6 Multivariate Varianzanalyse

6.1 Mathematische Grundlagen

Gegeben sei das *Modell der einfaktoriellen multivariaten Varianzanalyse*

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{mit} \quad \varepsilon_{ij} \sim N(0_m, \Sigma) \quad \text{u.i.v.} \,, \tag{1}$$

wobei $i = 1, \ldots, p$ ein Index über Gruppen, $j = 1, \ldots, n_i$ ein Index über experimentelle Einheiten innerhalb einer Gruppe und m die Dimensionalität eines Datenvektors $y_{ij} \in \mathbb{R}^m$ ist.

Wir betrachten die *multivariate Quadratsummenzerlegung* für dieses Modell mit

$$\begin{aligned} T &= \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}\right) \left(y_{ij} - \bar{y}\right)^{\mathrm{T}} \\ B &= \sum_{i=1}^{p} n_i \left(\bar{y}_i - \bar{y}\right) \left(\bar{y}_i - \bar{y}\right)^{\mathrm{T}} \\ W &= \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right) \left(y_{ij} - \bar{y}_i\right)^{\mathrm{T}} \end{aligned} \tag{2}$$

und die darauf basierende *Wilks'-Lambda-Statistik*

$$\Lambda = \frac{|W|}{|B + W|} \,, \tag{3}$$

wobei B die *between sum-of-squares Matrix*, W die *within sum-of-squares Matrix* und $T = B + W$ die *total sum of-squares Matrix* ist.

Das Theorem zu speziellen Verteilungen von Wilks'-Λ-Transformationen gibt Spezialfälle von Datendimension m und Gruppenanzahl p an, in denen die Verteilung einer Transformation von Λ unter der Nullhypothese $H_0 : \mu_1 = \ldots = \mu_p$ exakt bekannt ist. Es besagt, dass für den in den ersten beiden Tabellenspalten aufgeführten Spezialfall die in der dritten Tabellenspalte genannte Teststatistik bei Vorliegen der Nullhypothese einer f-Verteilung mit den Freiheitsgradparametern in der vierten Tabellenspalte folgt:

Datendimension m	Gruppenanzahl p	Statistik	f-Verteilungsparameter
beliebig	2	$\frac{1-\Lambda}{\Lambda} \frac{n-p-m+1}{m}$	$m,\ n-p-m+1$
beliebig	3	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n-p-m+1}{m}$	$2m,\ 2(n-p-m+1)$
1	beliebig	$\frac{1-\Lambda}{\Lambda} \frac{n-p}{p-1}$	$p-1,\ n-p$
2	beliebig	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n-p-1}{p-1}$	$2(p-1),\ 2(n-p-1)$

Damit gilt z.B. für $m > 2$ und $p = 3$ (zweite Tabellenzeile): Überschreitet $\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n-p-m+1}{m}$ den kritischen Wert k_{α_0}, der mit einem Signifikanzniveau α_0 aus der inversen kumulativen Verteilungsfunktion der f-Verteilung mit Freiheitsgradparametern $2m$ und $2(n-p-m+1)$ errechnet werden kann, wird die Nullhypothese identischer Erwartungswerte über Gruppen abgelehnt. Anderenfalls wird sie nicht abgelehnt.

6.2 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',             # Definition Variablen
          'memory.FADE', 'memory.SAME')
n     = sum(rows)                                    # Anzahl Datenpunkte
m     = length(cols)                                 # Anzahl Variablen
Y     = t(as.matrix(D[rows,cols]))                   # m x n Datenmatrix
print(dim(Y))                                        # Überprüfung Datenmatrix
```

6.3 Erste Programmieraufgabe

Wir wollen systematische Unterschiede der FADE-Scores und SAME-Scores zwischen den drei Altersgruppen "junge Personen" (bis 35 Jahre), "mittelalte Personen" (50-59 Jahre) und "ältere Personen" (ab 60 Jahre) untersuchen.

Wir werden dazu zunächst eine Variable x definieren, die die unabhängige Variable "Altersgruppe" repräsentiert und die Zugehörigkeit der Datenpunkte zu den verschiedenen Studiengruppen kodiert. Gehen Sie dazu wie folgt vor:

- Extrahieren Sie die Variable "age" aus dem Dataframe, aber nur für diejenigen Datenpunkte, die durch die Variable `rows` beschrieben werden.
- Bilden Sie drei logische Variablen `g1`, `g2`, `g3`, d.h. Vektoren von derselben Länge wie die Datenmatrix, die jeweils `TRUE` sind, wenn der entsprechende Datenpunkt zu Gruppe 1, 2, 3 gehört. Beispielsweise ist `g1 = (x <= 35)`.
- Bilden Sie einen Vektor, der für jeden Datenpunkt den entsprechenden Gruppenindex enthält, d.h. einen Vektor, der 1 ist, wenn Altersgruppe `g1` vorliegt, 2 ist, wenn Altersgruppe `g2` vorliegt und 3 ist, wenn Altersgruppe `g1` vorliegt. Sie erreichen dies z.B., indem Sie die

logischen Vektoren mit den Gruppenindizes multiplizieren und aufsummieren. Speichern Sie das Ergebnis in die Variable `x`.

- Speichern Sie das Maximum von `x`, d.h. die Anzahl der Gruppen, in die Variable `p`.
- Geben Sie die ersten 100 Einträge des Vektors `x` aus. Sie sollten folgende Ergebnisse erhalten:

