# Übung 5 Hauptkomponentenanalyse II & Neuronale Netze

## Aufgabe 1

Die folgende Tabelle enthält die Ergebnisse der 25 Teilnehmerinnen beim Siebenkampf der Frauen bei den Olympischen Sommerspielen 1988 in Seoul:

| | hurdles | highjump | shot | run200m | longjump | javelin | run800m | score |
|---|---|---|---|---|---|---|---|---|
| Joyner-Kersee (USA) | 12.69 | 1.86 | 15.80 | 22.56 | 7.27 | 45.66 | 128.51 | 7291 |
| John (GDR) | 12.85 | 1.80 | 16.23 | 23.65 | 6.71 | 42.56 | 126.12 | 6897 |
| Behmer (GDR) | 13.20 | 1.83 | 14.20 | 23.10 | 6.68 | 44.54 | 124.20 | 6858 |
| Sablovskaite (URS) | 13.61 | 1.80 | 15.23 | 23.92 | 6.25 | 42.78 | 132.24 | 6540 |
| Choubenkova (URS) | 13.51 | 1.74 | 14.76 | 23.93 | 6.32 | 47.46 | 127.90 | 6540 |
| Schulz (GDR) | 13.75 | 1.83 | 13.50 | 24.65 | 6.33 | 42.82 | 125.79 | 6411 |
| Fleming (AUS) | 13.38 | 1.80 | 12.88 | 23.59 | 6.37 | 40.28 | 132.54 | 6351 |
| Greiner (USA) | 13.55 | 1.80 | 14.13 | 24.48 | 6.47 | 38.00 | 133.65 | 6297 |
| Lajbnerova (CZE) | 13.63 | 1.83 | 14.28 | 24.86 | 6.11 | 42.20 | 136.05 | 6252 |
| Bouraga (URS) | 13.25 | 1.77 | 12.62 | 23.59 | 6.28 | 39.06 | 134.74 | 6252 |
| Wijnsma (HOL) | 13.75 | 1.86 | 13.01 | 25.03 | 6.34 | 37.86 | 131.49 | 6205 |
| Dimitrova (BUL) | 13.24 | 1.80 | 12.88 | 23.59 | 6.37 | 40.28 | 132.54 | 6171 |
| Scheider (SWI) | 13.85 | 1.86 | 11.58 | 24.87 | 6.05 | 47.50 | 134.93 | 6137 |
| Braun (FRG) | 13.71 | 1.83 | 13.16 | 24.78 | 6.12 | 44.58 | 142.82 | 6109 |
| Ruotsalainen (FIN) | 13.79 | 1.80 | 12.32 | 24.61 | 6.08 | 45.44 | 137.06 | 6101 |
| Yuping (CHN) | 13.93 | 1.86 | 14.21 | 25.00 | 6.40 | 38.60 | 146.67 | 6087 |
| Hagger (GB) | 13.47 | 1.80 | 12.75 | 25.47 | 6.34 | 35.76 | 138.48 | 5975 |
| Brown (USA) | 14.07 | 1.83 | 12.69 | 24.83 | 6.13 | 44.34 | 146.43 | 5972 |
| Mulliner (GB) | 14.39 | 1.71 | 12.68 | 24.92 | 6.10 | 37.76 | 138.02 | 5746 |
| Hautenauve (BEL) | 14.04 | 1.77 | 11.81 | 25.61 | 5.99 | 35.68 | 133.90 | 5734 |
| Kytola (FIN) | 14.31 | 1.77 | 11.66 | 25.69 | 5.75 | 39.48 | 133.35 | 5686 |
| Geremias (BRA) | 14.23 | 1.71 | 12.95 | 25.50 | 5.50 | 39.64 | 144.02 | 5508 |
| Hui-Ing (TAI) | 14.85 | 1.68 | 10.00 | 25.23 | 5.47 | 39.14 | 137.30 | 5290 |
| Jeong-Mi (KOR) | 14.53 | 1.71 | 10.83 | 26.61 | 5.50 | 39.26 | 139.17 | 5289 |
| Launa (PNG) | 16.42 | 1.50 | 11.78 | 26.16 | 4.88 | 46.38 | 163.43 | 4566 |

