Prof. Dr. Fred Böker 12.04.2012

Klausur zur Vorlesung Analyse mehrdimensionaler Daten, WS 2011/2012 6 Kreditpunkte, 90 min

Hinweis:

- **Bitte runden Sie alle Ergebnisse auf drei Dezimalstellen.**
- **Runden Sie jedoch nur die Endergebnisse und keine Zwischenergebnisse.**
- **Wenn Sie bereits abgefragte Ergebnisse in folgenden Berechnungen benötigen, verwenden Sie jedoch bitte die gerundeten Ergebnisse.**
- **Im Anhang finden Sie Tabellen der benötigten Verteilungen.**

Aufgabe 1 (Punkte: 11)
Gegeben seien je 5 Beobachtungen von vier Variablen, die in **R** in der Matrix `DaTab` gespeichert[1] sind.

```
> DaTab
     [,1] [,2] [,3] [,4]
[1,]   32   64   65   67
[2,]   61   37   62   65
[3,]   59   40   45   43
[4,]   36   62   34   35
[5,]   62   46   43   40
```

a) **(Punkte: 1)**
Mit den folgenden **R**-Befehlen wurde die Kovarianzmatrix und die Korrelationsmatrix berechnet:

```
> COV<-round(cov(DaTab),digits=2)
> COV
        [,1]    [,2]   [,3]   [,4]
[1,]  216.50 -175.25 -10.75 -28.50
[2,] -175.25  156.20 -15.05  -6.75
[3,]  -10.75  -15.05 174.70 195.00
[4,]  -28.50   -6.75 195.00 222.00
> KORR<-round(cor(DaTab),digits=2)
> KORR
      [,1]  [,2]  [,3]  [,4]
[1,]  1.00 -0.95 -0.06 -0.13
[2,] -0.95  1.00 -0.09 -0.04
[3,] -0.06 -0.09  1.00  ?.??
[4,] -0.13 -0.04  ?.??  1.00
```

[1] Der hypothetische Datensatz ist aus Tabachnik und Fidell (2001), Using Multivariate Statistics, 4. Auflage, Allyn und Bacon, S. 592.

Berechnen Sie die in der Korrelationsmatrix durch `?.??` ersetzten Werte.

Lösung: Gesucht ist $r_{34} = r_{43} = \dfrac{\sigma_{34}}{\sigma_3 \cdot \sigma_4} = \dfrac{195}{\sqrt{174.70 \cdot 222}} = 0.9901743 \approx 0.99$

b) **(Punkte: 1)**
Mit **R** wurden die Eigenwerte der Korrelationsmatrix berechnet:

```
> round(eigen(KORR)$values,digits=3)
[1] 2.017 1.938 x.xxx y.yyy
```

Wie groß ist die Summe der beiden durch `x.xxx` bzw. `y.yyy` dargestellten Eigenwerte.

Lösung: Da es sich um eine Korrelationsmatrix handelt, ist die Summe aller Eigenwerte 4. Die Summe der beiden ersten Eigenwerte ist $2.017 + 1.938 = 3.955$. Also ist die Summe der beiden letzten Eigenwerte $4 - 3.955 = 0.045$.

c) **(Punkte: 1)**
Wir berechnen mit **R** die Matrix der Eigenvektoren:

```
> AKORR<-round(eigen(KORR)$vectors,digits=3)
> AKORR
       [,1]   [,2]   [,3]   [,4]
[1,]  0.348 -0.617  0.666  0.233
[2,] -0.243  0.667  0.678  0.190
[3,] -0.631 -0.314  0.267 -0.657
[4,] -0.649 -0.275 -0.157  0.692
```

Jetzt multiplizieren wir die Korrelationsmatrix mit dem dritten Eigenvektor, d.h. mit `AKORR[,3]`:

```
> AKORR[,3]
[1]  0.666  0.678  0.267 -0.157
> KORR%*%AKORR[,3]
         [,1]
[1,]  0.02629
[2,]  0.02755
[3,]  0.01059
[4,] -0.00637
```

Bestimmen Sie aus diesen Berechnungen den dritten Eigenwert. (**Hinweis:** Beachten Sie, dass es Ungenauigkeiten geben kann, weil die obigen **R**-Ergebnisse gerundet wurden.)

Wie groß ist dann der vierte Eigenwert?

