9 Dimensionsreduktion

9.1 Mathematische Grundlagen

Gegeben seien eine *Datenmatrix* $X \in \mathbb{R}^{m \times n}$ und ihre *Stichprobenkovarianzmatrix* $C \in \mathbb{R}^{m \times m}$ (siehe Abschnitt 5.1 des Arbeitsblatts (5) *Multivariate Deskriptivstatistik*) mit

$$C := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^{\mathrm{T}} = \frac{1}{n-1} \left(X \left(I_n - \frac{1}{n} 1_{nn} \right) X^{\mathrm{T}} \right) , \tag{1}$$

wobei x_i die i-te Spalte von X und $\bar{x} \in \mathbb{R}^m$ das *multivariate Stichprobenmittel* bezeichne.

Die *Hauptkomponentenanalyse* eines Datensatzes ist definiert als die Orthonormalzerlegung seiner Stichprobenkovarianzmatrix

$$C = Q \Lambda Q^{\mathrm{T}} , \tag{2}$$

wobei Λ eine Diagonalmatrix ist, deren Diagonaleinträge die (absteigend nach Größe geordneten) Eigenwerte von C sind, und Q eine orthogonale Matrix ist, deren Spalten die (zugehörigen) Eigenvektoren von C sind.

Der *Hauptkomponentenanalyse-transformierte Datensatz* ist definiert als das Matrixprodukt

$$\tilde{X} = Q^{\mathrm{T}} X , \tag{3}$$

wobei die Spalten von Q als die *Hauptkomponenten* von C bezeichnet werden.

In der Vorlesung haben wir gesehen, dass für diese Hauptkomponentenanalyse folgendes gilt:

- Die Spalten von Q bilden eine Orthonormalbasis von $\mathbb{R}^m$, d.h. die Multiplikation mit Q^{T} transformiert die Datenmatrix-Einträge bezüglich der kanonischen Basis von $\mathbb{R}^m$ in Koordinaten bezüglich der Hauptkomponenten von C.
- Die Matrix Q spezifiziert die Hauptkomponenten und bildet originale Variablen auf transformierte Variablen ab, d.h. der Eintrag q_{ij} (i-te Zeile, j-Spalte) gibt an, wie stark die i-te originale Variable (X) in der j-ten transformierten Variable ($\tilde{X}$) repräsentiert ist.
- Die Stichprobenkovarianzmatrix des transformierten Datensatzes ist Λ:

$$\tilde{C} := \frac{1}{n-1} \sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\tilde{x}} \right) \left(\tilde{x}_i - \bar{\tilde{x}} \right)^{\mathrm{T}} = \Lambda . \tag{4}$$

- Die Stichprobenkorrelationsmatrix des transformierten Datensatzes ist I_m:

$$\tilde{R} := \left(\frac{(\tilde{C})_{ij}}{\sqrt{(\tilde{C})_{ii}}\sqrt{(\tilde{C})_{jj}}} \right)_{1 \leq i,j \leq m} = I_m \, . \tag{5}$$

  Dies folgt daraus, dass die Stichprobenkovarianzmatrix $\tilde{C}$ des transformierten Datensatzes $\tilde{X}$ eine Diagonalmatrix ist (Λ) und mithin ihre Nicht-Diagonaleinträge 0 sind. Damit sind alle paarweisen Kovarianzen $(\tilde{C})_{ij}$, $i \neq j$ und mithin auch alle paarweisen Korrelationen $(\tilde{R})_{ij}$, $i \neq j$ gleich 0.

In der vorliegenden Übung wollen wir diese mathematischen Eigenschaften der Hauptkomponentenanalyse anhand der Anwendung auf einen realen Datensatz nachvollziehen.

