Multivariate Verfahren

Clusteranalyse

Annika Hoyer

Sommersemester 2020

Clusteranalyse - Inhalt

Grundlagen

Ausgangssituation

- Beobachtung von n Objekten $\Omega = \{a_1, \ldots, a_n\}$ mit zugehörigen Merkmalsvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Ziel: Einteilung in Klassen/Cluster, so dass
  - große Homogenität innerhalb der Cluster
  - große Heterogenität über die Cluster hinweg
- Gesucht: Partition (= disjunkte, vollständige Zerlegung) $C_1, \ldots, C_g$ mit

$$\bigcup_{k=1}^{g} C_k = \{a_1, \ldots, a_n\} \quad \text{wobei} \quad C_k \cap C_\ell = \emptyset \quad \forall k \neq \ell$$

Beispiele

- Jugendliche in verschiedene Typen unterteilen
- Kunden gemäß ihrer Gewohnheiten einteilen (Kundensegmentierung)
- Entwicklungsmerkmale von Kindern in Klassen einteilen
- Bären gemäß ihrer Charakterzüge in Gruppen unterteilen

Merkmale der Clusteranalyse

- Klassen, insbesondere deren Anzahl, ist vorab nicht bekannt, sondern wird gesucht
- Clusteranalyse ist Verfahren des ”unsupervised learning”
- Grundlage für Clusterbildung ist Ähnlichkeits- bzw. Distanzmaß
- Auswahl der Merkmale, die zur Clusterbildung herangezogen werden, ist von zentraler Bedeutung

Verfahren der Clusteranalyse

1. Hierarchische Klassifikationsverfahren
2. Optimale Partitionen
3. Mischverteilungsansätze
4. Stochastische Partitionsverfahren

Merke: Für die Verfahren 2 bis 4 muss die Clusteranzahl im Vorfeld festgelegt werden.

Wiederholung: Ähnlichkeitsmaße

Für Ähnlichkeitsmaße für die Menge $\Omega = \{a_1, \ldots, a_n\}$

$$s : \Omega \; x \; \Omega \to \mathbb{R}$$
$$(a_i, a_\ell) \to s(a_i, a_\ell)$$

postuliert man

$$s(a_i, a_\ell) = s(a_\ell, a_i) \qquad \text{Symmetrie}$$
$$s(a_i, a_\ell) \leq s(a_i, a_i) \qquad \text{Selbstähnlichkeit}$$

Wiederholung: Distanzmaße

Für metrische Distanzmaße für die Menge $\Omega = \{a_1, \ldots, a_n\}$

$$d : \Omega \; x \; \Omega \to \mathbb{R}$$
$$(a_i, a_\ell) \to d(a_i, a_\ell)$$

postuliert man

$$d(a_i, a_\ell) = d(a_\ell, a_i)$$
$$d(a_i, a_i) = 0$$
$$d(a_i, a_\ell) \geq 0 \quad \forall\, i, \ell$$
$$d(a_i, a_\ell) \leq d(a_i, a_r) + d(a_r, a_\ell) \qquad \text{(Dreiecksungleichung)}$$

Minkowski-Distanz

Ein klassisches Distanzmaß für metrische Merkmale ist die Minkowski-Distanz oder L_q-Metrik. Statt $d(a_i, a_\ell)$ werden Merkmalsvektoren eingesetzt, betrachte also $d(\mathbf{x}_i, \mathbf{x}_\ell)$:

$$d_q(\mathbf{x}_i, \mathbf{x}_\ell) = \left(\sum_{j=1}^{p} |\mathbf{x}_{ij} - \mathbf{x}_{\ell j}|^q \right)^{\frac{1}{q}}$$

$$d_1(\mathbf{x}_i, \mathbf{x}_\ell) = \sum_{j=1}^{p} |\mathbf{x}_{ij} - \mathbf{x}_{\ell j}| \quad \text{(Manhattan-Metrik)}$$

$$d_2(\mathbf{x}_i, \mathbf{x}_\ell) = \sqrt{\sum_{j=1}^{p} (\mathbf{x}_{ij} - \mathbf{x}_{\ell j})^2} \quad \text{(Euklidische Distanz)}$$

$$d_\infty(\mathbf{x}_i, \mathbf{x}_\ell) = \sup_j |\mathbf{x}_{ij} - \mathbf{x}_{\ell j}| \quad \text{(Supremumsnorm)}$$

Hierarchische Klassifikationsverfahren

Hierarchische Klassifikationsverfahren

- Bilde Hierarchie von Partitionen $\mathbb{C} = \{C_1, \cdots, C_g\}$ der Menge $\Omega = \{a_1, \cdots, a_n\}$
- Partition C_i besteht aus i Klassen/Clustern
- Partitionen C_i und C_{i+1} haben $i - 1$ Klassen gemeinsam

Agglomerative Verfahren

Starte mit der Partition $\mathbb{C}^{(0)} = \{\{a_1\}, \ldots, \{a_n\}\}$, bei der jede Beobachtung ein eigenes Cluster bildet und fasse die Cluster sukzessive zusammen.