Lösung: Wenn P die Korrelationsmatrix und $\boldsymbol{a}$ ein Eigenvektor, so gilt für den zugehörigen Eigenwert λ die Gleichung $P\boldsymbol{a} = \lambda\boldsymbol{a}$, d.h. hier ist

$$\boldsymbol{a} = \begin{pmatrix} 0.666 \\ 0.678 \\ 0.276 \\ -0.157 \end{pmatrix} \quad \text{und} \quad P\boldsymbol{a} = \lambda\boldsymbol{a} = \begin{pmatrix} 0.02629 \\ 0.02755 \\ 0.01059 \\ -0.00637 \end{pmatrix}$$

Damit ist $\lambda = 0.02629/0.666 = 0.03947447 \approx 0.039$. Verwenden wir statt der ersten Komponente eine der anderen drei Komponenten ergeben sich leicht unterschiedliche Werte, wie die folgende **R**-Berechnung zeigt:

```
> KORR%*%AKORR[,3]/AKORR[,3]
           [,1]
[1,] 0.03947447
[2,] 0.04063422
[3,] 0.03966292
[4,] 0.04057325
> round(KORR%*%AKORR[,3]/AKORR[,3],digits=3)
      [,1]
[1,] 0.039
[2,] 0.041
[3,] 0.040
[4,] 0.041
```

Wenn $\lambda_3 = 0.039$, so ist $\lambda_4 = 0.045 - 0.039 = 0.006$. (Bachten Sie, dass es leicht andere Ergebnisse gibt, wenn Sie mit den anderen Werten für λ_3 rechnen.)

Gehen Sie in den folgenden Teilaufgaben davon aus, dass Sie eine Hauptkomponentenanalyse mit der Korrelationsmatrix durchführen.

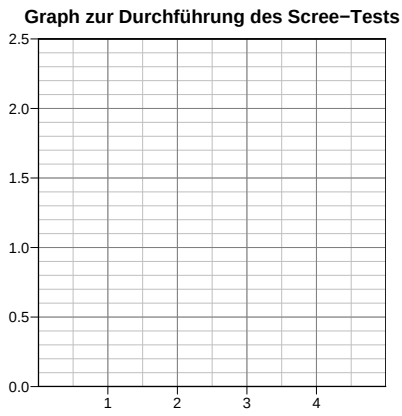
d) **(Punkte: 1)**
Wieviel Prozent der Variation werden durch die einzelnen Hauptkomponenten erklärt?

Lösung: Wir rechnen mit $\lambda_3 = 0.039$ und $\lambda_4 = 0.006$. Die Anteile sind dann
$100 \cdot 2.017/4\% = 50.425\%$
$100 \cdot 1.938/4\% = 48.450\%$
$100 \cdot 0.039/4\% = \ \ 0.975\%$
$100 \cdot 0.006/4\% = \ \ 0.150\%$

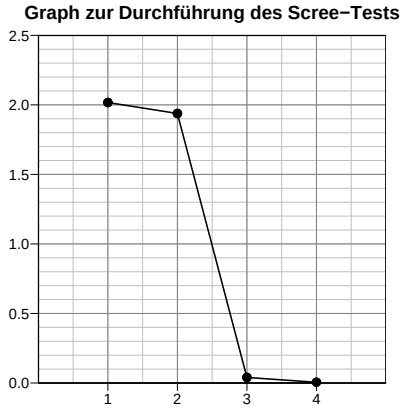
e) **(Punkte: 1)**
Wie viele Hauptkomponenten würden Sie verwenden, wenn mindestens 90% bzw. mindestens 99% der Variation durch die Hauptkomponenten erklärt werden sollen?

Lösung: Man müsste zwei bzw. drei Hauptkomponenten verwenden, denn die erste erklärt 50.425%, die beiden ersten zusammen 98.875%, die drei ersten 99.850% der Variation.

f) **(Punkte: 1)**
Führen Sie einen Scree-Test durch. Verwenden Sie dabei die folgende Abbildung. Erklären Sie, welche Werte Sie abtragen und welche Schlüsse Sie aus diesem Test ziehen.



Lösung:



Auf der x-Achse werden die Nummern der Eigenwerte $1, 2, 3, 4$ abgetragen, auf der y-Achse der Eigenwert. Hier ist ein deutlicher Knick nach dem 2. Eigenwert, d.h. man würde zwei Hauptkomponenten verwenden.

g) **(Punkte: 1)**
Verwenden Sie ein weiteres Kriterium zur Bestimmung der Anzahl zu verwendender Hauptkomponenten. (Gemeint ist **nicht** der Bartlett-Test.)

