Multivariates Testen & Clusteranalyse

Aufgabe 1: MANOVA

In einer kalifornischen Studie wurden in 409 Schuldistrikten die mittleren Lese- und Mathefähigkeiten aller Grundschüler erhoben. Mithilfe einer multivariate Varianzanalyse soll nun überprüft werden, wie sich die Fähigkeiten der Schüler unterscheiden, wenn man die Distrikte anhand ihres kategorisierten Einkommensniveaus vergleicht (73 Distrikte haben ein niedriges Niveau, 280 ein mittleres, 67 ein hohes).

a) Notieren Sie die Modellgleichung des MANOVA-Modells in Matrixform. Geben Sie dabei insbesondere den Aufbau und die Dimensionen der einzelnen Matrizen an.

b) Durch die Streuungszerlegung erhalten Sie die Matrizen

$$\boldsymbol{B} = \begin{pmatrix} 71446 & 63907 \\ 63907 & 57855 \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} 97968 & 81919 \\ 81919 & 89514 \end{pmatrix}.$$

Testen Sie auf Basis von Wilks' Λ und mit einem Signifikanzniveau von $\alpha = 0.05$, ob sich die Gruppen signifikant unterscheiden.

Hinweis: Quantile von Wilks' Λ-Verteilung werden üblicherweise nur über Approximationen bestimmt, z.B. anhand von

$$-\left(n - 1 - \frac{p+g}{2}\right) \log \Lambda(p, n-g, g-1) \sim \chi^2_{p(g-1)},$$

wobei p die Anzahl der untersuchten Merkmale, n die Anzahl der Beobachtungen und g die Anzahl der Gruppen.

c) Welche alternativen Möglichkeiten hätten Sie, den möglichen Gruppenunterschied auf Basis einer MANOVA zu testen?

d) Vollziehen Sie den berechneten Test in `R` nach:

(i) Laden Sie den Datensatz `CASchools` aus dem Paket `AER` und kategorisieren Sie das mittlere Einkommen in der Spalte `income` in die drei Kategorien $\{(5, 10], (10, 20], (20, 60]\}$.

(ii) Verschaffen Sie sich einen Überblick über die Daten (die relevanten Zielgroessen sind `read` und `math`). Überprüfen Sie insbesondere, ob die in der MANOVA gewählte Annahme der Varianzhomogenität für die vorliegenden Daten sinnvoll erscheint.

(iii) Schätzen Sie mithilfe der Funktion `stats::manova` die MANOVA und testen Sie mittels `stats::summary.manova` die Fragestellung unter Verwendung des Signifikanzniveaus von $\alpha = 0.05$. Wie können Sie auswählen, welcher Test durchgeführt wird?

a) Was wir wissen:

$$p = 2, g = 3, n = \sum_{k=1}^{g} n_k = 409$$

Modellgleichung:

$$\underbrace{\boldsymbol{Y}}_{\mathbb{R}_{n\times p}} = \underbrace{\boldsymbol{X}}_{\mathbb{R}_{n\times g}} \underbrace{\boldsymbol{M}}_{\mathbb{R}_{g\times p}} + \underbrace{\boldsymbol{\epsilon}}_{\mathbb{R}_{n\times p}}$$

Die einzelnen Komponenten:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_{11}^T \\ \vdots \\ \boldsymbol{y}_{1n_1}^T \\ \vdots \\ \boldsymbol{y}_{g1}^T \\ \vdots \\ \boldsymbol{y}_{gn_g}^T \end{pmatrix} = \begin{pmatrix} y_{11L} & y_{11M} \\ \vdots & \vdots \\ y_{1n_1L} & y_{1n_1M} \\ \vdots & \vdots \\ y_{g1L} & y_{g1M} \\ \vdots & \vdots \\ y_{gn_gL} & y_{gn_gM} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \boldsymbol{\mu}_3^T \end{pmatrix} = \begin{pmatrix} \mu_{1L} & \mu_{1M} \\ \mu_{2L} & \mu_{2M} \\ \mu_{3L} & \mu_{3M} \end{pmatrix},$$