9.2 Analyse in R

Die Datei `FADE_SAME.csv` enthält den in der ersten Seminarsitzung vorgestellten Datensatz. Erklären Sie die Funktion des folgenden R-Codes:

```
# Daten einlesen
fname = 'FADE_SAME.csv'                                  # Dateiname
D     = read.csv(fname)                                  # Dataframe

# Datenmatrix extrahieren
rows  = startsWith(D$subject, 'subA')                    # Personen aus Studie A
cols  = c('novelty.FADE','novelty.SAME','memory.FADE','memory.SAME',
          'age','memory','V.HC.left','V.HC.right')       # Definition Variablen
n     = sum(rows)                                        # Anzahl Datenpunkte
m     = length(cols)                                     # Anzahl Variablen
X     = t(as.matrix(D[rows,cols]))                       # m x n Datenmatrix
print(dim(X))                                            # Überprüfung Datenmatrix

# Altersgruppeneffekte herausrechnen
x     = D$age[rows]                                      # gruppendefinierende Variable
X_x   = cbind(1*(x <= 35), 1*(x > 35))                   # n x 2 Designmatrix
R_x   = diag(n) - X_x %*% solve(t(X_x) %*% X_x) %*% t(X_x)  # n x n residuenbildende Matrix
E_x   = R_x %*% t(X)                                     # n x m Residuen
X     = t(E_x)                                           # korrigierte Datenmatrix
print(dim(X))                                            # Überprüfung Datenmatrix
```

9.3 Erste Programmieraufgabe

Führen Sie für die im vorherigen Abschnitt geladene und vorverarbeitete Datenmatrix $X \in \mathbb{R}^{m \times n}$ die Hauptkomponentenanalyse durch. Gehen Sie dazu wie folgt vor:

- Standardisieren Sie die Datenmatrix, d.h. ziehen Sie von jeder Spalte ihr Stichprobenmittel ab (`rowMeans(X)`) und dividieren Sie jede Spalte durch ihre Stichprobenstandardabweichung (`apply(X,1,sd)`). Diese Transformation sorgt dafür, dass in der resultierenden Datenmatrix jede Zeile das Stichprobenmittel 0 und die Stichprobenvarianz 1 hat.
- Berechnen Sie die Stichprobenkovarianzmatrix und die Stichprobenkorrelationsmatrix der Datenmatrix. Orientieren Sie sich hierbei an Abschnitt 5.3 des Arbeitsblatts (5) *Multivariate Deskriptivstatistik*. Speichern Sie die Ergebnisse als `C` und `R`.
- Führen Sie eine Eigenanalyse der Stichprobenkovarianzmatrix durch. Orientieren Sie sich hierbei an Abschnitt 3.2 des Arbeitsblatts (3) *Eigenanalyse*. Speichern Sie die Eigenwerte als Vektor `lambda` und die Eigenvektoren als Matrix `Q`.
- Berechnen Sie den Hauptkomponenten-transformierten Datensatz gemäß der oben angegebenen Formel. Speichern Sie die transformierte Datenmatrix als `X_tilde`.
- Berechnen Sie die Stichprobenkovarianzmatrix und die Stichprobenkorrelationsmatrix des transformierten Datensatzes. Orientieren Sie sich erneut an Abschnitt 5.3 des Arbeitsblatts (5) *Multivariate Deskriptivstatistik*. Speichern Sie die Ergebnisse als `C_tilde` und `R_tilde`.
- Geben Sie die ersten 5 Spalten der transformierten Datenmatrix aus. Sie sollten folgende Ergebnisse erhalten:

```
           [,1]        [,2]       [,3]         [,4]        [,5]
[1,] -0.5720948 -0.84877102 -3.0746290  0.298976337  1.04176184
[2,] -1.1069561 -0.42840640  0.6393084  0.311179977 -0.51561271
[3,] -0.5299393 -2.21840258 -0.1635837 -0.009469275  1.52128352
[4,]  0.9610312  0.88008201  1.0674911  0.044727674 -0.50397609
[5,]  1.2803564  0.16080050  2.5580011 -0.980029110 -0.07857034
[6,]  0.3002272  0.53581290 -1.8837037  0.689082544  0.48049802
[7,]  0.1607294  0.21986225 -0.3499547 -0.353602418 -0.04119800
[8,]  0.3040577  0.02811444  0.1027157 -0.287257955 -0.42884352
```