Lösung: Die ersten beiden Eigenwerte sind > 1, die weiteren sind < 1, d.h. man sollte die ersten beiden Hauptkomponenten verwenden.

h) **(Punkte: 2)**
In der folgenden **R**-Ausgabe wird ein Bartlett-Test durchgeführt. Wie viele Hauptkomponenten würden Sie bei einem Signifikanzniveau von $\alpha = 0.05$ verwenden?

Mit welchem **R**-Befehl kann man den zweiten der beiden P-Werte, der in der Ausgabe mit `0.1379` angegeben ist, berechnen?

```
> Bartlett.fun(DaTab)
$m1
[1] 1
$B
[1] 27.2159
$PWert
[1] 1e-04
$FG
[1] 5

$m1
[1] 2
$B
[1] 3.9621
$PWert
[1] 0.1379
$FG
[1] 2
```

Lösung: Für $m_1 = 2$ wird die Nullhypothese erstmals nicht verworfen, also sollte man zwei Hauptkomponenten verwenden.

Die Prüfgröße ist χ^2-verteilt mit $(m-m_1+2)(m-m_1-1)/2 = (4-2+2)(4-2-1)/2 = 4 \cdot 1/2 = 2$ Freiheitsgraden, d.h. der P-Wert kann mit `1 - pchisq(3.9621,2)` berechnet werden. In der Tat ergibt sich

```
> 1-pchisq(3.9621,2)
[1] 0.1379243
```

i) **(Punkte: 1)**
Berechnen Sie den Wert der ersten Hauptkomponente für den ersten Merkmalsträger.

Lösung: Die erste Hauptkomponente berechnet sich als

$Z_1 = 0.348Y_1 - 0.243Y_2 - 0.631Y_3 - 0.649Y_4 = 0.348 \cdot 32 - 0.243 \cdot 64 - 0.631 \cdot 65 - 0.649 \cdot 67 = -88.914$

j) **(Punkte: 1)**
Welche Informationen enthält die folgende **R**-Ausgabe?

Was bedeutet z.B. die Zahl -0.86 in der 1. Zeile und 2. Spalte?

```
> LAMBDAKORR<-diag(round(eigen(KORR)$values,digits=3))
> round(AKORR%*%sqrt(LAMBDAKORR),digits=2)
      [,1]  [,2]  [,3]  [,4]
[1,]  0.49 -0.86  0.13  0.02
[2,] -0.35  0.93  0.14  0.01
[3,] -0.90 -0.44  0.05 -0.05
[4,] -0.92 -0.38 -0.03  0.05
```

Lösung: Gegeben ist die Matrix der Komponentenladungen. Sie enthät die Korrelationen der j-ten Hauptkomponenten (Spalte) mit der i-ten standardisierten Originalvariablen (Zeile).

Die Korrelation zwischen ersten standardisierten Originalvariablen und der zweiten Hauptkomponente ist -0.86.

Aufgabe 2 (Punkte: 13)
Der Datensatz `DAFA` enthält $n = 100$ Beobachtungen von $m = 5$ Variablen. Mit dem **R**-Programm `factanal` wurde eine Faktorenanlyse durchgeführt:

```
> factanal(DAFA,factors=2)

Uniquenesses:
[1] 0.005 0.164 0.191 0.010 0.150

Loadings:
     Factor1 Factor2
[1,]          0.998
[2,]         -0.912
[3,]  0.899
```

```
[4,]  0.995
[5,]  0.918

                Factor1 Factor2
SS loadings       2.646   1.834
Proportion Var    0.529   0.367
Cumulative Var    0.529   0.896

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 65.86 on 1 degree of freedom.
The p-value is 4.84e-16

> LAMBDA<-round(factanal(DAFA,factors=2)$loadings[,1:2],digits=2)
>LAMBDA
     Factor1 Factor2
[1,]    0.00    1.00
[2,]   -0.07   -0.91
[3,]    0.90   -0.03
[4,]    0.99   -0.01
[5,]    0.92    0.08

> LAMBDA%*%t(LAMBDA)

       [,1]    [,2]    [,3]    [,4]    [,5]
  [1,]  1.00 -0.9100 -0.0300 -0.0100  0.0800
  [2,] -0.91  0.8330 -0.0357 -0.0602 -0.1372
  [3,] -0.03 -0.0357  0.8109  0.8913  0.8256
  [4,] -0.01 -0.0602  0.8913  0.9802  0.9100
  [5,]  0.08 -0.1372  0.8256  0.9100  0.8528
```

a) **(Punkte: 2)**
Erklären Sie die Ausgabe:

```
Uniquenesses:
[1] 0.005 0.164 0.191 0.010 0.150
```

Was genau bedeuten diese Zahlen im Zusammenhang mit dem Modell der Faktorenanalyse? Geben Sie dazu das hier verwendete Modell an.

