

7 Kanonische Korrelationsanalyse

7.1 Mathematische Grundlagen

Mit dem Begriff der univariaten Korrelation bzw. des Korrelationskoeffizienten können Zusammenhänge zwischen einer einzelnen unabhängigen Variable (UV) und einer einzelnen abhängigen Variable (AV) erfasst werden. Der Begriff der kanonischen Korrelation verallgemeinert diesen Begriff auf multivariate UVs und AVs, d.h. auf Zusammenhänge zwischen Mengen von UVs und Mengen von AVs.

Das Ziel der *kanonischen Korrelationsanalyse* ist, für Realisierungen von Zufallsvektoren

$$x \in \mathbb{R}^{m_x} \quad \text{und} \quad y \in \mathbb{R}^{m_y} \quad (1)$$

mit $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ Linearkombinationen der Form

$$\xi = a^T x \quad \text{und} \quad v = b^T y \quad (2)$$

zu finden, sodass die Korrelation $\rho(\xi, v)$ maximal ist. In der Praxis einer faktischen Datenanalyse wird dazu wie folgt vorgegangen.

Gegeben seien eine Datenmatrix $X \in \mathbb{R}^{m_x \times n}$, deren Spalten als Realisierungen von x verstanden werden, und eine Datenmatrix $Y \in \mathbb{R}^{m_y \times n}$, deren Spalten als Realisierungen von y verstanden werden.

X und Y lassen sich zu einer Datenmatrix $Z \in \mathbb{R}^{m \times n}$ zusammenfügen lassen, wobei $m = m_x + m_y$. Wir bestimmen für diese Datenmatrix Z die *Stichprobenkovarianzmatrix* C , die sich entlang der Trennung zwischen den m_x UVs und den m_y AVs *partitionieren* lässt

$$C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m}, \quad (3)$$

wobei $C_{xx} \in \mathbb{R}^{m_x \times m_x}$, $C_{xy} \in \mathbb{R}^{m_x \times m_y}$, $C_{yx} \in \mathbb{R}^{m_y \times m_x}$ und $C_{yy} \in \mathbb{R}^{m_y \times m_y}$. Kern der kanonischen Korrelationsanalyse ist dann die Matrix

$$\hat{K} = C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2} \in \mathbb{R}^{m_x \times m_y} \quad (4)$$

und es kann gezeigt werden, dass sich die *kanonischen Koeffizientenvektoren*, d.h. die optimalen Linearkombinationen von x und y , als

$$\hat{a}_i = C_{xx}^{-1/2} \hat{\alpha}_i \quad \text{und} \quad \hat{b}_i = C_{yy}^{-1/2} \hat{\beta}_i \quad (5)$$

schätzen lassen und dass sich die *kanonischen Korrelationen*, d.h. die maximierten Korrelationen zwischen ξ und v , als

$$\hat{\rho}_i = \sqrt{\hat{\lambda}_i} \quad (6)$$

schätzen lassen, wobei $\hat{\alpha}_i$ der i -te Eigenvektor von $\hat{K}\hat{K}^T$, $\hat{\beta}_i$ der i -te Eigenvektor von $\hat{K}\hat{K}^T$ und $\sqrt{\hat{\lambda}_i}$ der i -te Singulärwert von $\hat{K}$ ist.

7.2 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

# Datenmatrizen extrahieren
rows   = startsWith(D$subject, 'subA') # Personen aus Studie A
colX   = c('novelty.FADE','novelty.SAME','memory.FADE', 'memory.SAME')
colY   = c('V.left.Hippocampal.tail', 'V.left.Subiculum', 'V.left.CA1', 'V.left.CA3', 'V.left.CA4',
           'V.right.Hippocampal.tail', 'V.right.Subiculum', 'V.right.CA1', 'V.right.CA3', 'V.right.CA4')
n      = sum(rows)                 # Anzahl Datenpunkte
X      = t(as.matrix(D[rows,colX])) # 4 x n Datenmatrix X
Y      = t(as.matrix(D[rows,colY])) # 10 x n Datenmatrix Y
print(dim(X))                     # Überprüfung Datenmatrix X
print(dim(Y))                     # Überprüfung Datenmatrix Y
```

7.3 Erste Programmieraufgabe

Berechnen Sie die univariate Stichprobenkorrelation (d.h. einen einfachen Korrelationskoeffizienten) zwischen dem Novelty-FADE-Score und dem Novelty-SAME-Score. Gehen Sie dazu wie folgt vor:

- Extrahieren Sie die erste und die zweite Zeile der Datenmatrix X als Variablen x_i und y_i .
- Berechnen Sie die Stichprobenmittel der x_i und der y_i , indem alle Werte aufsummieren und das Ergebnis mit $1/n$ multiplizieren.
- Berechnen Sie die Stichprobenstandardabweichungen der x_i und der y_i , indem Sie die Terme $(x_i - \bar{x})^2$ bzw. $(y_i - \bar{y})^2$ aufsummieren, dann mit $1/(n - 1)$ multiplizieren und schließlich die Quadratwurzel ziehen.

- Berechnen Sie die Stichprobenkovarianz der x_i mit den y_i , indem Sie die Terme $(x_i - \bar{x})(y_i - \bar{y})$ aufsummieren und das Ergebnis mit $1/(n - 1)$ multiplizieren.
- Berechnen Sie die Stichprobenkorrelation zwischen den x_i und den y_i , indem Sie die Stichprobenkovarianz durch das Produkt der Stichprobenstandardabweichungen teilen.
- Überprüfen Sie Ihre Berechnung der Stichprobenkorrelation, indem Sie die R-Funktion `cor()` auf `x_i` und `y_i` anwenden.
- Geben Sie die Resultate ihrer Analyse aus. Sie sollten folgende Ergebnisse erhalten:

Stichprobenkorrelation (manuell):