9.4 Abbildung in R

Die originale und die transformierte Datenmatrix sollen im Folgenden dargestellt werden. Erklären Sie dazu den folgenden R-Code und die Abbildung, die er erzeugt:

```
# Abbildungsparameter
library(latex2exp)
library(plot.matrix)
par(
    family      = "sans",
    mfcol       = c(2,1),
    pty         = "m",
    bty         = "l",
    lwd         = 1,
    las         = 1,
    mgp         = c(2,1,0),
    xaxs        = "i",
    yaxs        = "i",
    font.main   = 1,
    cex         = 1,
    cex.main    = 2)

# Visualisierung X
plot(X[,1:30],
    breaks      = c(-5,5),
    col         = topo.colors,
    fmt.key     = "%.0f",
    polygon.key = NULL,
    axis.key    = NULL,
    xlab        = "",
    ylab        = "",
    main        = TeX("$X$"))

# Visualisierung \tilde{X}
plot(X_tilde[,1:30],
    breaks      = c(-5,5),
    col         = topo.colors,
    fmt.key     = "%.0f",
    polygon.key = NULL,
    axis.key    = NULL,
    xlab        = "",
    ylab        = "",
    main        = TeX("$\\tilde{X}$"))

# Speichern
dev.copy2pdf(
    file        = "Abbildungen/Dimensionsreduktion_1.pdf",
    width       = 15,
    height      = 10)
```