Lösung: Hier wird ein Modell mit zwei Faktoren verwendet, d.h.

$$Y_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + e_j \qquad j = 1, 2, \ldots, 5$$

Dabei sind e_j die spezifischen Faktoren und die `Uniquenesses` sind die Varianzen der spezifischen Faktoren, d.h.

$$\text{Var}(e_1) = 0.005;\ \text{Var}(e_2) = 0.164;\ \text{Var}(e_3) = 0.191;\ \text{Var}(e_4) = 0.010;\ \text{Var}(e_5) = 0.150$$

b) **(Punkte: 2)**
Bestimmen Sie allein mit der Ausgabe in a) die Kommunalitäten. Gehen Sie davon aus, dass die Originalvariablen standardisiert sind, d.h. alle Varianzen sind 1.

Welche Bedeutung haben die Kommunalitäten?

Lösung: Die Summe aus den Kommunalitäten und den spezifischen Varianzen ist 1. Daher sind die Kommunlitäten

$$1-0.005 = 0.995;\ 1-0.164 = 0.836;\ 1-0.191 = 0.809;\ 1-0.010 = 0.990;\ 1-0.150 = 0.850$$

Die Kommunalität ist die Varianz, die durch die gemeinsamen Faktoren erklärt wird.

c) **(Punkte: 3)**
Erklären Sie die Ausgabe:

```
Loadings:
     Factor1 Factor2
[1,]           0.998
[2,]          -0.912
[3,]  0.899
[4,]  0.995
[5,]  0.918
```

Was bedeuten die Zahlen in dieser Ausgabe?
Warum sind einige Felder leer?
Mit welchen Variablen korreliert der erste Faktor, mit welchen der zweite Faktor hoch?
Ist die Originalvariable Y_2 eher klein oder eher groß, wenn der zweite Faktor groß ist?

Lösung: Die Ausgabe enthält die Faktorladungen $\lambda_{jk};\ j = 1, 2, \ldots 5; k = 1, 2$, d.h. die Ladungen der j-ten Variablen auf den k-ten Faktor. Die Ladungen sind gleich den Korrelationskoeffizienten zwischen der j-ten Variablen und dem k-ten Faktor.

Es werden nur die bedeutenden Faktorladungen angegeben, kleine Ladungen werden weggelassen.

Der erste Faktor korreliert hoch mit der dritten, vierten und fünften Variablen. Der zweite Faktor korreliert stark positiv mit der ersten Variablen und stark negativ mit der zweiten Variablen.

Wenn der zweite Faktor groß ist, ist die zweite Variable eher klein, da sie stark negativ korreliert sind.

d) **(Punkte: 2)**
Gehen Sie davon aus, dass die Daten einer gemeinsamen standardisierten Normalverteilung entstammen mit der Kovarianzmatrix:

$$\Sigma = \begin{pmatrix} 1.0 & -0.9 & -0.1 & -0.1 & 0.0 \\ -0.9 & 1.0 & -0.1 & 0.0 & 0.0 \\ -0.1 & -0.1 & 1.0 & 0.9 & 0.8 \\ -0.1 & 0.0 & 0.9 & 1.0 & 0.9 \\ 0.0 & 0.0 & 0.8 & 0.9 & 1.0 \end{pmatrix}$$

Schreiben Sie Σ unter Verwendung der Notation der Vorlesung in der Gestalt

$$\Sigma = \Lambda\Lambda^t + \Psi$$

Dabei sollen zwei Faktoren verwendet werden. Benutzen Sie dazu die Ausgabe:

```
>LAMBDA
     Factor1 Factor2
[1,]    0.00    1.00
[2,]   -0.07   -0.91
[3,]    0.90   -0.03
[4,]    0.99   -0.01
[5,]    0.92    0.08
```

Hinweis: Verwenden Sie auch die weiteren Berechnungen, die oben, d.h. vor den einzelnen Teilaufgaben mit LAMBDA durchgeführt wurden.

