Prof. Dr. Fred Böker 28.02.2013

Klausur zur Vorlesung Analyse mehrdimensionaler Daten, WS 2012/2013 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.**

Gesamtpunkte: 44

Aufgabe 1 (Punkte: 7)
Gegeben sei die folgende Korrelationsmatrix der Zufallsvariablen $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^t$.

$$P = \begin{pmatrix} 1.00 & 0.90 & 0.80 & 0.05 \\ 0.90 & 1.00 & 0.80 & -0.02 \\ 0.80 & 0.80 & 1.00 & 0.15 \\ 0.05 & -0.02 & 0.15 & 1.00 \end{pmatrix}$$

a) **(Punkte: 2)**
Berechnen Sie die Kovarianz $\sigma_{23} = \text{Cov}(Y_2, Y_3)$, wenn die Varianzen von Y_2 und Y_3 durch $\sigma_2^2 = 9$ und $\sigma_3^2 = 196$ gegeben sind.

Lösung: Es gilt $\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} \iff \sigma_{ij} = \rho_{ij} \cdot \sigma_i \cdot \sigma_j$, d.h. hier gilt $\sigma_{23} = \rho_{23} \cdot \sigma_2 \cdot \sigma_3 = 0.8 \cdot 3 \cdot 14 = 33.6.$

b) **(Punkte: 2)**
Nehmen Sie an, dass alle Standardabweichungen $\sigma_1, \ldots, \sigma_4$ bekannt sind. Mit welchen Matrizenoperationen können Sie dann aus der Korreleationsmatrix P die Kovarianzmatrix Σ berechnen? Geben Sie insbesondere die dazu benötigte Matrix an.

Lösung: Wir bilden die Diagonalmatrix D, die in der Diagonale die Standardabweichungen σ_i enthält:

$$D = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{pmatrix}$$

Dann gilt: $\Sigma = DPD$.

c) **(Punkte: 3)**

Nehmen Sie an, dass es sich bei der obigen Korrelationsmatrix um eine geschätzte Korrelationsmatrix handelt. Die Stichprobengröße sei $n = 102$. Prüfen Sie die Nullhypothese $\rho_{34} = 0$ gegen die Alternativhypothese $\rho_{34} \neq 0$.

Berechnen Sie den Wert der Prüfgröße und geben Sie die Verteilung der Prüfgröße unter der Nullhypothese an.

Berechnen Sie den P-Wert mit Hilfe einer geeigneten approximierenden Verteilung, deren Verteilungsfunktion im Anhang bei den Tabellen zu finden ist. Runden Sie dazu den Wert der Prüfgröße auf zwei Stellen nach dem Dezimalpunkt.

Lösung: Die Prüfgröße ist $PG = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.15 \cdot \sqrt{100}}{\sqrt{1-0.15^2}} = 1.517165 \approx 1.52$. Die Prüfgröße ist unter der Nullhypothese t-verteilt mit $n-2 = 100$ Freiheitsgraden. Bei großen Freiheitsgraden (≥ 30) kann die t-Verteilung durch eine $N(0,1)$-Verteilung approximiert werden.

Wenn $Z \sim N(0,1)$, so gilt $P(Z > 1.52) = 1 - 0.064$, so dass der P-Wert $2 \cdot 0.064 = 0.128$ ist.

Aufgabe 2 (Punkte: 10)
Die in Aufgabe 1 verwendete Korrelationsmatrix wurde in **R** unter dem Namne `Korrmat` gespeichert, während unter `IMat` die Einheitsmatix der Ordnung 4 gespeichert wurde.

```
> KorrMat
     [,1]  [,2] [,3]  [,4]
[1,] 1.00  0.90 0.80  0.05
[2,] 0.90  1.00 0.80 -0.02
[3,] 0.80  0.80 1.00  0.15
[4,] 0.05 -0.02 0.15  1.00
> IMat
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

a) **(Punkte: 2)**
In **R** wurden mit diesen Matrizen die folgenden Berechnungen durchgeführt:

```
> det(KorrMat-2.674*IMat)
[1] 0.002209530
> det(KorrMat-1.013*IMat)
[1] -0.000522674
> det(KorrMat-0.217*IMat)
[1] 7.485132e-05
> det(KorrMat-0.097*IMat)
[1] -1.116132e-05
```

Alle Werte sind nahezu Null. (Sie wären noch näher bei Null, hätten wir bei den verwendeten Zahlen statt drei noch weitere Stellen nach dem Dezimalpunkt verwendet.) Was folgt daraus für die Koeffizienten von `IMat`?

Lösung: Wenn man die Koeffizienten von `IMat` mit λ bezeichnet, so gilt ungefähr $\det(P-\lambda I) = 0$, d.h. die Eigenwerte von P sind ungefähr $\lambda_1 = 2.674;\ \lambda_2 = 1.013;\ \lambda_3 = 0.217;\ \lambda_4 = 0.097$.

b) **(Punkte: 2)**
Zeigen Sie dass der zweite Eigenvektor von P ungefähr $\boldsymbol{a} = (-0.05, -0.13, 0.08, 0.99)^t$ ist. Runden Sie dazu den zugehörigen Eigenwert auf eine ganze Zahl.

Hinweis: Beachten Sie, dass die Ergebnisse wegen der Rundungen nur **ungefähr** sein können.

Lösung: Der zweite Eigenwert ist nach a) ungefähr 1. Demnach ist zu zeigen, dass $(P - I)\boldsymbol{a} = 0$. Wir machen das hier mit **R**:

```
> EigenVek2<-round(eigen(KorrMat)$vectors[,2],2)
> EigenVek2
[1] -0.05 -0.13  0.08  0.99
> KorrMat-IMat
     [,1]  [,2] [,3]  [,4]
[1,] 0.00  0.90 0.80  0.05
[2,] 0.90  0.00 0.80 -0.02
[3,] 0.80  0.80 0.00  0.15
[4,] 0.05 -0.02 0.15  0.00
> (KorrMat-IMat)%*%EigenVek2
        [,1]
[1,] -0.0035
[2,] -0.0008
[3,]  0.0045
[4,]  0.0121
```

c) **(Punkte: 2)**
Interpretieren Sie das Ergebnis aus b). Schauen Sie sich dazu den zweiten Eigenwert, den zweiten Eigenvektor im Zusammenhang mit einer bestimmten Spalte der Korrelationsmatrix P an.

Lösung: Der zweite Eigenwert ist ungefähr 1, d.h. ungefähr gleich der Varianz der standardisierten Variablen. Die vierte Variable Y_4 ist nahezu unkorreliert mit den anderen Variablen. Der zweite Eigenvektor hat ungefähr den Wert 1 als vierten Wert und stimmt nahezu mit Y_4 überein (siehe Skript. S. 53).

d) **(Punkte: 2)**
Wieviel Prozent der Totalvariation erklärt die erste Hauptkomponente, wieviel die beiden ersten zusammen?

Lösung: Die erste Hauptkomponente erklärt $\frac{2.674}{4} \cdot 100\% = 66.85\%$, die ersten beiden $\frac{2.674 + 1.013}{4} \cdot 100\% = 92.175\%$. Beachten Sie: Die Summe der Eigenwerte ist 4.

e) **(Punkte: 2)**
Mit **R** wurden die folgenden Berechnungen durchgeführt:

```
> EigenVeks<-round(eigen(KorrMat)$vectors,2)
> EigenWerte<-round(eigen(KorrMat)$values,2)
> round(EigenVeks%*%sqrt(diag(EigenWerte)),2)
      [,1]  [,2]  [,3]  [,4]
[1,] -0.95 -0.05  0.20  0.22
[2,] -0.95 -0.13  0.16 -0.23
[3,] -0.92  0.08 -0.38  0.01
[4,] -0.10  0.99  0.07 -0.02
```

Welche Werte enthält die hier ausgegebene Matrix? Interpretieren Sie insbesondere die beiden ersten Spalten.

Lösung: Ausgegeben wird die Matrix der Komponentenladungen. Sie enthält die Korrelation der Hauptkomponenten (in den Spalten) mit den standardisierten Originalvariabeln (in den Zeilen). Die erste Hauptkomponente ist sehr stark negativ korreliert mit den drei ersten Variablen und sehr schwach korreliert mit der vierten Variablen. Die zweite Hauptkomponente ist sehr schwach korreliert mit den ersten drei Variablen und nahezu vollständig korreliert mit der vierten Variablen.

Aufgabe 3 (Punkte: 8)
Gegeben sei die folgende geschätzte Korrelationsmatrix der Zufallsvariablen $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^t$.

$$R = \begin{pmatrix} 1.00 & 0.90 & 0.80 \\ 0.90 & 1.00 & 0.80 \\ 0.80 & 0.80 & 1.00 \end{pmatrix}$$

Diese Matrix wurde in **R** unter dem Namen `korrmat` gespeichert. Anschließend wurden die folgenden Berechnungen durchgeführt:

```
> korrmat<-matrix(c(1,0.9,0.8,0.9,1,0.8,0.8,0.8,1),nrow=3)
> korrmat
     [,1] [,2] [,3]
[1,]  1.0  0.9  0.8
[2,]  0.9  1.0  0.8
[3,]  0.8  0.8  1.0
> eigenvek<-round(eigen(korrmat)$vectors,2)
> eigenvek
      [,1]  [,2]  [,3]
[1,] -0.59 -0.40  0.71
[2,] -0.59 -0.40 -0.71
[3,] -0.56  0.83  0.00
> eigenwerte<-round(eigen(korrmat)$values,2)
> eigenwerte
[1] 2.67 0.23 0.10
> round(eigenvek%*%sqrt(diag(eigenwerte)),3)
       [,1]   [,2]   [,3]
[1,] -0.964 -0.192  0.225
[2,] -0.964 -0.192 -0.225
[3,] -0.915  0.398  0.000
```

Es soll eine Faktorenanalyse mit zwei Faktoren nach der Hauptkomponentenmethode durchgeführt werden.

a) **(Punkte: 2)**
Schreiben Sie zunächst allgemein das Modell der Faktorenanalyse für die obige Situation auf und geben Sie alle Annahmen an.

Lösung: Siehe Kap. 5.2 im Skript.

b) **(Punkte: 1)**
Bestimmen Sie dann die Matrix $\Lambda_{(2)}$ für das Modell mit 2 Faktoren aus den obigen Berechnungen mit **R**.

Lösung: $\Lambda_{(2)}$ enthält die beiden ersten Spalten der obigen Matrix

```
> round(eigenvek%*%sqrt(diag(eigenwerte)),3)
       [,1]   [,2]   [,3]
[1,] -0.964 -0.192  0.225
[2,] -0.964 -0.192 -0.225
[3,] -0.915  0.398  0.000
```

In **R** erhalten wir $\Lambda_{(2)}$ mit:

```
> Lambda2<-round(eigenvek%*%sqrt(diag(eigenwerte)),3)[,1:2]
> Lambda2
       [,1]   [,2]
[1,] -0.964 -0.192
[2,] -0.964 -0.192
[3,] -0.915  0.398
```

c) **(Punkte: 3)**
Zerlegen Sie R in $R = \Lambda_{(2)}\Lambda_{(2)}^t + \Psi$. (Runden Sie auf drei Stellen nach dem Dezimalpunkt). Geben Sie die Kommunalitäten und die Einzelrestvarianzen an.

Lösung:

```
> Lambda2%*%t(Lambda2)
         [,1]     [,2]     [,3]
[1,] 0.966160 0.966160 0.805644
[2,] 0.966160 0.966160 0.805644
[3,] 0.805644 0.805644 0.995629
> round(Lambda2%*%t(Lambda2),3)
      [,1]  [,2]  [,3]
[1,] 0.966 0.966 0.806
[2,] 0.966 0.966 0.806
[3,] 0.806 0.806 0.996
```

Man erhält Ψ aus $R - \Lambda_{(2)}\Lambda_{(2)}^t$:

```
> korrmat- round(Lambda2%*%t(Lambda2),3)
       [,1]   [,2]   [,3]
[1,]  0.034 -0.066 -0.006
[2,] -0.066  0.034 -0.006
[3,] -0.006 -0.006  0.004
```

Die Kommunalitäten sind die Diagonalelemente in $\Lambda_{(2)}\Lambda^t_{(2)}$, d.h. die Kommunalitäten sind hier $0.966; 0.966$ und 0.996. Somit sind die Einzelrestvarianzen $0.34; 0.34$ und 0.04.

d) **(Punkte: 2)**
Die Ladungsmatrix $\Lambda_{(2)}$ wurde in **R** unter `Lambda2` gespeichert. Was bedeutet die folgende **R**-Ausgabe? Welche Schlüsse ziehen Sie aus diesem Ergebnis?

```
> round(varimax(Lambda2)$rotmat,2)
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

Lösung: Mit diesem Befehl wird der optimale Drehwinkel α bestimmt. Die Matrix enthält die Drehmatrix

$$T = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

Da $\cos(0) = 1$ und $\sin(0) = 0$ ist der optimale Drehwinkel 0. Es ist keine Drehung notwendig.

Aufgabe 4 (Punkte: 14)
Der Datensatz `DatMV` enthält $n = 24$ Beobachtungen von $m = 4$ Variablen aus einer multivariaten Normalverteilung.

```
> DatMV
      [,1] [,2]  [,3]  [,4]
 [1,] 7.06 0.70  1.26  0.48
 [2,] 6.20 1.57 -0.03 -0.42
 [3,] 6.47 1.75 -0.94 -1.05
 [4,] 5.46 2.58 -0.50 -0.67
 [5,] 7.42 0.41  0.55  0.74
 [6,] 5.81 2.55  0.95  0.66
 [7,] 6.17 2.26  1.13  1.41
 [8,] 6.31 1.87  0.97  0.69
 [9,] 4.16 2.24  2.44  2.46
[10,] 5.52 2.34  3.39  3.89
[11,] 4.72 3.16  1.70  1.12
[12,] 5.35 2.11  3.33  3.62
[13,] 5.35 2.39  0.78  0.85
[14,] 6.47 1.49  0.62  0.97
[15,] 4.90 3.74 -0.78  0.11
[16,] 4.01 4.06  1.75  2.04
[17,] 5.21 3.17 -0.34 -0.74
[18,] 6.07 2.03  2.20  2.40
[19,] 5.54 2.65  2.37  2.87
[20,] 6.16 1.45  1.39  1.11
[21,] 7.20 0.60  1.32  1.56
[22,] 4.92 3.47  2.08  2.68
[23,] 7.24 0.87  0.54  0.67
[24,] 4.64 2.55  1.54  1.39
```

Es soll die Nullhypothese

$$H_0 : \boldsymbol{\mu}^t = (\mu_1, \mu_2, \mu_3, \mu_4) = (6, 2, 1, 1)$$

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

```
> xquer<-round(apply(DatMV,2,mean),2)
> xquer
[1] 5.76 2.17 1.16 1.20
> MUnull<-c(6,2,1,1)
> MUnull
[1] 6 2 1 1
> xquer-MUnull
[1] -0.24  0.17  0.16  0.20
> SINVDatMV<-round(solve(var(DatMV)),2)
> SINVDatMV
      [,1]  [,2]  [,3]  [,4]
[1,]  7.04  6.27  3.77 -2.35
[2,]  6.27  6.74  4.11 -2.85
[3,]  3.77  4.11 10.83 -8.81
[4,] -2.35 -2.85 -8.81  7.85
> round(SINVDatMV%*%(xquer-MUnull),2)
      [,1]
[1,] -0.49
[2,] -0.27
[3,] -0.24
[4,]  0.24
```

a) **(Punkte: 3)**
Prüfen Sie die obige Nullhypothese mit der Prüfgröße $\mathcal{T}^2$ unter Verwendung der obigen Berechnungen mit **R**.

Geben Sie zwei Formeln für die Prüfgröße $\mathcal{T}^2$ an.

Welche Formel haben Sie verwendet? Welcher Teil dieser Formel ist durch die obigen Berechnungen mit **R** schon gegeben?