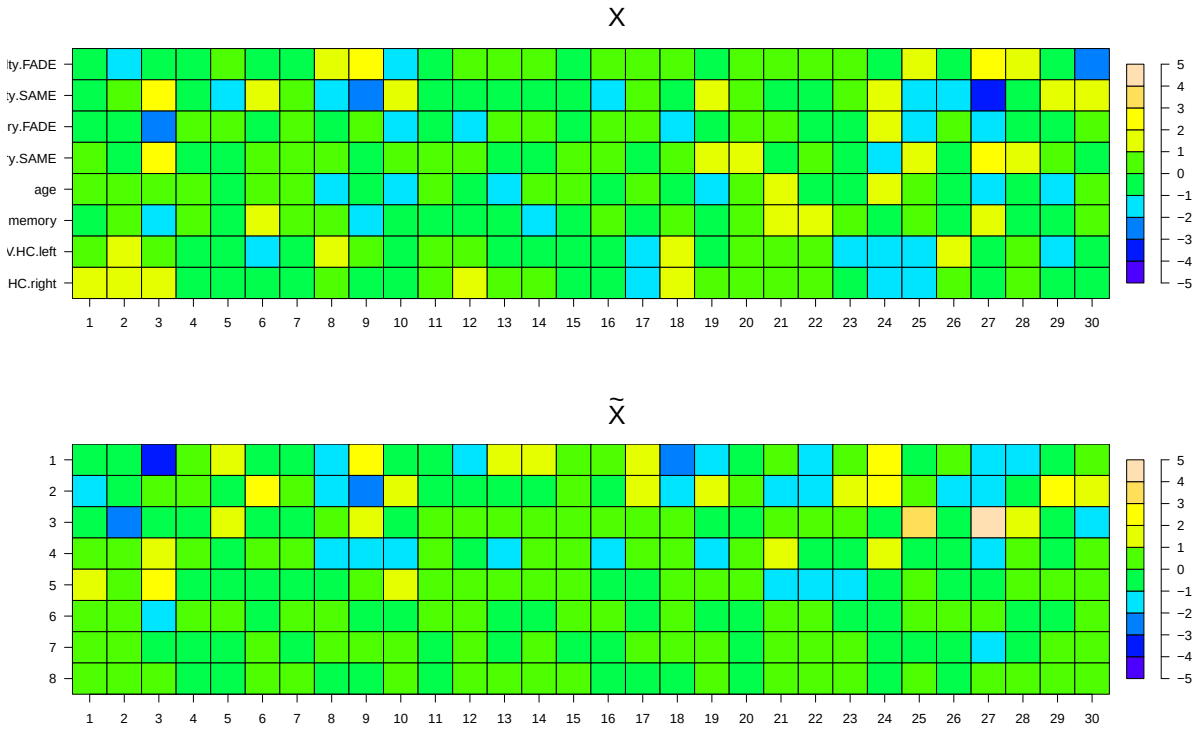

Abbildung 1. Datensatz X und Hauptkomponentenanalyse-transformierter Datensatz $\tilde{X}$.

9.5 Zweite Programmieraufgabe

Des Weiteren sollen nun nicht nur die Datenmatrizen, sondern auch die Transformationsmatrizen der Hauptkomponentenanalyse visualisiert werden. Gehen Sie dazu wie folgt vor:

- Setzen Sie mithilfe logischer Indizierung diejenigen Werte von `C_tilde` und `R_tilde`, die kleiner sind als 0.001, auf 0. Dies verhindert im Folgenden Artefakte in der Darstellung.
- Definieren Sie dann die folgenden Abbildungsparameter, die eine Figure mit 2×3 Panels vorbereitet:

```
# Abbildungsparameter
par(
    family      = "sans",
    mfcol       = c(2,3),
    pty         = "m",
    bty         = "l",
    lwd         = 1,
    las         = 1,
    mgp         = c(2,1,0),
    xaxs        = "i",
    yaxs        = "i",
    font.main   = 1,
    cex         = 1,
    cex.main    = 2)
```

- Der folgende Code visualisiert die Einträge einer quadratischen Matrix A auf einer diskreten Farbskala:

```
plot(A,
    col         = topo.colors,
    digits      = 2,
    key         = NULL,
    cex         = 0.8,
    polygon.key = NULL,
    axis.key    = NULL,
    xlab        = "",
    ylab        = "",
    main        = TeX("$A$"))
```

- Visualisieren Sie auf diese Weise nacheinander die Matrizen Q, Λ (`diag(lambda)`), C, $\tilde{C}$, R und $\tilde{R}$.
- Speichern Sie die resultierende Graphik wie im vorherigen Abschnitt. Sie sollten folgende Abbildung erhalten:

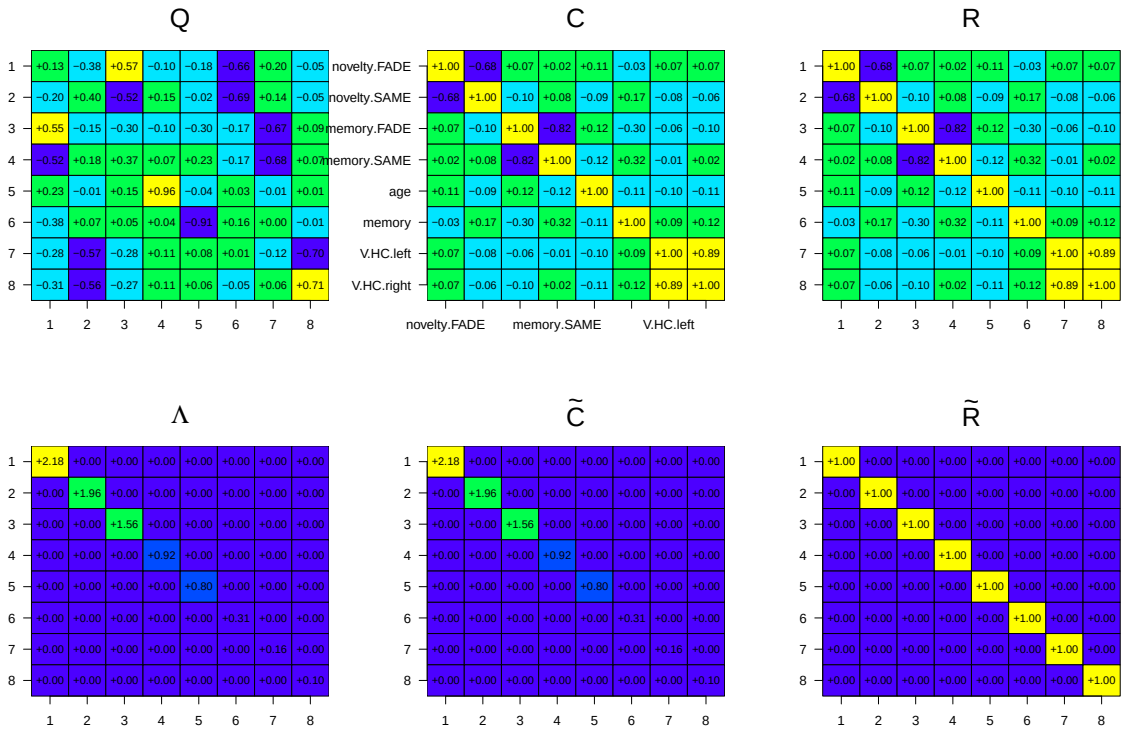

Abbildung 2. Mit der Hauptkomponentenanalyse des Datensatzes X assoziierte Matrizen.

9.6 Lückentext

Füllen Sie mit den in der Übung gewonnenen Erkenntnissen den folgenden Lückentext aus und präsentieren Sie die Ergebnisse im Seminar:

Lückentext: Die Datenmatrix besteht aus _____ Datenpunkten und _____ Variablen. Anhand der Stichprobenkorrelationsmatrix lässt sich sehen, dass die Variablenpaare _____ sowie _____ besonders stark negativ korrelieren ($r < -0.5$), während das Variablenpaar _____ besonders stark positiv korreliert ($r > +0.5$). Die ersten drei Hauptkomponenten werden dementsprechend maßgeblich durch diese drei Variablenpaare bestimmt (Spalten _____ bis _____ der Matrix Q). Die vierte Hauptkomponente besteht maßgeblich aus der Variable _____, während die fünfte Hauptkomponente maßgeblich auf der Variable _____ basiert. Die Summe der Stichprobenvarianzen (d.h. der Diagonaleinträge der Stichprobenkovarianzmatrix) beträgt sowohl für die originale Datenmatrix als auch die transformierte Datenmatrix _____. Die Varianzen der transformierten Variablen 6, 7 und 8 sind vernachlässigbar. Qualitativ gesprochen wird der ursprünglich aus _____ Variablen bestehende Datensatz also auf _____ Hauptkomponenten-transformierte Variablen dimensionsreduziert.

9.7 Mögliche Klausurfrage

Präsentieren Sie im Seminar folgende Klausurfrage und erklären Sie die richtige Antwort:

Frage: $\tilde{X} = Q^{\mathrm{T}}X$ sei der aus einer Hauptkomponentenanalyse eines Datensatzes $X \in \mathbb{R}^{m \times n}$ hervorgegangene transformierte Datensatz. Welche Aussage trifft **nicht** zu?

a) Die Dimensionalität des transformierten Datensatzes $\tilde{X}$ ist $n \times m$.

b) Die paarweisen Strichprobenkorrelationen der Features in $\tilde{X}$ sind Null.

c) Die Gesamtvarianz der Features in $\tilde{X}$ entspricht der Gesamtvarianz der Features in X.

d) Die erste Zeile von $\tilde{X}$ entspricht dem transformierten Feature mit der größten Varianz.

9.8 Kinderwitz

Mit welchem Fahrzeug fährt der Panda am liebsten?

Antwort: Mit dem Bam-Bus.