Lösung: In den Berechnungen ist $\Lambda\Lambda^t$ gegeben:

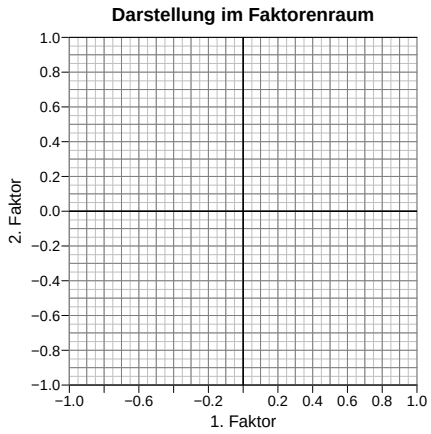
```
> LAMBDA%*%t(LAMBDA)

      [,1]    [,2]    [,3]    [,4]    [,5]
[1,]  1.00 -0.9100 -0.0300 -0.0100  0.0800
[2,] -0.91  0.8330 -0.0357 -0.0602 -0.1372
[3,] -0.03 -0.0357  0.8109  0.8913  0.8256
[4,] -0.01 -0.0602  0.8913  0.9802  0.9100
[5,]  0.08 -0.1372  0.8256  0.9100  0.8528
```

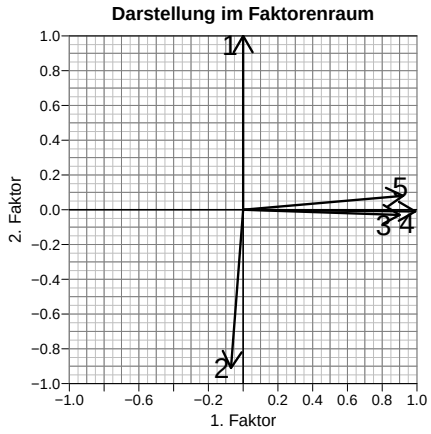
Damit kann Ψ als Differenz $\Psi = \Sigma - \Lambda\Lambda^t$ berechnent werden:

$$\Psi = \begin{pmatrix} 0.0000 & 0.0100 & -0.0700 & -0.0900 & -0.0800 \\ 0.0100 & 0.1670 & -0.0643 & 0.0602 & 0.1372 \\ -0.0700 & -0.0643 & 0.1891 & 0.0087 & -0.0256 \\ -0.0900 & 0.0602 & 0.0087 & 0.0198 & -0.0100 \\ -0.0800 & 0.1372 & -0.0256 & -0.0100 & 0.1472 \end{pmatrix}$$

e) **(Punkte: 2)**
Stellen Sie die Variablen unter Verwendung der folgenden Abbildung im Faktorenraum dar.



Lösung:



f) **(Punkte: 2)**
Was bedeuten die Zahlen in der folgenden Ausgabe?

```
                Factor1 Factor2
SS loadings       2.646   1.834
Proportion Var    0.529   0.367
Cumulative Var    0.529   0.896
```

Lösung: In der Zeile `SS loadings` stehen die durch den oben angegebenen Faktor erklärte Varianzen, d.h. 2.646 durch den ersten und 1.834 durch den zweiten Faktor.

In der Zeile `Proportion Var` steht der Anteil der durch diesen Faktor erklärten Varianz, d.h. der erste Faktor erklärt 52.9% der Varianz, der zweite 36.7%. Die Zahlen berechnen sich aus : $2.646/5$ bzw. $1.834/5$. In der Zeile `Cumulative Var` steht der kumulierte Anteil der erklärten Varianz, d.h. der erste Faktor erklärt 52.9% der Varianz, die beiden ersten erklären zusammen 89.6%.

Aufgabe 3 (Punkte: 12)
Der Datensatz `DAMVN` enthält $n = 124$ Beobachtungen von $m = 4$ Variablen aus einer multivariaten Normalverteilung. Es soll die Nullhypothese