In der von Ihnen (vermutlich) verwendeten Formel kommt ein Vektor vor, der eine besondere Bedeutung hat. Welche?

Lösung: Wir verwenden **R** zur Berechnung von $\mathcal{T}^2$:

```
> 24*(xquer-MUnull)%*%round(SINVDatMV%*%(xquer-MUnull),2)
       [,1]
[1,] 1.9512
```

$\mathcal{T}^2 = 1.9512$

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

Der Vektor $\boldsymbol{a}^*$ ist durch

```
> round(SINVDatMV%*%(xquer-MUnull),2)
```

gegeben. Der Vektor $\boldsymbol{a}^*$ ist der Vektor $\boldsymbol{a}$, der in der ersten Formel für $\mathcal{T}^2$ zum Maximum führt.

b) **(Punkte: 4)**
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?

Nutzen Sie die folgende **R**-Ausgabe, um den P-Wert annähernd zu bestimmen. Dabei sind `df1` und `df2` die Freiheitsgrade der F-Prüfgröße unter der Nullhypothese.

```
> round(qf((1:9)/10,df1,df2),3)
[1] 0.261 0.409 0.551 0.699 0.863 1.055 1.295 1.627 2.195
```

Lösung: $\mathcal{F} = \dfrac{(n-m)\mathcal{T}^2}{m(n-1)} = \dfrac{(24-4)\cdot 1.9512}{4\cdot(24-1)} = \dfrac{20\cdot 1.9512}{92} = 0.4241739 \approx 0.424$

$\mathcal{F}$ hat $m = 4$ und $n - m = 24 - 4 = 20$ Freiheitsgrade.

Der Ablehnungsbereich ist $A = (2.87, \infty)$.

Da $F = 0.424 \notin A$, kann die Nullhypothese nicht verworfen werden.

Der Wert von F liegt am nächsten bei 0.409 und die obige Ausgabe besagt:
$P(F < 0.409) = 0.2$. Daher ist der P-Wert etwas kleiner als 0.8. Eine genaue Berechnung mit **R** besagt:

```
> 1-pf( 0.4241739,4,24)
[1] 0.7896349
```

c) **(Punkte: 2)**
Die obige Nullhypothese $(\mu_1, \mu_2, \mu_3, \mu_4) = (6, 2, 1, 1)$ ist ein Spezialfall der allgemeineren Nullhypothese

$$H_0 : \begin{array}{ccccccccc} \mu_1 & - & \mu_2 & & & & & = & 4 \\ & & & & \mu_3 & + & \mu_4 & = & 2 \\ \mu_1 & & & + & \mu_3 & & & = & 7 \end{array}$$

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: Das obige Gleichungssystem kann in Matrixschreibweise so formuliert werden

$$\boldsymbol{C}^t\boldsymbol{\mu} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 7 \end{pmatrix} = \boldsymbol{\phi}$$

Die Prüfgröße ist

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

d) **(Punkte: 2)**
Die Berechnung der Prüfgröße mit **R** ergab $\mathcal{T}^2 \approx 1.9$.

Berechnen Sie den zugehörigen Wert von F und geben Sie die Verteilung von F unter der Nullhypothese an.

Wie kann man in **R** den zugehörigen P-Wert berechnen?

Lösung: Es ist $\frac{n-p}{p(n-p)}\mathcal{T}^2 \sim F(p, n-p) = F(3, 21)$, denn hier ist $n = 24$ und $p = 3$.
Damit ist $F = \frac{n-p}{p(n-1)} \cdot \mathcal{T}^2 = \frac{21}{3 \cdot 23} \cdot 1.9 = 0.5782609 \approx 0.578$.
Der P-Wert kann also mit `1-pf(0.578,3,21)` berechnet werden.

e) **(Punkte: 3)**
Es sollen simultane und „*gewöhnliche Konfidenzintervalle*" für die Erwartungswerte $\mu_i;\ i = 1, 2, \ldots, 4$ berechnet werden. Rechnen Sie die Konfidenzintervalle nicht aus, sondern geben Sie für beide Fälle die benötigten Formeln und die benötigten Quantile an, wenn $1 - \alpha = 0.9$.

