2350
 0.1488  0.2086 -0.2398 -0.2148
-0.2175 -0.2398  0.5189  0.0667
-0.2350 -0.2148  0.0667  0.5013
```

Für $m_1 = 2$ hätte man als Korrelationsmatrix $\tilde{R} = R_1 + R_2$

```
R1+R2
```
0.9700 0.9704 0.3861 0.4082
0.9704 **0.9708** 0.3841 0.4063
0.3861 0.3841 **0.7836** 0.7780
0.4082 0.4063 0.7780 **0.7733**

Die Fehlermatrix wäre dann:

$R^* = R - R_1 - R_2 = R_3 + R_4$

```
R3+R4
 0.0300 -0.0295  0.0140 -0.0144
-0.0295  0.0292 -0.0069  0.0071
 0.0140 -0.0069  0.2164 -0.2215
-0.0144  0.0071 -0.2215  0.2267
```

Für $m_1 = 3$ wäre $\tilde{R} = R_1 + R_2 + R_3$

```
R1+R2+R3
```
0.9711 0.9698 0.4017 0.3922
0.9698 **0.9711** 0.3756 0.4150
0.4017 0.3756 **0.9999** 0.5566

0.3922 0.4150 0.5566 **0.9999**

Die Fehlermatrix wäre dann:

$R^* = R - R_1 - R_2 - R_3 = R_4$

```
R4
 0.0289 -0.0289 -0.0016  0.0016
-0.0289  0.0289  0.0016 -0.0016
-0.0016  0.0016  0.0001 -0.0001
 0.0016 -0.0016 -0.0001  0.0001
```

Man beachte, wie sich die Kommunalitäten mit wachsendem m_1 dem Wert Eins nähern.

Wir geben für dieses Beispiel noch eine grafische Darstellung der Ladungen der beiden ersten Hauptkomponenten. Auf der Abszisse werden die Komponentenladungen der ersten Hauptkomponente, auf der Ordinate die der zweiten Komponente abgetragen (siehe Abbildung 4.2).

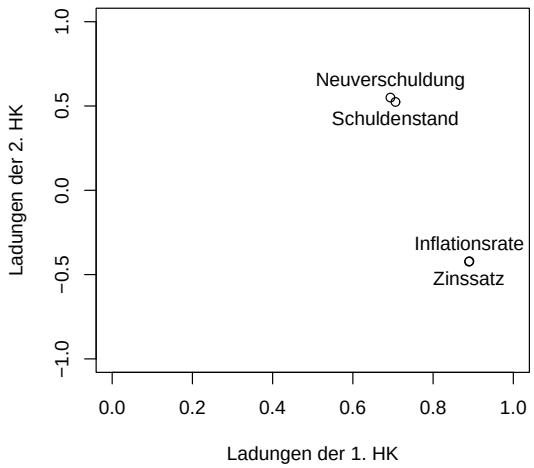

Abbildung 4.2: Grafische Darstellungen der Komponentenladungen für das Beispiel der EWU-Konvergenzkriterien

Wir erreichen diese Abbildung mit folgenden Befehlen:

```
EWUCL<-EWUcor.Vektor%*%diag(EWUcor.eigen)^(1/2)
EWUnames<-c("Inflationsrate","Zinssatz","Neuverschuldung",
"Schuldenstand")
LadeHK1<-EWUCL[,1]
LadeHK2<-EWUCL[,2]
plot(LadeHK1,LadeHK2,xlim=c(0,1),ylim=c(-1,1),xlab="Ladungen der
1. HK", ylab="Ladungen der 2. HK")
identify(LadeHK1,LadeHK2,labels=EWUnames)
```

Abbildung 4.2 zeigt eine Zweiteilung der vier Variablen. Die eine Gruppe wird von den beiden den Staatshaushalt charakterisierenden Variablen *öffentliche Neuverschuldung* und *öffentlicher Schuldenstand* gebildet. Die anderen beiden rein ökonomischen Variablen *Inflationsrate* und *langfristiger Zinssatz* fallen praktisch in einem Punkt zusammen.

In Abbildung 4.2 haben wir die Variablen dargestellt und können Ähnlichkeiten zwischen den Variablen entdecken, d.h. wir haben eine spaltenorientierte Methode.

In Abbildung 4.3 können wir Ähnlichkeiten zwischen den Zeilen, hier den Staaten, beobachten. Dort sind die ersten beiden Hauptkomponenten gegeneinander dargestellt.