$$H_0 : \boldsymbol{\mu}^t = (\mu_1, \mu_2, \mu_3, \mu_4) = (5, 3, 1, 1)$$

getetstet werden. Es wurden die folgenden Berechnungen mit **R** durchgeführt:

```
> DAMVN  # Datenmatrix, die drei ersten und drei letzten Zeilen
       [,1] [,2]  [,3]  [,4]
  [1,] 4.76 3.11  2.21  2.50
  [2,] 5.84 2.23  0.99  0.81
  [3,] 5.33 3.32  0.62  0.71


[122,] 5.07 3.65  0.44  1.56
[123,] 7.56 0.36  0.22  0.60
[124,] 3.94 4.41  1.38  1.18
> round(apply(DAMVN,2,mean),2)  # Mittelwerte der Spalten
```

```
[1] 5.06 3.13 0.91 1.08
> muNULL<-c(5,3,1,1) # Erwartungswerte unter Nullhypothese
> muNULL
[1] 5 3 1 1
> round(apply(DAMVN,2,mean),2)-muNULL
[1]  0.06  0.13 -0.09  0.08
> round(var(DAMVN),2) # Geschätzte Kovarianzmatrix
      [,1]  [,2]  [,3]  [,4]
[1,]  1.03 -0.91 -0.09 -0.08
[2,] -0.91  1.01 -0.10 -0.01
[3,] -0.09 -0.10  0.99  0.89
[4,] -0.08 -0.01  0.89  1.00
> SINVDAMVN<-round(solve(var(DAMVN)),2) # Inverse
> SINVDAMVN
      [,1]  [,2]  [,3]  [,4]
[1,]  6.99  6.64  3.71 -2.66
[2,]  6.64  7.34  3.99 -2.93
[3,]  3.71  3.99  7.47 -6.31
[4,] -2.66 -2.93 -6.31  6.36
> round(SINVDAMVN%*%round(apply(DAMVN,2,mean),2)-muNULL,2)
       [,1]
[1,]  51.66
[2,]  54.04
[3,]  30.24
[4,] -22.50
```

a) **(Punkte: 1)**
Welche Bedeutung hat der mit

```
> round(SINVDAMVN%*%round(apply(DAMVN,2,mean),2)-muNULL,2)
       [,1]
[1,]  51.66
[2,]  54.04
[3,]  30.24
[4,] -22.50
```

berechnete Vektor für die Berechnung der Prüfgröße

$$\mathcal{T}^2 = \max_{\boldsymbol{a}} \frac{n[\boldsymbol{a}^t(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)]^2}{\boldsymbol{a}^t S \boldsymbol{a}} \quad ?$$

Wie wurde dieser Vektor in der Vorlesung bezeichnet?

Geben Sie eine andere Formel an für die Berechnung der Prüfgröße $\mathcal{T}^2$ mit Hilfe des obigen Vektors.

Lösung: Es ist der Vektor $\boldsymbol{a}^*$, der die Prüfgröße $\mathcal{T}^2$ maximiert. Es gilt $\mathcal{T}^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\boldsymbol{a}^*$.

b) **(Punkte: 1)**
Berechnen Sie die Prüfgröße $\mathcal{T}^2$ unter Verwendung der obigen Berechnungen mit **R**.

Lösung: $\mathcal{T}^2 = 124(0.06, 0.13, -0.09, 0.08) \cdot (51.66, 54.04, 30.24, -22.50)$, wobei mit $\cdot$ das Skalarprodukt bezeichnet wird. Es folgt
$\mathcal{T}^2 = 124 \cdot (0.06 \cdot 51.66 + 0.13 \cdot 54.04 - 0.09 \cdot 30.24 - 0.08 \cdot 22.50) = 124 \cdot 5.6032 = 694.7968$.

c) **(Punkte: 3)**
Rechnen Sie $\mathcal{T}^2$ um in eine F-verteilte Prüfgröße $\mathcal{F}$. Welche Freiheitsgrade hat $\mathcal{F}$?

Bestimmen Sie den Ablehnungsbereich für $\mathcal{F}$, wenn $\alpha = 0.05$.

Würden Sie die Nullhypothese ablehnen?

Lösung: $\mathcal{F} = \frac{(n-m)\mathcal{T}^2}{m(n-1)} = \frac{(124-4)\cdot 694.7968}{4\cdot(124-1)} = \frac{120\cdot 694.7968}{492} = 169.4626 \approx 169.463$

$\mathcal{F}$ hat $m = 4$ und $n - m = 124 - 4 = 120$ Freiheitsgrade.

Der Ablehnungsbereich ist $A = (2.45, \infty)$.

Da $169.463 \in A$, ist die Nullhypothese zu verwerfen.

d) **(Punkte: 3)**
Es sollen simultane und „*gewöhnliche Konfidenzintervalle*" für die Erwartungswerte $\mu_i;\ i = 1, 2, \ldots, 4$ berechnet werden. Rechnen Sie beide Konfidenzintervalle für $i = 4$ aus, wenn $1 - \alpha = 0.90$.