$\boldsymbol{\epsilon}$ ist analog zu $\boldsymbol{Y}$ (mit $\boldsymbol{\epsilon}_{ij}^T$ statt $\boldsymbol{y}_{ij}^T$).

b) $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \quad vs. \quad H_1 : \exists i, j$ mit $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$

Berechne Wilks' Λ mittels

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|}.$$

Berechne Determinanten:

$$|\boldsymbol{W}| = \left| \begin{pmatrix} 97968 & 81919 \\ 81919 & 89514 \end{pmatrix} \right| = 97968 \cdot 89514 - 81919^2 = 2058784991$$

$$|\boldsymbol{B} + \boldsymbol{W}| = \left| \begin{pmatrix} 169414 & 145826 \\ 145826 & 147369 \end{pmatrix} \right| = \ldots = 3701149490$$

Insgesamt:

$$\Lambda = \frac{2058784991}{3701149490} \approx 0.56 \overset{H_0}{\sim} \Lambda(2, 409 - 3, 3 - 1) = \Lambda(2, 406, 2)$$

Umrechnung der Teststatistik, s.d. die approximierte χ^2-Verteilung benutzt werden kann:

$$T = -\left(409 - 1 - \frac{2+3}{2}\right) \log(0.56) \approx 235.$$

Testentscheidung:

$$T = 235 > 9.49 \approx \chi_4^2(0.95)$$

$\Rightarrow H_0$ kann abgelehnt werden!

c) Weitere Testmöglichkeiten sind auf Basis ...

- ... des größten Eigenwerts von $\boldsymbol{W}^{-1}\boldsymbol{B}$
- ... der Spur von $\boldsymbol{W}^{-1}\boldsymbol{B}$ (Lawley-Hotelling)
- ... der Spur von $(\boldsymbol{B}+\boldsymbol{W})^{-1}\boldsymbol{B}$ (Pillai-Spur)

d) siehe `R`!

Aufgabe 2: Hierarchische Clusteranalyse

Für vier Filialen einer Supermarktkette erhält man für die Merkmale Umsatz und Verkaufsfläche, jeweils gemessen in geeigneten Einheiten, die folgende Datenmatrix:

Filiale	1	2	3	4
Umsatz	8	5	10	4
Verkaufsfläche	24	22	25	21

a) Führen Sie ein hierarchisches Clustering mit dem *Single Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die quadrierte euklidische Distanz.

b) Führen Sie ein hierarchisches Clustering mit dem *Zentroid* Verfahren durch.

c) Geben Sie für beide Verfahren das vollständige Dendrogramm an.

Hierarchische Clusteranalyse

- Gegeben: n Objekte mit Messungen $x_1, \ldots, x_n$
- Clustering: Bilden von geeigneten Clustern / Klassen / Gruppen
- Unterteilung von Clusterverfahren
  - agglomerativ, hierarchisch: Teilklassen werden sukzessive zusammengefasst
  - divisiv: Start mit allen Objekten in 1 Cluster, welcher sukzessive aufgesplitted wird
- Vorgehen: Im ersten Schritt bilden alle Objekte ein eigenes Cluster. Fasse Cluster basierend auf Distanzmaßen solange zusammen, bis alle Objekte in einem Cluster zusammengefasst werden.
- $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j) \widehat{=}$ Distanz zwischen den Objekten i und j bei beobachteten Merkmalsvektoren $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$
- $D(C_i, C_j) \widehat{=}$ Distanz zwischen den Clustern C_i und C_j.
- $\mathcal{C}^\nu$ ist definiert als die Partition im ν-ten Schritt.
- $h_\nu \widehat{=}$ Distanz der beiden im Schritt ν fusionierten Cluster (im Dendrogramm abzutragen).