Zum Zwecke einer einfacheren Interpretation der Ergebnisse wurden die Variablen transformiert, so dass bei allen sieben Disziplinen größere Werte für bessere Leistungen stehen. Dies liefert die folgende Tabelle:

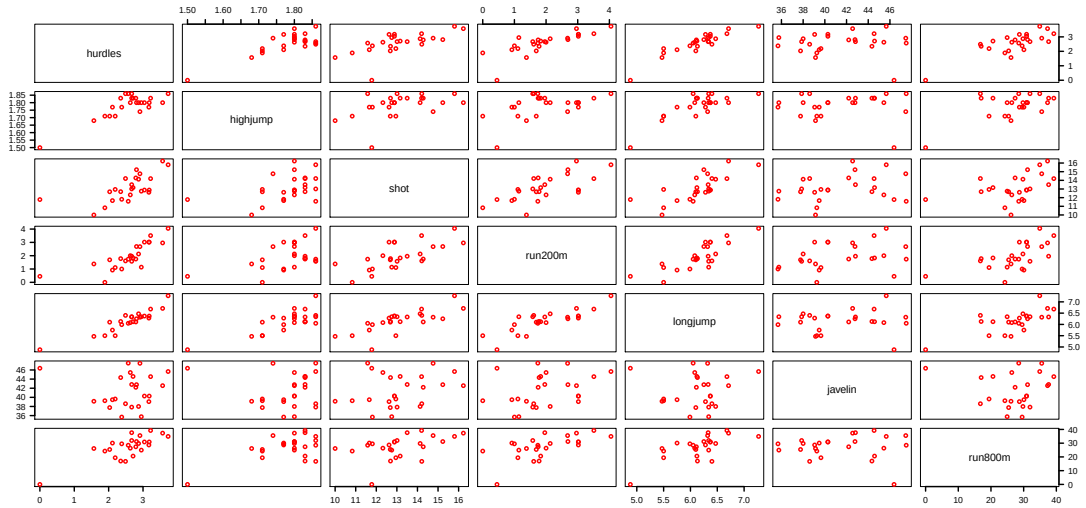
```
> data("heptathlon",package="HSAUR2")
> heptathlon$hurdles <- with(heptathlon, max(hurdles)-hurdles)
> heptathlon$run200m <- with(heptathlon, max(run200m)-run200m)
> heptathlon$run800m <- with(heptathlon, max(run800m)-run800m)
> heptathlon
                     hurdles highjump  shot run200m longjump javelin run800m score
Joyner-Kersee (USA)    3.73     1.86 15.80    4.05     7.27   45.66   34.92  7291
John (GDR)             3.57     1.80 16.23    2.96     6.71   42.56   37.31  6897
Behmer (GDR)           3.22     1.83 14.20    3.51     6.68   44.54   39.23  6858
Sablovskaite (URS)     2.81     1.80 15.23    2.69     6.25   42.78   31.19  6540
Choubenkova (URS)      2.91     1.74 14.76    2.68     6.32   47.46   35.53  6540
Schulz (GDR)           2.67     1.83 13.50    1.96     6.33   42.82   37.64  6411
Fleming (AUS)          3.04     1.80 12.88    3.02     6.37   40.28   30.89  6351
Greiner (USA)          2.87     1.80 14.13    2.13     6.47   38.00   29.78  6297
Lajbnerova (CZE)       2.79     1.83 14.28    1.75     6.11   42.20   27.38  6252
Bouraga (URS)          3.17     1.77 12.62    3.02     6.28   39.06   28.69  6252
Wijnsma (HOL)          2.67     1.86 13.01    1.58     6.34   37.86   31.94  6205
Dimitrova (BUL)        3.18     1.80 12.88    3.02     6.37   40.28   30.89  6171
Scheider (SWI)         2.57     1.86 11.58    1.74     6.05   47.50   28.50  6137
Braun (FRG)            2.71     1.83 13.16    1.83     6.12   44.58   20.61  6109
Ruotsalainen (FIN)     2.63     1.80 12.32    2.00     6.08   45.44   26.37  6101
Yuping (CHN)           2.49     1.86 14.21    1.61     6.40   38.60   16.76  6087
Hagger (GB)            2.95     1.80 12.75    1.14     6.34   35.76   24.95  5975
Brown (USA)            2.35     1.83 12.69    1.78     6.13   44.34   17.00  5972
Mulliner (GB)          2.03     1.71 12.68    1.69     6.10   37.76   25.41  5746
Hautenauve (BEL)       2.38     1.77 11.81    1.00     5.99   35.68   29.53  5734
Kytola (FIN)           2.11     1.77 11.66    0.92     5.75   39.48   30.08  5686
Geremias (BRA)         2.19     1.71 12.95    1.11     5.50   39.64   19.41  5508
Hui-Ing (TAI)          1.57     1.68 10.00    1.38     5.47   39.14   26.13  5290
Jeong-Mi (KOR)         1.89     1.71 10.83    0.00     5.50   39.26   24.26  5289
Launa (PNG)            0.00     1.50 11.78    0.45     4.88   46.38    0.00  4566
```