Hinweis: Vernachlässigen Sie das Problem, dass Sie die das für die „*gewöhnlichen Konfidenzintervalle*" benötigte Quantil nicht explizit aus der Tabelle ablesen können.

Lösung: Die simultanen Konfidenzintervalle (siehe Skript, S. 101) haben die Gestalt

$$\mu_i \in \bar{x}_i \pm K_{0.05}\sqrt{\frac{s_i^2}{n}} \qquad i = 1, 2, \ldots, 4$$

Dabei sind $\bar{x}_i$ die Mittelwerte $(5.06, 3.13, 0.91, 1.08)$ und s_i^2 die Varianzen $(1.03, 1.01, 0.99, 1.00)$ in den Gruppen $i = 1, 2, 3, 4$. Die Varianzen sind die Diagonalemente in der Varianz-Kovarianzmatrix. Es gilt

$K_{0.05} = \left(\frac{m(n-1)}{n-m}F_{0.1}(m, n-m)\right)^{1/2} = \left(\frac{4\cdot 123}{120}\cdot F_{0.1}(4, 120)\right)^{1/2} = \left(\frac{4\cdot 123}{120}\cdot 1.99\right)^{1/2} = \sqrt{8.159} = 2.856396 \approx 2.856.$

Das Intervall für μ_4 ist $\left(1.08 - 2.856396\sqrt{\frac{1.00}{124}}, 1.08 - 2.856396\sqrt{\frac{1.00}{124}}\right) = (1.08 - 0.2565120, 1.08 + 0.2565120) = (0.823488, 1.336512) \approx (0.823, 1.337)$

Die „*gewöhnlichen*" Konfidenzintervalle haben die Gestalt

$$\mu_i \in \bar{x}_i \pm t_{123,0.05}\sqrt{\frac{s_i^2}{n}} \qquad i = 1, 2, \ldots, 4$$

Dabei ist $t_{123,0.05} \approx 1.66$.

Das Intervall für μ_4 ist $\left(1.08 - 1.66\sqrt{\frac{1.00}{124}}, 1.08 - 1.66\sqrt{\frac{1.00}{124}}\right) = (1.08 - 0.1490724, 1.08 + 0.1490724) = (0.9309276, 1.229072) \approx (0.931, 1.229)$

e) **(Punkte: 2)**
Die obige Nullhypothese $(\mu_1, \mu_2, \mu_3, \mu_4) = (5, 3, 1, 1)$ ist ein Spezialfall der allgemeineren Nullhypothese

$$H_0: \qquad \mu_1 = \mu_2 + 2 \qquad \mu_2 = \mu_3 + 2 \qquad \mu_3 = \mu_4$$

Geben Sie eine geeignete Matrix $\boldsymbol{C}$ und einen geeigneten Vektor $\boldsymbol{\phi}$ an, mit denen man die Nullhypothese formulieren kann. Geben Sie dann die Formel zur Berechnung der Prüfgröße an.

Lösung: Die Nullhypothese ist äquivalent zu den Gleichungen

$$\begin{aligned} \mu_1 - \mu_2 &= 2 \\ \mu_2 - \mu_3 &= 2 \\ \mu_3 - \mu_4 &= 0 \end{aligned}$$

Dies kann geschrieben werden als

$$\boldsymbol{C}^t\boldsymbol{\mu} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix} = \boldsymbol{\phi}$$

Die Prüfgröße ist

$$\mathcal{T}^2 = n(C^t\bar{\boldsymbol{X}} - \boldsymbol{\phi})^t(C^tSC)^{-1}(C^t\bar{\boldsymbol{X}} - \boldsymbol{\phi})$$

f) **(Punkte: 2)**
Rechnet man die Prüfgröße aus, so ergibt sich der Wert $\mathcal{T}^2 \approx 18.3$. Geben Sie den **R**-Befehl an, mit dem man den P-Wert mit Hilfe der F-Verteilung berechnen kann.

Lösung: Es ist $\frac{n-p}{p(n-p)}\mathcal{T}^2 \sim F(p, n-p)$ und hier ist $n = 124$ und $p = 3$. Damit ist $F = \frac{n-p}{p(n-1)} \cdot \mathcal{T}^2 = \frac{121}{3 \cdot 123} \cdot 18.3 = 6.000813$.
Der P-Wert kann also mit `1-pf(6.000813,3,121)` berechnet werden.