a) Single-Linkage mit quad. euklidischer Distanz $d_{ij} = ||x_i - x_j||^2$: im ν-ten Schritt werden diejenigen Cluster C_i, $C_j \in \mathcal{C}^{(\nu-1)}$ fusioniert, für die gilt

$$D(C_i, C_j) = (h_\nu =) \min_{l \neq k} D(C_l, C_k) = \min_{l \neq k} \left\{ \min_{r \in C_l,\, s \in C_k} \{d_{rs}\} \right\}$$

(1) Distanzmatrix der Partition $\mathcal{C}^{(0)} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$:

$$\text{z.B. } d_{12} = \left|\left| \begin{pmatrix} 8 \\ 24 \end{pmatrix} - \begin{pmatrix} 5 \\ 22 \end{pmatrix} \right|\right|^2 = \left|\left| \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right|\right|^2 = 3^2 + 2^2 = 13$$

$$\begin{array}{r|cccc} & 1 & 2 & 3 & 4 \\ 1\,| & 0 & 13 & 5 & 25 \\ 2\,| & & 0 & 34 & \textcircled{2} \\ 3\,| & & & 0 & 52 \\ 4\,| & & & & 0 \end{array} \quad \Rightarrow \quad h_1 = \min_{l \neq k} \left\{ \min_{r \in C_l,\, s \in C_k} \{d_{rs}\} \right\} = 2 \widehat{=} D(\{2\}, \{4\})$$

$\Rightarrow$ Schritt 1: Fusion von $\{2\}$ und $\{4\}$

$\Rightarrow \mathcal{C}^{(1)} = \{\{1\}, \{2,4\}, \{3\}\}$

(2) Distanzmatrix der Partition $\mathcal{C}^{(1)}$:

$$\begin{array}{r|ccc} & 1 & 2,4 & 3 \\ 1 & 0 & 13 & \textcircled{5} \\ 2,4 & & 0 & 34 \\ 3 & & & 0 \end{array} \quad \Rightarrow \quad h_2 = \min_{l \neq k} \left\{ \min_{r \in C_l,\, s \in C_k} \{d_{rs}\} \right\} = 5 \widehat{=} D(\{1\}, \{3\})$$

$\Rightarrow$ Schritt 2: Fusion von $\{1\}$ und $\{3\}$

$\Rightarrow \mathcal{C}^{(2)} = \{\{1,3\}, \{2,4\}\}$

(3) Distanz zwischen $\{1,3\}$ und $\{2,4\}$:

$h_3 = \min\limits_{r \in \{1,3\},\, s \in \{2,4\}} \{d_{rs}\} = 13 \widehat{=} D(\{1,3\}, \{2,4\})$

$\Rightarrow$ Schritt 3: Fusion von $\{1,3\}$ und $\{2,4\}$

$\Rightarrow \mathcal{C}^{(3)} = \{\{1,2,3,4\}\}$

b) Zentroid-Verfahren mit quad. euklidischer Distanz: Fusioniert werden im ν-ten Schritt diejenigen Cluster C_i, $C_j \in \mathcal{C}^{(\nu-1)}$, für die gilt:

$$D(C_i, C_j) = (h_\nu =) \min_{l \neq k} D(C_l, C_k) = \min_{l \neq k} ||\bar{x}_l - \bar{x}_k||^2 \,, \quad \text{wobei} \quad \bar{x}_r = \frac{1}{n_r} \sum_{s \in C_r} x_s$$

(1) Distanzmatrix der Partition $\mathcal{C}^{(0)} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$:

$$\begin{array}{r|cccc} & 1 & 2 & 3 & 4 \\ 1\,| & 0 & 13 & 5 & 25 \\ 2\,| & & 0 & 34 & \textcircled{2} \\ 3\,| & & & 0 & 52 \\ 4\,| & & & & 0 \end{array} \quad \Rightarrow \quad h_1 = \min_{l \neq k} ||\bar{x}_l - \bar{x}_k||^2 = 2 \widehat{=} D(\{2\}, \{4\})$$