```
EWUHK<-princomp(EWU.frame[,2:5],cor=T)$scores
plot(EWUHK[,1], EWUHK[,2], type = "n", xlab = "1. Hauptkomponente",
ylab = "2. Hauptkomponente")
text(EWUHK[,1],EWUHK[,2], labels=as.vector(Staat), cex=2)
```

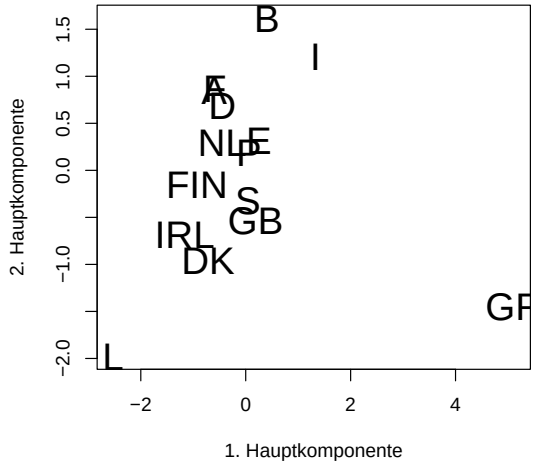

Abbildung 4.3: Grafische Darstellung der ersten beiden Hauptkomponenten für das Beispiel der EWU-Konvergenzkriterien

Die Hauptkomponentenanalyse ermöglicht es also, die Daten in zwei Dimensionen darzustellen. Diese Darstellung zeigt die zwei Ausreißer *L* und *GR*. Gelegentlich ist es auch möglich, Gruppen in den Daten zu erkennen. Abbildung 4.4 zeigt diese Darstellung für den Datensatz `teil01.frame`. Wir haben dort die Zahlen 0 (für Männer) und 1 (für Frauen) als Plotsymbole gewählt. Die Befehle für diese Grafik sind (Beachten Sie, dass die Hauptkomponenten für dieses Beispiel in der Matrix Z stehen.).

```
plot(Z[,1],Z[,2], type="n", xlab = "1. Hauptkomponente",ylab = "2.
Hauptkomponente")
text(Z[,1], Z[,2], labels=frag.frame[rownames(Xstand),2], cex=1.2)
```

Wir sehen die Nullen überwiegend in der rechten Hälfte der Grafik, die Einsen in der linken. Die erste Hauptkomponente separiert also sehr gut nach Geschlecht. Schauen wir uns noch einmal die Ladungen an.