Die Korrelationsmatrix der transformierten Daten ist gegeben durch:

```
> score <- which(colnames(heptathlon) == "score")
> round(cor(heptathlon[,-score]), 2)

         hurdles highjump shot run200m longjump javelin run800m
hurdles     1.00     0.81 0.65    0.77     0.91    0.01    0.78
highjump    0.81     1.00 0.44    0.49     0.78    0.00    0.59
shot        0.65     0.44 1.00    0.68     0.74    0.27    0.42
run200m     0.77     0.49 0.68    1.00     0.82    0.33    0.62
longjump    0.91     0.78 0.74    0.82     1.00    0.07    0.70
javelin     0.01     0.00 0.27    0.33     0.07    1.00   -0.02
run800m     0.78     0.59 0.42    0.62     0.70   -0.02    1.00
```

a) Interpretieren Sie die Werte der Korrelationsmatrix.

b) Die folgende Abbildung zeigt die Streudiagrammmatrix für die sieben Variablen. Interpretieren Sie die Streudiagrammmatrix.




c) Wenn die Teilnehmerin aus Papua-Neuguinea (PNG) von der Betrachtung ausgeschlossen wird, erhält man die untenstehende Korrelationsmatrix. Interpretieren Sie das Ergebnis.

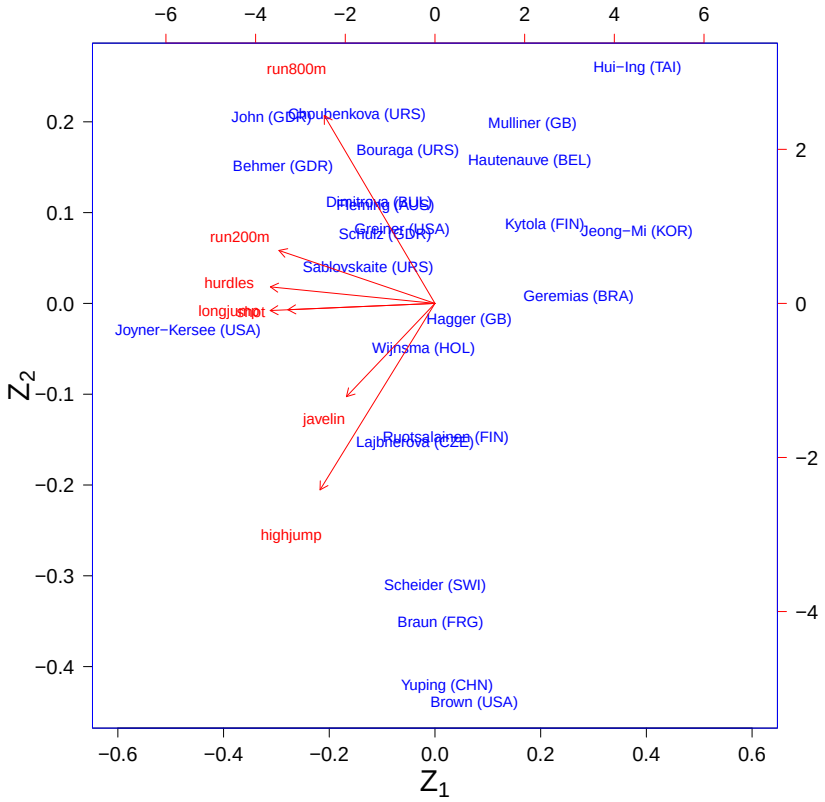
```
> heptathlon<-heptathlon[-grep("PNG", rownames(heptathlon)),]
> score<-which(colnames(heptathlon) == "score")
> round(cor(heptathlon[,-score]), 2)

         hurdles highjump shot run200m longjump javelin run800m
hurdles     1.00     0.58 0.77    0.83     0.89    0.33    0.56
highjump    0.58     1.00 0.46    0.39     0.66    0.35    0.15
shot        0.77     0.46 1.00    0.67     0.78    0.34    0.41
run200m     0.83     0.39 0.67    1.00     0.81    0.47    0.57
longjump    0.89     0.66 0.78    0.81     1.00    0.29    0.52
javelin     0.33     0.35 0.34    0.47     0.29    1.00    0.26
run800m     0.56     0.15 0.41    0.57     0.52    0.26    1.00
```