Aufgabe 4 (Punkte: 2)
Die Beobachtung $\boldsymbol{x}_0$ soll einer von drei Populationen $\mathcal{P}_1, \mathcal{P}_2$ oder $\mathcal{P}_3$ zugeordnet werden. Die Kosten einer Fehlzuordnung seien durch die Matrix

$$\boldsymbol{C} = (C_{ij} = C(i|j)) = \begin{pmatrix} 0 & 100 & 500 \\ 10 & 0 & 200 \\ 20 & 50 & 0 \end{pmatrix}$$

gegeben. Die Apriori-Wahrscheinlichkeiten seien:

$$\pi_1 = 0.1 \qquad \pi_2 = 0.3 \qquad \pi_3 = 0.6$$

Die Werte der gemeinsamen Dichtefunktion an der Stelle $\boldsymbol{x}_0$ seien:

$$f_1(\boldsymbol{x}_0) = 0.1 \qquad f_2(\boldsymbol{x}_0) = 0.8 \qquad f_3(\boldsymbol{x}_0) = 1.2$$

Berechnen Sie die in der Vorlesung definierten Diskriminanzwerte w_i.

Beachten Sie, dass keine Normalverteilung vorausgesetzt wurde!

Welcher Population würden Sie die Beobachtung $\boldsymbol{x}_0$ zuordnen?

Lösung: Die Diskriminanzwerte sind $w_i = -\sum_{j=1}^{3} \pi_j C(i|j) f_j(\boldsymbol{x}_0)$. Hier ist

$w_1 = -\pi_2 C(1|2) f_2(\boldsymbol{x}_0) - \pi_3 C(1|3) f_3(\boldsymbol{x}_0) = -0.3 \cdot 100 \cdot 0.8 - 0.6 \cdot 500 \cdot 1.2 = -24 - 360 = -384$

$w_2 = -\pi_1 C(2|1) f_1(\boldsymbol{x}_0) - \pi_3 C(2|3) f_3(\boldsymbol{x}_0) = -0.1 \cdot 10 \cdot 0.1 - 0.6 \cdot 200 \cdot 1.2 = -0.1 - 144 = -144.1$

$w_3 = -\pi_1 C(3|1) f_1(\boldsymbol{x}_0) - \pi_2 C(3|2) f_2(\boldsymbol{x}_0) = -0.1 \cdot 20 \cdot 0.1 - 0.3 \cdot 50 \cdot 0.8 = -0.2 - 12 = -12.2$

Damit ist w_3 am größten. Man würde die Beobachtung der Population $\mathcal{P}_3$ zuordnen.

Aufgabe 5 (Punkte: 2)
Nehmen Sie an: Sie führen eine Diskriminanzanalyse mit zwei Populationen $\mathcal{P}_1$ und $\mathcal{P}_2$ durch. Es gelte

$$\pi_1 = \pi_2 \qquad \text{und} \qquad C(1|2) = C(2|1)$$

Sie haben aus einer Lernstichprobe die Werte

$$(\boldsymbol{a}^*)^t = (0.4, 0.8, 0.1) \qquad \bar{v}_1 = \boldsymbol{L}^t \bar{\boldsymbol{x}}_1 = 122 \qquad \bar{v}_2 = \boldsymbol{L}^t \bar{\boldsymbol{x}}_2 = 88$$

erhalten. Es liegt eine neue Beobachtung $\boldsymbol{x}^t = (180, 40, 70)$ vor.

Berechnen Sie $\boldsymbol{L}^t \boldsymbol{x}$ und ordnen Sie diese Beobachtung einer der beiden Populationen zu. Verwenden Sie dazu den in der Vorlesung mit $\hat{m}$ bezeichneten Wert.

Lösung: Nach (8.13) ist $\hat{m} = \frac{1}{2}(\bar{v}_1 + \bar{v}_2) = \frac{1}{2}(122 + 88) = \frac{210}{2} = 105$.

Für die neue Beobachtung gilt $\boldsymbol{L}^t \boldsymbol{x} = (\boldsymbol{a}^*)^t \boldsymbol{x} = 0.4 \cdot 180 + 0.8 \cdot 40 + 0.1 \cdot 70 = 72 + 32 + 7 = 111 > 105 = \hat{m}$, d.h. die Beobachtung wird der Population $\mathcal{P}_1$ zugeordnet.

Ende der Klausur