```
[1] -0.6626767
```

Stichprobenkorrelation (automatisch):

```
[1] -0.6626767
```

7.4 Analyse in R

Im zweiten Teil wollen wir nun mit den Datenmatrizen X und Y eine kanonische Korrelationsanalyse durchführen. Erklären Sie dazu zunächst den folgenden R-Code und die Variablen, die er erzeugt:

```
# Berechnung der Stichprobenkovarianzmatrix
m_x      = nrow(X)                # Dimensionalität der x_i
m_y      = nrow(Y)                # Dimensionalität der y_i
Z        = rbind(X,Y)             # (m_x+m_y) x n Datenmatrix
I_n      = diag(n)                # Einheitsmatrix I_n
J_n      = matrix(rep(1,n^2), nrow = n) # Einsmatrix 1_{nn}
C        = (1/(n-1))*(Z %*% (I_n-(1/n)*J_n) %*% t(Z)) # Stichprobenkovarianzmatrix

# Partitionierung der Stichprobenkovarianzmatrix
C_xx     = C[1:m_x,1:m_x]
C_xy     = C[1:m_x,(m_x+1):(m_x+m_y)]
C_yx     = C[(m_x+1):(m_x+m_y),1:m_x]
C_yy     = C[(m_x+1):(m_x+m_y),(m_x+1):(m_x+m_y)]
```

7.5 Zweite Programmieraufgabe

Berechnen Sie die multivariate Stichprobenkorrelation (d.h. die kanonischen Korrelationen) zwischen FADE/SAME-Scores und Hippocampus-Subregionen-Volumen. Gehen Sie dazu wie folgt vor:

- Laden Sie das R-Paket `expm`.
- Berechnen Sie $\hat{K}$ anhand der obigen Formel. Die inverse Quadratwurzel einer Matrix ($A^{-1/2}$) lässt sich mit in R mittels `sqrtn(solve)` berechnen.
- Führen Sie eine Singulärwertzerlegung der Matrix $\hat{K}$ durch. Orientieren Sie sich hierbei an Abschnitt 3.5 des Arbeitsblatts (3) *Eigenanalyse* und speichern Sie die Ergebnisse in die Variablen `A_hat`, `Lambda_hat` und `B_hat`.
- Berechnen Sie die kanonischen Koeffizientenvektoren $\hat{a}_i$ und $\hat{b}_i$ für $i = 1, \dots, \min(m_x, m_y)$, indem Sie die Matrixprodukte `sqrtn(solve(C_xx)) %*% A_hat` bzw. `sqrtn(solve(C_yy)) %*% B_hat` bilden.
- Berechnen Sie die kanonischen Korrelationen $\hat{\rho}_i$, indem Sie die Werte aus `Lambda_hat` extrahieren.
- Geben Sie die zur ersten kanonischen Korrelation gehörenden Werte aus. Sie sollten folgende Ergebnisse erhalten:

```
rho_hat_1 : 0.3723013
a_hat_1   : -0.8918519 -0.6201265 0.4046517 -0.7096886
b_hat_1   : -0.01139014 0.006110925 -0.00320041 0.02827114 -0.02225319 0.00314746 0.005155459 -0.01369069 0.0041
```

- Überprüfen Sie Ihre Berechnung der kanonischen Korrelationen, indem Sie folgenden Code mit der in R eingebauten Funktion `cancor()` ausführen:

```
# Kanonische Korrelationsanalyse (automatisch)
XT = t(X)
YT = t(Y)
cca = cancor(XT, YT)
cat( "rho_hat_1 : ", cca$cor[1],
     "\nx_coef   : ", cca$xcoef[,1],
     "\ny_coef   : ", cca$ycoef[,1], "\n")
```

7.6 Lückentext

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

Lückentext: Die Stichprobenkorrelation zwischen Novelty-FADE-Score und Novelty-SAME-Score beträgt _____. Der aus der CSV-Datei extrahierte Datensatz enthält insgesamt _____ Datenpunkte bzw. Personen. Die Datenmatrix X besteht aus _____-dimensionalen Zufallsvektoren und die Datenmatrix Y besteht aus _____-dimensionalen Zufallsvektoren, sodass die Datenmatrix Z die Dimensionalität _____ (Zeilen $\times$ Spalten) hat. Die erste kanonische Korrelation ist _____, dies bedeutet: _____. Die kanonischen Koeffizienten weisen darauf hin, dass im Bereich der FADE/SAME-Scores die Variable _____ und im Bereich der Hippocampus-Subregion-Volumen die Variable _____ am stärksten in diese Linearkombinationen eingehen.

7.7 Mögliche Klausurfrage

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

Frage: Was versteht man, im Rahmen einer Kanonischen Korrelationsanalyse (KKA), unter dem “besten Prädiktor”?

- a) Der beste Prädiktor ist die nach Schätzung der KKA sich ergebende Linearkombination der unabhängigen Variablen, die die Korrelation mit einer Linearkombination der abhängigen Variablen maximiert.
- b) Der beste Prädiktor ist die nach Schätzung der KKA sich ergebende Linearkombination der abhängigen Variablen, die die Korrelation mit einer Linearkombination der unabhängigen Variablen maximiert.
- c) Der beste Prädiktor ist diejenige abhängige oder unabhängige Variable, die im Rahmen der Schätzung der KKA das höchste Gewicht im kanonischen Koeffizientenvektor erhält.
- d) Der beste Prädiktor ist der geschätzte Wert der ersten kanonischen Korrelation.

7.8 Kinderwitz

Wie heißt eine gelbe Kanone?

Antwort: Banane.