```
CL<-A%*%Lambda^(1/2)
round(CL,digits=4)
0.9384 -0.2315  0.2564
0.9444 -0.1598 -0.2872
0.9148  0.4024  0.0335
```

Die erste Hauptkomponente korreliert mit allen drei Variablen `Groesse, Schuh, Gewicht` sehr stark . Man könnte sie als allgemeinen Index für Größe bezeichen. Diese Hauptkomponente ist groß, wenn alle drei Variablen auch groß sind. Die 2. Hauptkomponente korreliert negativ mit `Groesse` und `Schuh` und positiv mit `Gewicht`. Dementsprechend könnte man diese Komponente als einen Index für Übergewicht interpretieren. Das Gewicht ist im Verhältnis zur Körpergröße und Schuhgröße zu groß, wenn die 2. Hauptkomponente groß ist.

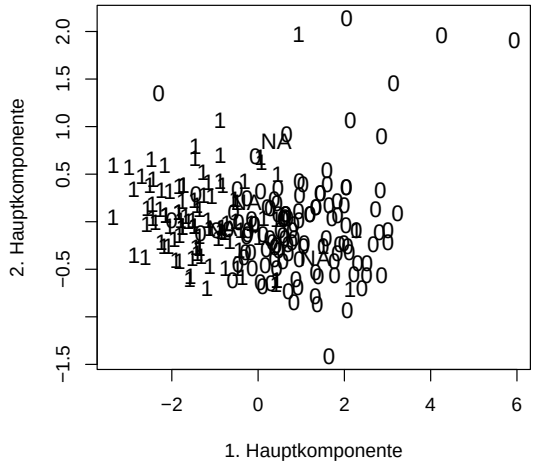

Abbildung 4.4: Grafische Darstellungen der ersten beiden Hauptkomponenten für den Datensatz `teil01.frame`

Die 2. Hauptkomponente ist klein, wenn das Gewicht im Verhältnis zu den beiden anderen Variablen klein ist. Diese Interpretation wird durch die folgenden Daten bestätigt.

```
round(cbind(sort(Z[,2]), Xstand[names(sort(Z[,2])),]),digits=4)
```

Z_2	Groesse	Schuh	Gewicht
-1.4117	1.6415	1.3998	-0.2223
-0.9330	1.7456	1.3998	0.4217
-0.8719	1.2253	1.0651	0.0640
-0.8503	0.9131	0.7303	-0.2223
-0.7846	0.8090	1.3998	0.0640
-0.7343	0.0806	1.3998	-0.2938
-0.7080	1.6415	1.3998	0.6364
-0.6996	1.6415	1.7346	0.7795
-0.6959	-0.2316	-0.6088	-1.2240
-0.6948	0.6009	1.0651	-0.0792
0.8958	1.2253	1.3998	2.3536
0.9234	-0.0235	0.0607	1.1372
1.0646	0.6009	1.0651	2.0674
1.0671	-0.6479	-1.2784	0.4217
1.3531	-2.7291	-0.9436	-0.2938
1.4523	0.8090	1.7346	2.9261
1.9022	2.0578	3.4085	4.8580
1.9598	1.9537	1.3998	4.0709
1.9719	0.0806	-0.6088	2.2105
2.1334	0.2887	0.3955	2.9261

Diese Tabellen enthalten in der ersten Spalte die der Größe nach geordneten Werte der 2.

Hauptkomponente, dann in den Spalten 2 - 4 die standardisierten Originalvariablen. Wir haben nur die ersten 10 und letzten 10 Werte ausgedruckt. Bei den ersten Werten, d.h. hier ist die 2. Hauptkomponente am kleinsten, ist die 3. standardisierte Variable deutlich kleiner als die beiden anderen (das sind die Untergewichtigen), bei den letzten 10 ist die dritte Variable deutlich größer als die beiden anderen (das sind die Übergewichtigen). Für dieses Beispiel geben wir jetzt noch die grafische Darstellung der Komponentenladungen, d.h. eine Darstellung der Variablen (siehe Abbildung 4.5). Es ist üblich, die Variablen durch Pfeile vom Ursprung aus darzustellen.