```
  [1] 3 3 1 3 3 3 3 2 2 2 1 1 2 3 3 1 3 1 2 1 1 2 1 3 1 1 1 3 1 3 3 3 3 3 1 3 2
 [38] 3 1 1 3 1 1 2 3 3 1 3 3 3 1 3 1 2 1 1 1 2 3 1 2 1 1 2 1 3 3 3 1 3 3 3 3 3
 [75] 1 1 3 2 3 3 1 3 3 2 1 3 1 1 1 3 3 3 3 1 1 1 1 1 3 3
```

6.4 Analyse in R

Im zweiten Teil wollen wir nun mit der Datenmatrix Y und dem Gruppenvektor x eine einfaktorielle multivariate Varianzanalyse durchführen. Erklären Sie dazu zunächst den folgenden R-Code und die Funktionen, die er definiert:

```r
# Parameter schätzen
estimate = function(Y,x) {

    # Datendimensionalität
    n = ncol(Y)                                     # Anzahl Datenpunkte
    m = nrow(Y)                                     # Anzahl Variablen
    p = max(x)                                      # Anzahl Gruppen

    # Parameterschätzer
    mu_hat    = matrix(rep(0,m*p), nrow = m)        # Erwartungswertparameter
    Sigma_hat = matrix(rep(0,m*m), nrow = m)        # Kovarianzmatrixparameter
    for (i in 1:p) {
        Y_i        = Y[,x==i]                       # Datenmatrix Gruppe i
        n_i        = ncol(Y_i)                      # Anzahl Datenpunkte Gruppe i
        mu_hat[,i] = as.matrix(rowMeans(Y_i))       # Erwartungswerte Gruppe i
        for (j in 1:n_i) {
            Sigma_hat = Sigma_hat + (Y_i[,j] - mu_hat[i]) %*% t(Y_i[,j] - mu_hat[i])
        }
    }
    Sigma_hat = (1/(n-p)) * Sigma_hat

    # Funktionswerte
    return(list(mu_hat = mu_hat, Sigma_hat = Sigma_hat))
}

# Multivariate Quadratsummenzerlegung
sumofsqr = function(Y,x) {

    # Datendimensionalität
    n = ncol(Y)                                     # Anzahl Datenpunkte
    m = nrow(Y)                                     # Anzahl Variablen
```

```
    p = max(x)                                          # Anzahl Gruppen

    # sum-of-squares Matrizen
    y_bar   = as.matrix(rowMeans(Y))                    # Gesamtstichprobenmittel
    y_i_bar = matrix(rep(0,m*p), nrow = m)              # Gruppenstrichprobenmittel
    B       = matrix(rep(0,m*m), nrow = m)              # between-group sum-of-squares Matrix
    W       = matrix(rep(0,m*m), nrow = m)              # within-group sum-of-squares Matrix
    for (i in 1:p) {
        Y_i         = Y[,x==i]                          # Datenmatrix Gruppe i
        n_i         = ncol(Y_i)                         # Anzahl Datenpunkte Gruppe i
        y_i_bar[,i] = as.matrix(rowMeans(Y_i))          # Erwartungswerte Gruppe i
        for (j in 1:n_i) {
            B = B + (y_i_bar[,i] - y_bar) %*% t(y_i_bar[,i] - y_bar)
            W = W + (Y_i[,j] - y_i_bar[,i]) %*% t(Y_i[,j] - y_i_bar[,i])
        }
    }
    T         = B + W                                   # total sum-of-squares Matrix

    # Funktionswerte
    return(list(T = T, B = B, W = W))
}
```

6.5 Zweite Programmieraufgabe

Nutzen Sie die Funktionen der Parameterschätzung und der multivariaten Quadratsummenzerlegung, um die MANOVA-Modellparameter zu schätzen. Gehen Sie dazu wie folgt vor:

- Wenden Sie die Funktion `estimate` auf Y und x an und extrahieren Sie die Felder `mu_hat` sowie `Sigma_hat`.
- Wenden Sie die Funktion `sumofsqr` auf Y und x an und extrahieren Sie die Felder `T` sowie `B` und `W`.
- Geben Sie die Schätzer für die Gruppenerwartungswertparameter und den Kovarianzmatrixparameter aus. Sie sollten folgende Ergebnisse erhalten:

```
          young middle.aged      older
[1,] -1.88580755  -1.9468262 -1.8505495
[2,] -0.07335094  -0.2948429 -0.4256090
[3,] -1.46600660  -0.9582262 -0.8179117
[4,] -0.06110755  -1.0596524 -1.1492261
```

```
     novelty.FADE novelty.SAME memory.FADE memory.SAME
[1,]    1.0195042   -0.3632044   0.1869198   0.2602081
[2,]   -0.3632044    2.2154334   0.6098182   1.5811919
[3,]    0.1869198    0.6098182   0.6675496   0.2481929
[4,]    0.2602081    1.5811919   0.2481929   2.0553726
```

Führen Sie nun eine einfaktorielle multivariate Varianzanalyse (MANOVA) durch. Gehen Sie dazu wie folgt vor:

- Legen Sie das Signifikanzniveau $\alpha_0 = 0.05$ fest.
- Berechnen Sie die Wilks'-Lambda-Statistik Λ anhand der obigen Formel. Die Determinante ist in R als die Funktion `det()` implementiert.
- Suchen Sie aus der Tabelle in Abschnitt 6.1 den für diese Datenanalyse passenden Spezialfall heraus, berechnen Sie aus Λ die Teststatistik und speichern Sie das Ergebnis in die Variable `eff`.
- Berechnen Sie die zum Spezialfall gehörenden Freiheitsgradparameter und speichern Sie diese in die Variablen `df_1` und `df_2`.
- Ermitteln Sie den kritischen Wert k_{α_0} des Tests. Die inverse kumulative Verteilungsfunktion der F-Verteilung ist in R als `qf()` implementiert. Informieren Sie sich über die Eingabeparameter dieser Funktion, um den kritischen Wert korrekt zu bestimmen.
- Ermitteln Sie den p-Wert für den Test. Die kumulative Verteilungsfunktion der F-Verteilung ist in R als `pf()` implementiert. Informieren Sie sich über die Eingabeparameter dieser Funktion, um den p-Wert korrekt zu bestimmen.
- Ermitteln Sie das Testergebnis. Es ist 1, wenn die Teststatistik den kritischen Wert überschrittet. Andernfalls ist es 0.
- Geben Sie die Resultate ihrer Analyse aus. Sie sollten folgende Ergebnisse erhalten:

```
Wilks' Lambda          :  0.596186
Teststatistik          :  18.66617
Freiheitsgradparameter :  8 506
Signifikanzniveau      :  0.05
kritischer Wert        :  1.956693
Testergebnis           :  1
p-Wert                 :  0
```

- Überprüfen Sie Ihre multivariate Varianzanalyse, indem Sie folgenden Code mit der in R eingebauten Funktion `Manova()` ausführen:

```r
# Einfaktorielle multivariate Varianzanalyse (automatisch)
library(car)                                          # R-Paket "car"
group  = as.factor(x)                                 # faktorielle Variable
model  = lm(t(Y) ~ group)                             # Modellformulierung
result = Manova(model, test.statistic = "Wilks")      # Modellevaluation
print(result)                                         # Ergebnisausgabe
```

6.6 Lückentext

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

Lückentext: Die Anzahl der beobachteten experimentellen Einheiten ist _____, die Anzahl der abhängigen Variablen pro Einheit ist _____ und die Anzahl der Level der unabhängigen Variable ist _____. Die Wilks'-Lambda-Statistik hat den Wert _____, die daraus sich ergebende Teststatistik ist _____. Der kritische Wert für die multivariate ANOVA ergibt sich aus _____ und beträgt _____. Die Nullhypothese identischer Gruppenerwartungswerte wird daher bei einem Signifikanznivea von 5 Prozent _____ (abgelehnt/nicht abgelehnt). Der p-Wert für die multivariate ANOVA ergibt sich aus _____ und beträgt _____.

6.7 Mögliche Klausurfrage

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

Frage: Welche Aussage über die Parameterschätzer $\hat{\mu}_i$ und $\hat{\Sigma}$ im Modell der einfaktoriellen multivariaten Varianzanalyse ist **nicht** korrekt?

a) $\hat{\mu}_i$ ist das multivariate Stichprobenmittel der i-ten Gruppe.

b) $\hat{\mu}_i$ entspricht dem Gesamtstichprobenmittel $\bar{y}$.

c) $\hat{\Sigma}$ entspricht einer skalierten Form der der within-group sum-of-squares Matrix W.

d) $\hat{\Sigma}$ ist ein unverzerrter Schätzer des Kovarianzmatrixparameters Σ.

6.8 Kinderwitz

Gestern haben wir zwei Biber beim Essen beobachtet:

Gab Steg.