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 und s_i^2 die Varianzen in den Gruppen $i = 1, 2, 3, 4$. Die Varianzen sind die Diagonalemente in der Varianz-Kovarianzmatrix. Es gilt
$K_{0.05} = \left(\frac{m(n-1)}{n-m}F_{0.1}(m, n-m)\right)^{1/2} = \left(\frac{4 \cdot 23}{20} \cdot F_{0.1}(4, 20)\right)^{1/2} = \left(\frac{4 \cdot 23}{20} \cdot 2.25\right)^{1/2} = \sqrt{10.35} = 3.217142 \approx 3.217$.

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 $t_{23,0.05} = 1.71$.

Aufgabe 5 (Punkte: 5)
Im Rahmen einer Diskriminanzanalyse sollen neue Individuen einer von zwei Gruppen zugeordnet werden. Es gelte $\pi_1 = \pi_2$ und $C(1|2) = C(2|1)$. Für Fishers lineare Diskriminanzfunktion ergab sich

$$\boldsymbol{L}^t\boldsymbol{x} = 28x_1 + 0.2x_2$$

Die Mittelwertvektoren seien

$$\bar{\boldsymbol{x}}_1 = (3.5, 550) \qquad \bar{\boldsymbol{x}}_2 = (2.5, 450)$$

a) **(Punkte: 2)**
Berechnen Sie $\hat{m}$ und entscheiden Sie mit Hilfe von $\hat{m}$, welcher Gruppe ein Individuuum mit den Merkmalen $(3, 540)$ zuzuordnen ist.

Lösung: $\hat{m} = \frac{1}{2}\boldsymbol{L}^t(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(28, 0.2)\begin{pmatrix} 6 \\ 1\,000 \end{pmatrix} = \frac{1}{2}(168 + 200) = \frac{1}{2} \cdot 368 = 184.$

Für das neue Individuum gilt $\boldsymbol{L}^t\boldsymbol{x} = 28 \cdot 3 + 0.2 \cdot 540 = 84 + 108 = 192 > 184$, d.h. das Individuum wird der Gruppe 1 zugeordnet.

b) **(Punkte: 3)**
Nehmen Sie an, es gebe eine dritte Gruppe.

Für das obige $\boldsymbol{L}$ gelte

$$\boldsymbol{L}^t = \boldsymbol{L}_1^t - \boldsymbol{L}_2^t = (106, 0.3) - (78, 0.1)$$

Ferner sei $\bar{\boldsymbol{x}}_3 = (3, 450)$ und $\boldsymbol{L}_3^t = (93, 0.15)$.

Nehmen Sie an, dass $\pi_1 = \pi_2 = \pi_3$.

Berechnen Sie für das obige neue Individuum mit den Merkmalen $(3, 540)$ die Diskriminanzwerte $w_i;\ i = 1, 2, 3$ und vernachlässigen Sie dabei den Term $\ln \pi_i$, da er für alle drei Gruppen gleich groß ist.

Welcher Gruppe wird das neue Individuum jetzt zugeordnet?

Lösung: Unter Vernachlässigung des Terms $\ln \pi_i$ gilt $w_i = \boldsymbol{L}_i^t \boldsymbol{x} - \frac{1}{2}\boldsymbol{L}_i^t \bar{\boldsymbol{x}}_i$ und damit

$$w_1 = 106 \cdot 3 + 0.3 \cdot 540 - \frac{1}{2}(106 \cdot 3.5 + 0.3 \cdot 550) = 318 + 162 - \frac{1}{2}(371 + 165) = 212$$

$$w_2 = 78 \cdot 3 + 0.1 \cdot 540 - \frac{1}{2}(78 \cdot 2.5 + 0.1 \cdot 450) = 234 + 54 - \frac{1}{2}(195 + 45) = 168$$

$$w_3 = 93 \cdot 3 + 0.15 \cdot 540 - \frac{1}{2}(93 \cdot 3 + 0.15 \cdot 450) = 279 + 81 - \frac{1}{2}(279 + 67.5) = 186.75$$

Da w_1 am größten ist, wird das neue Individuum der Gruppe 1 zugeordnet.