```
plot(CL[,1],CL[,2], type = "n", xlim=c(0,1), ylim=c(-1,1), xlab =
"Ladungen der 1. Hauptkomponente", ylab="Ladungen der 2. Hauptkomponente"
null<-rep(0,3)
arrows(null,null,CL[,1],CL[,2])
identify(CL[,1],CL[,2], labels=colnames(teil01.frame))
abline(h=0)
```

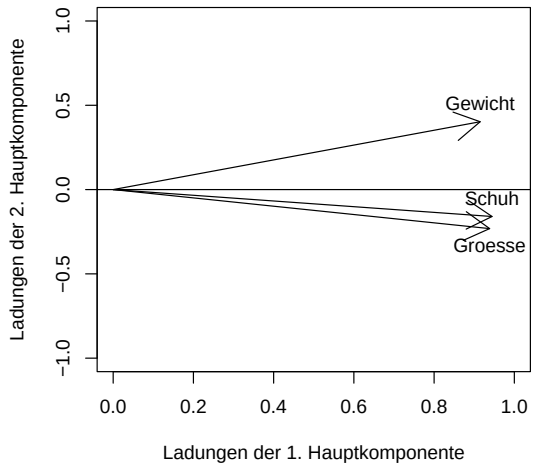

Abbildung 4.5: Grafische Darstellungen der Komponentenladungen für den Datensatz `teil01.frame`

Wir sehen an dieser Abbildung, dass die beiden Variablen `Groesse` und `Schuh` nahe zusammenliegen. Außerdem liegen sie nahe an der 1. Hauptkomponente. Dies zeigt die geringe Korrelation mit der zweiten Hauptkomponente. Aufgrund der negativen Korrelation mit der zweiten Hauptkomponente liegen sie unterhalb der Abszisse. Die Variable `Gewicht` zeigt in die positive Richtung der zweiten Hauptkomponente. Ist diese Variable groß, ist die zweite Hauptkomponente auch eher groß. Man beachte, dass alle Variablen etwa auf einem Kreis mit dem Radius 1 liegen, das liegt daran, dass die Summe (über die Spalten) der Quadrate der Komponentenladungen 1 ergibt. Gleichung 4.26 besagte, dass $1 = \sum_{j=1}^{m} (a_{ij}^*)^2$, d.h. die Varianz der standardisierten Originalvariablen ist gleich der Summe der Quadrate der Ladungen der Hauptkomponenten mit dieser Variablen. Da wir nur die ersten beiden Hauptkomponenten darstellen können, liegen die Variablen nicht genau auf einem Kreis. Die Summe der Quadrate der Ladungen der ersten beiden Hauptkomponenten mit dieser Variablen sind gerade die Kommunalitäten, die für dieses Beispiel so bestimmen können (die Ladungen standen in der Matrix `CL`):

```
sum(CL[1,1:2]^2)
```

```
0.934241
sum(CL[2,1:2]^2)
0.9174944
sum(CL[3,1:2]^2)
0.9988788
```

Diese Zahlen besagen, dass 93.42% der Varianz der Variablen `Groesse`, 91.75% der Varianz der Variablen `Schuh` und 99.89% der Varianz der Variablen `Gewicht` durch die beiden ersten Hauptkomponenten erklärt werden.

Wir werden jetzt eine Grafik kennenlernen, die die beiden bisherigen Grafiken - Darstellung der Merkmalsträger im Raum der beiden ersten Hauptkomponenten = Darstellung der Zeilen der Beobachtungsmatrix - und Darstellung der Variablen durch ihre Ladungen mit den beiden Hauptkomponenten = Darstellung der Spalten der Beobachtungsmatrix - miteinander in einer einzigen Grafik verbindet. Wir erzeugen die Grafik mit den folgenden Befehlen:

```
aus<-princomp(teil01.frame,cor=T)
biplot.princomp(aus, pc.biplot=T,
       xlab=frag.frame[rownames(teil01.frame),2])
```

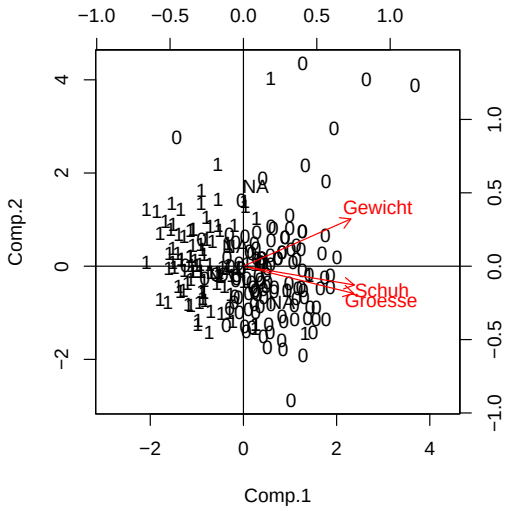

Abbildung 4.6: Biplot für den Datensatz `teil01.frame`

Standardmäßig, d.h. ohne das Argument `xlab = ...` werden die Zeilennamen der Datenmatrix als Plotsymbole für die Merkmalsträger (Zeilen) verwendet (siehe Abbildung 4.7).

Die Abbildungen 4.6 und 4.7 heißen *Biplot*. Ein Biplot ist eine grafische Darstellung einer $n \times m$-Matrix (siehe Gabriel (1971) und Gower und Hand (1996)). Jedoch ist es nicht ganz eindeutig, von welcher Matrix man ausgeht. Es gibt unterschiedliche Skalierungen. So könnte man z.B., wie es Rinne (2000, S. 36) macht, die mittelwertbereinigte Datenmatrix verwenden. Die Zeilen und Spalten dieser Matrix werden in einer gemeinsamen Grafik dargestellt. Jede Zeile dieser Matrix wird in der Grafik durch ein Symbol (z.B. die Zeilennummer) dargestellt. Die Variablen werden durch Pfeile dargestellt. Für die Darstellung der Zeilen und Spalten werden unterschiedliche Skalierungen verwendet. So gehört die linke und untere Skalierung in Abbildung 4.6 und 4.7 zu den Zeilen, d.h. Merkmalsträgern, während die rechte und obere Skalierung zu den Spalten, d.h. Variablen gehört. Im wesentlichen stellt man die beiden ersten Hauptkomponenten für die Zeilen und die Ladungen der beiden ersten Hauptkomponenten für die Spalten dar. Die Darstellung ist nur dann exakt, wenn die Matrix

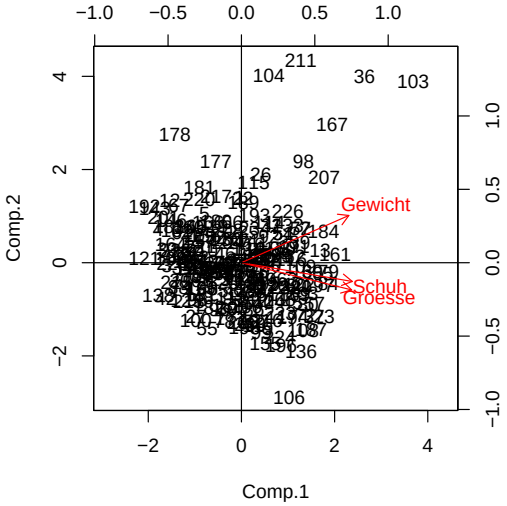

Abbildung 4.7: Biplot für den Datensatz `teil01.frame`

den Rang 2 hat. Folgendes ist an solchen Darstellungen zu beobachten:

a) Der euklidische Abstand zweier Punkte approximiert ein gewisses Abstandsmaß (die sogenannte Mahalanobis-Distanz) der Merkmalsträger. Je näher zwei Punkte im Biplot liegen, desto ähnlicher sind sich die Merkmalsträger.

b) Die Länge der Pfeile ist proportional zur Standardabweichung der dargestellten Variablen. Geht man von standardisierten Variablen aus, so ist die Standardabweichung natürlich 1.

c) Das Skalarprodukt zweier Vektoren (Pfeile) ist proportional zur Kovarianz der beiden Variablen.

d) Der Kosinus des Winkels zwischen zwei Variablen (Pfeilen) approximiert den Korrelationskoeffizienten dieser beiden Variablen. Ein Winkel von $90°$ oder $270°$ bedeutet Unkorreliertheit, einer von $0°$ oder $360°$ bedeutet perfekte positive Korrelation, einer von $180°$ perfekte negative Korrelation.

e) Betrachtet man die Lage des i-ten Punktes relativ zur Lage des j-ten Pfeils, so gilt folgendes:

- Liegt der i-te Punkt in derselben Richtung wie der j-te Pfeil, so ist der i-te Merkmalsträger bezüglich der j-ten Variablen überdurchschnittlich ausgeprägt.
- Liegt der i-te Punkt in entgegengesetzter Richtung wie der j-te Pfeil, so ist der i-te Merkmalsträger bezüglich der j-ten Variablen unterdurchschnittlich ausgeprägt.

Die Qualität eines Biplots hängt natürlich davon ab, welcher Anteil der Gesamtvariation durch die beiden ersten Hauptkomponenten wiedergegeben wird.

4.5 Hauptkomponentenanalyse für multivariat normalverteilte Daten

Bisher haben wir keine Annahmen über die Verteilung der Daten gemacht. Jetzt werden wir annehmen, dass die Beobachtungen einer multivariaten Normalverteilung entstammen:

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}; \Sigma)$$

An der Herleitung der Hauptkomponenten ändert sich nichts. Wenn wir die Matrix der Eigenvektoren mit A bezeichnen, so gilt für die Hauptkomponenten:

$$\boldsymbol{Z} = A^t(\boldsymbol{Y} - \boldsymbol{\mu}) \tag{4.28}$$

Wir haben angenommen, dass wir zunächst eine Korrektur am Erwartungswert vornehmen. Wir können jetzt, wenn wir eine Verteilung für $\boldsymbol{Y}$ voraussetzen, die Verteilung der Hauptkomponenten, also von $\boldsymbol{Z}$ bestimmen. Da jede Komponente von $\boldsymbol{Z}$ eine Linearkombination von $\boldsymbol{Y}$ ist, also eine Linearkombination von normalverteilten Zufallsvariablen, folgt, dass jede Komponente von $\boldsymbol{Z}$ normalverteilt ist. Der Erwartungswertvektor ist $\mathbf{0}$ und die Kovarianzmatrix ist Λ, wobei Λ wie schon früher eine Diagonalmatrix ist, deren i-tes Diagonalelement λ_i ist, d.h.

$$\boldsymbol{Z} \sim N(\mathbf{0}; \Lambda)$$

Die Hauptkomponenten haben im Falle einer gemeinsamen Normalverteilung eine anschauliche Interpretation. Sie stimmen mit den Hauptachsen eines Ellipsoids überein. Dieses Ellipsoid ist der geometrische Ort aller Vektoren mit gleicher Dichte. Im Falle einer bivariaten Normalverteilung (denken Sie an die Darstellung der gemeinsamen Dichtefunktion durch Höhenlinien wie in Abbildung 4.8 werden die Hauptachsen der Ellipse so gedreht, dass sie parallel zu den Koodinatenachsen verlaufen.

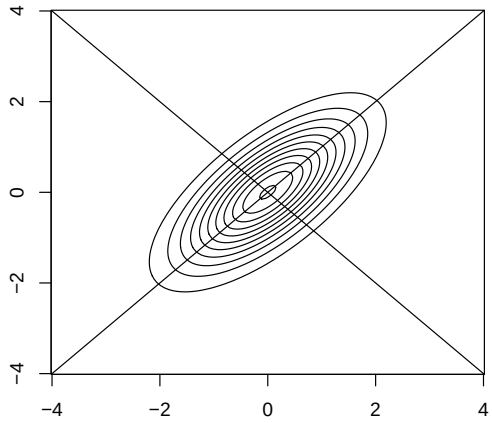

Abbildung 4.8: Höhenlininien für eine bivariate Normalverteilung ($\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0.7$*)*

In Abbildung 4.8 sind die Höhenlinien einer bivariaten Normalverteilung mit $\boldsymbol{\mu} = \mathbf{0}$ und $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$ dargestellt. Die Eigenwerte dieser Kovarianzmatrix sind $\lambda_1 = 1.7$ und $\lambda_2 = 0.3$. Demnach ist die Verteilung der Hauptkomponenten gegeben durch $N(\mathbf{0}, \Lambda)$ mit