d) Die Teilnehmerin aus Papua-Neuguinea (PNG) wird aus der Datenmatrix entfernt. Ferner wird die Datenmatrix zentriert und skaliert, so dass alle sieben Variablen die Varianz 1 haben. Eine Hauptkomponentenanalyse liefert das untenstehende Ergebnis. Interpretieren Sie die ersten beiden Hauptkomponenten $Z_1$ und $Z_2$.

```
> op<-options(digits = 2)
> heptathlon_pca<-prcomp(heptathlon[,-score], center=TRUE, scale = TRUE)
> print(heptathlon_pca)

Standard deviations:
[1] 2.08 0.95 0.91 0.68 0.55 0.34 0.26

Rotation:
             PC1    PC2   PC3    PC4    PC5    PC6    PC7
hurdles  -0.45  0.058 -0.17  0.048 -0.199  0.847 -0.070
highjump -0.31 -0.651 -0.21 -0.557  0.071 -0.090  0.332
shot     -0.40 -0.022 -0.15  0.548  0.672 -0.099  0.229
run200m  -0.43  0.185  0.13  0.231 -0.618 -0.333  0.470
longjump -0.45 -0.025 -0.27 -0.015 -0.122 -0.383 -0.749
javelin  -0.24 -0.326  0.88  0.060  0.079  0.072 -0.211
run800m  -0.30  0.657  0.19 -0.574  0.319 -0.052  0.077

> summary(heptathlon_pca)

Importance of components:
                         PC1   PC2   PC3    PC4    PC5    PC6     PC7
Standard deviation     2.079 0.948 0.911 0.6832 0.5462 0.3375 0.26204
Proportion of Variance 0.618 0.128 0.119 0.0667 0.0426 0.0163 0.00981
Cumulative Proportion  0.618 0.746 0.865 0.9313 0.9739 0.9902 1.00000
```

e) Der Korrelationskoeffizient zwischen den Scores der ersten Hauptkomponente $Z_1$ und der Variablen `score`, d.h. der im Siebenkampf erzielten Gesamtpunktzahl, beträgt $-0{,}99$. Interpretieren Sie dieses Ergebnis.

f) Die untenstehende Abbildung zeigt einen sog. Biplot. Er ermöglicht es in einer gemeinsamen Abbildung die Scores der ersten beiden Hauptkomponenten $Z_1$ und $Z_2$ für die 24 Beobachtungen/Teilnehmerinnen (ohne die Teilnehmerin aus Papua-Neuguinea (PNG)) sowie die Ladungen der 7 Variablen (Disziplinen) in den ersten beiden Hauptkomponenten $Z_1$ und $Z_2$ darzustellen.




Bei der Betrachtung von Biplots sind folgende Punkte zu beobachten:

- Je näher zwei Punkte im Biplot beieinander liegen, desto ähnlicher sind sich die Beobachtungen.
- Die Länge der Pfeile ist proportional zur Standardabweichnung der Variablen (Disziplinen).
- Der Kosinus des Winkels zwischen zwei Pfeilen (Variablen) approximiert die Korrelation zwischen diesen beiden Variablen. D.h. umso kleiner der Winkel ist, desto größer ist die positive Korrelation. Insbesondere bedeutet ein Winkel von 0° näherungsweise perfekte positive Korrelation, ein Winkel von 90° näherungsweise Unkorreliertheit und ein Winkel von 180° näherungsweise perfekte negative Korrelation.
- Liegt der $i$-te Punkt in derselben Richtung wie der $j$-te Pfeil, dann ist die $i$-te Beobachtung bzgl. der $j$-ten Variablen überdurchschnittlich.
- Liegt der $i$-te Punkt in entgegensetzter Richtung wie der $j$-te Pfeil, dann ist die $i$-te Beobachtung bzgl. der $j$-ten Variablen unterdurchschnittlich.
- Die Qualität eines Biplots hängt davon ab, welcher Anteil der Gesamtvariation durch die beiden ersten Hauptkomponenten erklärt wird.

Interpretieren Sie nun den Biplot.

**Aufgabe 2**

Es wird ein künstliches Neuronales Netz betrachtet, das nur aus einem Input und einem Output Layer besteht. Mit Hilfe der Inputs „Alter“ ($x_1$), „Einkommen“ ($x_2$) und „Anzahl der bisherigen Käufe“ ($x_3$) sagt es für die Kunden eines Unternehmens vorher, ob Werbung zum Kauf eines bestimmten Produkts führt. Für den Output gilt dabei

$$y = \begin{cases} 0 & \text{falls das Produkt nicht gekauft wird} \\ 1 & \text{falls das Produkt gekauft wird} \end{cases}.$$

Der Bias und die Gewichte für die Inputs wurden bereits mit Hilfe von Kundendaten geschätzt und sind gegeben durch

$$w_0 = 0, \quad w_1 = -0{,}1, \quad w_2 = 0{,}6 \quad \text{und} \quad w_3 = 0{,}7.$$

Als Output Aktivierungsfunktion wird die Logistische Funktion

$$g(x) = \frac{1}{1 + \exp(-x)}$$

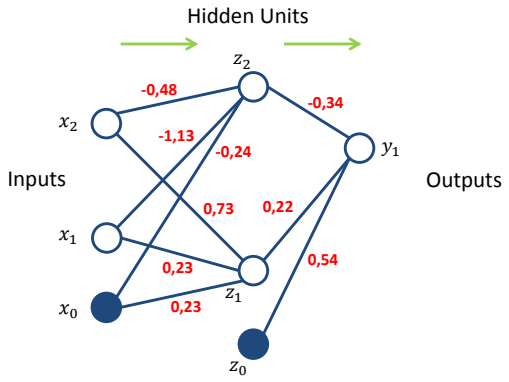
verwendet. Bestimmen Sie für die folgenden Kunden die Wahrscheinlichkeit (gerundet auf vier Nachkommastellen), dass sie das Produkt kaufen:

| | Kunde 1 | Kunde 2 | Kunde 3 |
|---|---|---|---|
| Alter $x_1$ | 20 | 30 | 40 |
| Einkommen $x_2$ | 6 | 5 | 1 |
| Anzahl bisheriger Käufe $x_3$ | 1 | 0 | 3 |

Interpretieren Sie ferner die erhaltenen Ergebnisse.

**Aufgabe 3**

Es wird ein Single Layer Perceptron mit zwei Inputvariablen $x_1$ und $x_2$, zwei Hidden Variablen $z_1$ und $z_2$ sowie einer Outputvariablen $y_1$ betrachtet. Die beiden Aktivierungsfunktionen sind $g^{(1)}(v) = g^{(2)}(v) = \tanh(v)$ und Initialisierungen für die Gewichte zwischen den Variablen sind durch die Werte im folgenden Netzwerkdiagramm gegeben:




a) Für die beiden Inputvariablen liegen die Beobachtungen $x_1 = 0{,}9$ und $x_2 = 0{,}9$ vor. Führen Sie eine Forward Propagation für das künstliche Neuronale Netz durch und runden Sie die dabei resultierenden Werte auf 2 Nachkommastellen.

b) Zusätzlich sei nun für die Outputvariable die Beobachtung $t_1 = -0{,}9$ gegeben. Führen Sie eine Backward Propagation für das künstliche Neuronale Netz durch und runden Sie die dabei resultierenden Werte auf 2 Nachkommastellen. Verwenden Sie dabei im Gradientenabstiegsverfahren die Schrittweite $\eta = 0{,}3$.