$\Rightarrow$ Schritt 1: Fusion von $\{2\}$ und $\{4\}$

$\Rightarrow \mathcal{C}^{(1)} = \{\{1\}, \{2,4\}, \{3\}\}$

Klassenschwerpunkte:

$$\bar{x}_{\{2,4\}} = \frac{1}{2}\left(\begin{pmatrix}5\\22\end{pmatrix} + \begin{pmatrix}4\\21\end{pmatrix}\right) = \begin{pmatrix}4,5\\21,5\end{pmatrix}$$

$$\Rightarrow \bar{X}^{(1)} = \begin{pmatrix} 8 & 4,5 & 10 \\ 24 & 21,5 & 25 \end{pmatrix}$$
$$\{1\} \quad \{2,4\} \quad \{3\}$$

(2) Distanzmatrix der Partition $\mathcal{C}^{(1)}$:

z.B. $D(\{1\},\{2,4\}) = \left|\left|\begin{pmatrix}8\\24\end{pmatrix} - \begin{pmatrix}4,5\\21,5\end{pmatrix}\right|\right|^2 = \left|\left|\begin{pmatrix}3,5\\2,5\end{pmatrix}\right|\right|^2 = 3,5^2 + 2,5^2 = 18,5$

		1	2,4	3
1	|	0	18,5	(5)
2,4	|		0	42,5
3	|			0

$\Rightarrow \quad h_2 = \min\limits_{l \neq k} ||\bar{x}_l - \bar{x}_k||^2 = 5 \widehat{=} D(\{1\},\{3\})$

$\Rightarrow$ Schritt 2: Fusion von $\{1\}$ und $\{3\}$

$\Rightarrow \mathcal{C}^{(2)} = \{\{1,3\},\{2,4\}\}$

Klassenschwerpunkte:

$$\Rightarrow \bar{X}^{(2)} = \begin{pmatrix} 9 & 4,5 \\ 24,5 & 21,5 \end{pmatrix}$$
$$\{1,3\} \quad \{2,4\}$$

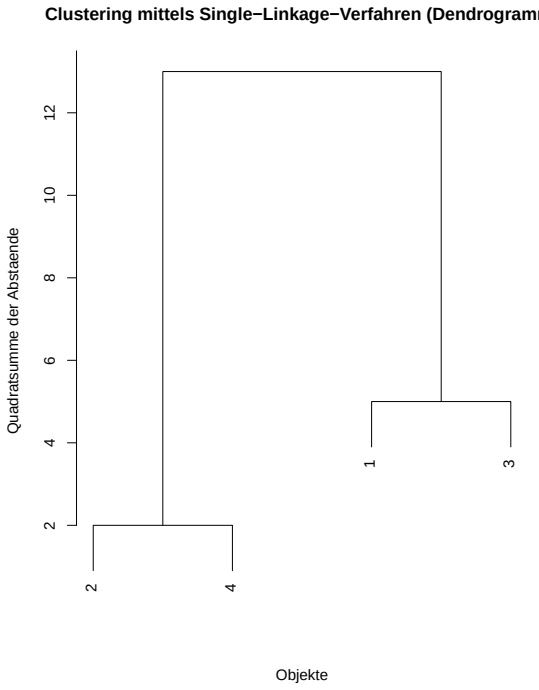
(3) Distanz zwischen $\{1,3\}$ und $\{2,4\}$:

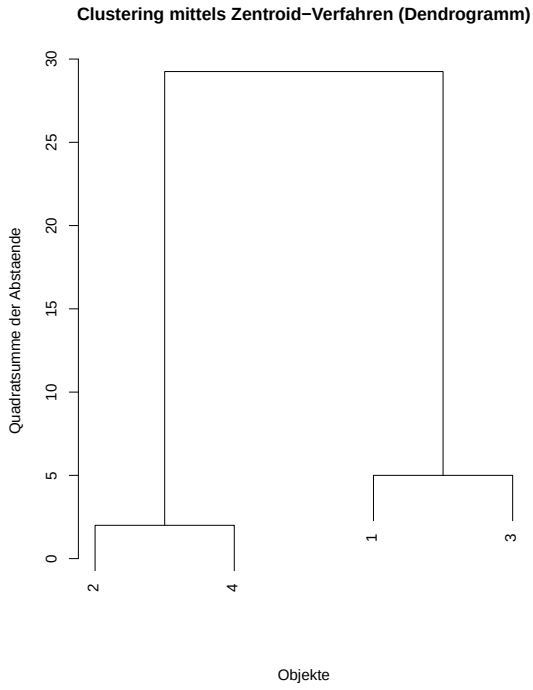
$h_3 = ||\bar{x}_{\{1,3\}} - \bar{x}_{\{2,4\}}||^2 = 4,5^2 + 3^2 = 29,25 \widehat{=} D(\{1,3\},\{2,4\})$

$\Rightarrow$ Schritt 3: Fusion von $\{1,3\}$ und $\{2,4\}$

$\Rightarrow \mathcal{C}^{(3)} = \{\{1,2,3,4\}\}$

c) Dendrogramme für die beiden Cluster-Verfahren


Clustering mittels Single−Linkage−Verfahren (Dendrogramm)
Quadratsumme der Abstaende
2
4
6
8
10
12
2
4
1
3
Objekte


Clustering mittels Zentroid−Verfahren (Dendrogramm)
Quadratsumme der Abstaende
0
5
10
15
20
25
30
2
4
1
3
Objekte

Aufgabe 3: Clusteranalyse in R I

Der Datensatz `europa.txt` enthält Daten zu $n = 24$ europäischen Ländern. Folgende Variablen wurden erhoben: `ober` (Oberfläche in km^2), `einw` (Einwohner in Millionen), `brut` (BIP pro Kopf in \$) und `arbl` (Arbeitslosenquote in %).

a) Lesen Sie den Datensatz in R ein und standardisieren Sie die Daten.

b) Führen Sie mithilfe der Funktion `hclust()` ein hierarchisches Clustering unter Einbeziehung aller vier Kovariablen mit dem *Single Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die quadrierte euklidische Distanz.

c) Führen Sie mithilfe der Funktion `hclust()` ein hierarchisches Clustering unter Einbeziehung aller vier Kovariablen mit dem *Zentroid* Verfahren durch.

d) Führen Sie mithilfe der Funktion `hclust()` ein hierarchisches Clustering unter Einbeziehung aller vier Kovariablen mit dem *Complete Linkage* Verfahren durch. Verwenden Sie als zugrundeliegende Distanz zwischen einzelnen Objekten die Mahalanobis-Distanz.

Hinweis: Die Funktion `mahalanobis()` könnte hilfreich sein.

e) Visualisieren und vergleichen Sie ihre Ergebnisse der Teilaufgaben b), c) und d) jeweils mithilfe eines Dendrogramms.

f) Führen Sie das k-means Clusterverfahren mit Hilfe der Funktion `kmeans()` aus dem Paket `cluster` durch. Wählen Sie dafür $k = 2, \ldots, 10$ und entscheiden Sie anschließend mittels des *elbow criterion* bezogen auf die Summe der quadratischen Abstände innerhalb der Cluster was eine geeignete Anzahl an Clustern sein könnte.

g) Wiederholen Sie die Verfahren aus den Teilaufgaben b), c), d) und f) nur unter Einbeziehung der beiden Variablen `arbl` und `brut`. Vergleichen Sie die Ergebnisse für $k = 4$, indem Sie jeweils die 4 Cluster in 4 verschiedenen Farben im zweidimensionalen Raum plotten.