$$\Lambda = \begin{pmatrix} 1.7 & 0 \\ 0 & 0.3 \end{pmatrix}$$

Die Höhenlinien dieser bivariaten Normalverteilung sind in Abbildung 4.9 dargestellt.

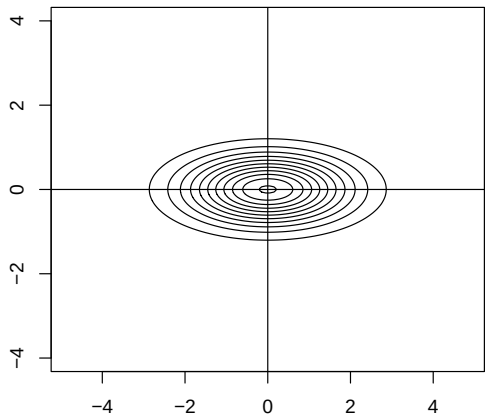

Abbildung 4.9: Höhenlininien für die Normalverteilung der Hauptkomponenten ($\mu_1 = \mu_2 = 0, \sigma_1^2 = 1.7, \sigma_2^2 = 0.3, \rho = 0$)

4.6 Zusammenfassung

Eine Hauptkomponentenanalyse ist eine orthogonale Transformation im m-dimensionalen Raum der Originalvariablen in eine neue Variablenmenge, die Hauptkomponenten genannt werden. Die Hauptkomponentenanalyse hat einige Nachteile. Die Ergebnisse sind abhängig von der Skalierung und daher nicht eindeutig. Die Hauptkomponenten sind schwierig zu interpretieren. Die Interpretation ist subjektiv. Auch die Anzahl der auszuwählenden bedeutenden Hauptkomponenten ist nicht eindeutig. Dennoch ist sie in der Lage, Einsichten in die Struktur der Daten zu bringen, insbesondere über die Korrelationsstruktur. Sie kann zu einer Reduktion der Anzahl der Variablen führen. Es ist oft sinnvoll, die wichtigsten Hauptkomponenten in weiterführenden Analysen zu benutzen. Die grafische Darstellung der Hauptkomponenten kann zu einem besseren Verständnis führen. Wir fassen die wesentlichen Schritte einer Hauptkomponentenanalyse zusammen (siehe Chatfield und Collins, 1991).

a) Entscheiden Sie, ob es lohnt alle Originalvariablen in die Analyse einzubeziehen und ob einige Variablen transformiert werden müssen.

b) Berechnen Sie die Kovarianz- oder Korrelationsmatrix und beachten Sie dabei, dass ein Korrelationskoeffizient nicht berechnet werden sollte, wenn die Beziehung zwischen zwei Variablen offensichtlich nichtlinear ist.

c) Betrachten Sie die Korrelationsmatrix und achten Sie darauf, ob es offensichtliche Gruppen in den Variablen mit hohen Korrelationen gibt. Wenn alle Korrelationen annähernd 0 sind, ist eine Hauptkomponentenanalyse nicht angebracht.

d) Berechnen Sie die Eigenwerte und Eigenvektoren der Korrelationsmatrix (oder Kovarianzmatrix).

e) Betrachten Sie die Eigenwerte und entscheiden Sie, wie viele der Eigenwerte wirklich „groß" sind. Diese Zahl gibt Ihnen die effektive Dimension der Daten an.

f) Schauen Sie, ob die Hauptkomponenten Ihnen Hinweise auf Gruppierungen der Variablen geben und versuchen Sie die Hauptkomponenten zu interpretieren.

g) Benutzen Sie die Hauptkomponenten für weitere Analysen, um damit die Dimension der Daten zu reduzieren.