Divisive Verfahren

Starte mit der Partition $\mathbb{C}^{(0)} = \{a_1, \ldots, a_n\}$, bei der alle Beobachtungen ein einziges Cluster bilden und teile die Cluster sukzessive auf.

Beispiel: agglomeratives Verfahren

- Betrachte Alter von 6 Personen: 43, 38, 6, 47, 37, 9
- Sei $\Omega = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$

$\rightarrow$ jede Person mit dem zugehörigen Alter bildet Klasse

- Folge von Partitionen, die durch agglomeratives Verfahren entstehen:

$$
\begin{aligned}
\mathbb{C}^0 &= \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}, \\
\mathbb{C}^1 &= \{\{1\}, \{3\}, \{4\}, \{6\}, \{2, 5\}\}, \\
\mathbb{C}^2 &= \{\{1\}, \{4\}, \{3, 6\}, \{2, 5\}\}, \\
\mathbb{C}^3 &= \{\{1, 4\}, \{3, 6\}, \{2, 5\}\}, \\
\mathbb{C}^4 &= \{\{1, 2, 4, 5\}, \{3, 6\}\}, \\
\mathbb{C}^5 &= \{\{1, 2, 3, 4, 5, 6\}\}
\end{aligned}
$$

- Partitionen $\mathbb{C}^2$ und $\mathbb{C}^3$ haben die Klassen $\{3, 6\}$ und $\{2, 5\}$ gemeinsam

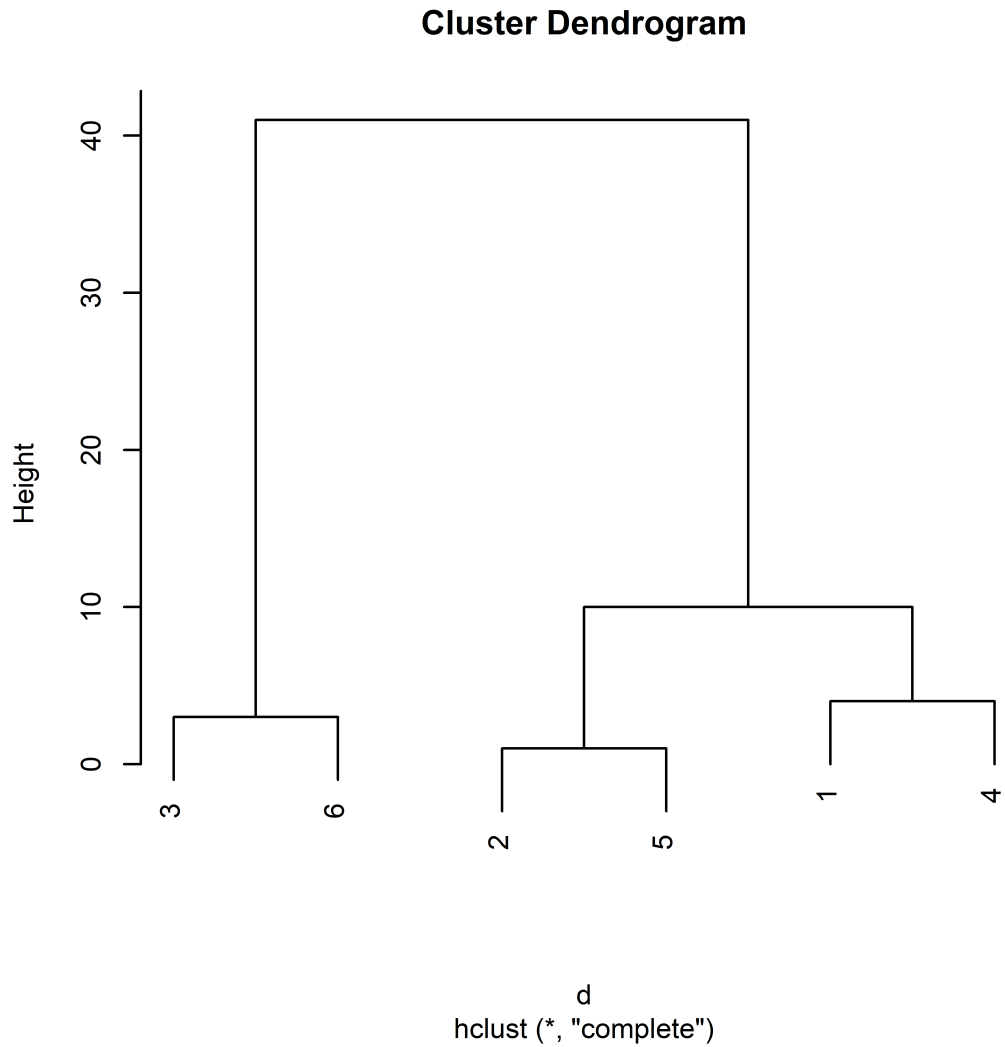
Beispiel: agglomeratives Verfahren

- Bestimme (euklidische) Distanz zwischen 2 Personen
- Zugehörige Distanzmatrix:

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}$$

Beispiel: agglomeratives Verfahren

Darstellung der Partitionen und Distanzen in Dendrogramm



Beispiel: agglomeratives Verfahren

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}$$

Idee: Verschmelze die beiden Klassen, mit dem geringsten Abstand

Beispiel: agglomeratives Verfahren

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}$$

Idee: Verschmelze die beiden Klassen, mit dem geringsten Abstand

Beispiel: agglomeratives Verfahren

1. Schritt: Fusioniere die Klassen 2 und 5 (Alter 37 und 38)

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$					
$\{3\}$		37			
$\{4\}$		4	41		
$\{6\}$		34	3	38	

Beispiel: agglomeratives Verfahren

Wie wird die Distanz zwischen $\{2, 5\}$ und den anderen Klassen bestimmt?

- Beispiel: Distanz zwischen $\{2, 5\}$ und 1
- Bestimme Distanz zwischen 2 und 1 (=5) und 5 und 1 (=6)
- Distanz wird je nach verwendeten Verfahren gewählt, z.B. kleinste oder größte
- Im Beispiel: Maximum der zwei Distanzen (=6)

Beispiel: agglomeratives Verfahren

1. Schritt: Fusioniere die Klassen 2 und 5 (Alter 37 und 38)

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$	6				
$\{3\}$	32	37			
$\{4\}$	10	4	41		
$\{6\}$	29	34	3	38	

Beispiel: agglomeratives Verfahren

2. Schritt: Fusioniere die Klassen 3 und 6 (Alter 6 und 9)

	$\{2,5\}$	$\{1\}$	$\{3,6\}$	$\{4\}$
$\{2,5\}$				
$\{1\}$	6			
$\{3,6\}$	32	37		
$\{4\}$	10	4	41	

Beispiel: agglomeratives Verfahren

3. Schritt: Fusioniere die Klassen 1 und 4 (Alter 43 und 47)

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	10		
$\{3,6\}$	32	41	

Beispiel: agglomeratives Verfahren

4. Schritt: Fusioniere die Klassen $\{1,4\}$ und $\{2,5\}$

	$\{1,2,4,5\}$	$\{3,6\}$
$\{1,2,4,5\}$		
$\{3,6\}$	41	

Agglomerative Verfahren

- Startpartition: $\mathbb{C}^{(0)} = \{C_1^{(0)} = \{a_1\}, \ldots, C_n^{(0)} = \{a_n\}\}$
- Fusioniere im ν-ten Schritt diejenigen Cluster

$$C_r^{(\nu)},\ C_s^{(\nu)}, \quad r \neq s\,,$$

  die die kleinste Distanz D aufweisen.
- Distanz zwischen den Objekten wird durch Distanzmaß bestimmt, die Distanz zwischen den Clustern durch Linkage.
- Verfahren unterscheiden sich bzgl. der Distanz zwischen den Clustern

Single-Linkage Verfahren

$$D_{SL}(C_r, C_s) = \min_{\substack{a_i \in C_r \\ a_\ell \in C_s}} \{d(a_i, a_\ell)\}$$

$\rightarrow$ Fusion bestimmt sich anhand der nächsten Nachbarn.

Single-Linkage Verfahren: Beispiel

Erinnerung: Distanzmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}$$

Single-Linkage Verfahren: Beispiel

1. Schritt: Fusioniere die Klassen 2 und 5
Es gilt:

- $D_{SL}(\{2,5\},\{1\}) = \min\{d_{21}, d_{51}\} = \min\{5,6\} = 5$
- $D_{SL}(\{2,5\},\{3\}) = \min\{d_{23}, d_{53}\} = \min\{32,31\} = 31$
- $D_{SL}(\{2,5\},\{4\}) = \min\{d_{24}, d_{54}\} = \min\{9,10\} = 9$
- $D_{SL}(\{2,5\},\{6\}) = \min\{d_{26}, d_{56}\} = \min\{29,28\} = 28$

Single-Linkage Verfahren: Beispiel

1. Schritt: Fusioniere die Klassen 2 und 5

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$	5				
$\{3\}$	31	37			
$\{4\}$	9	4	41		
$\{6\}$	28	34	3	38	

Single-Linkage Verfahren: Beispiel

2. Schritt: Fusioniere die Klassen 3 und 6
Es gilt:

- $D_{SL}(\{3,6\},\{2,5\}) = \min\{d_{325}, d_{625}\} = \min\{31,28\} = 28$
- $D_{SL}(\{3,6\},\{1\}) = \min\{d_{31}, d_{61}\} = \min\{37,34\} = 34$
- $D_{SL}(\{3,6\},\{4\}) = \min\{d_{34}, d_{64}\} = \min\{41,38\} = 38$

Single-Linkage Verfahren: Beispiel

2. Schritt: Fusioniere die Klassen 3 und 6

	$\{2,5\}$	$\{1\}$	$\{3,6\}$	$\{4\}$
$\{2,5\}$				
$\{1\}$	5			
$\{3,6\}$	28	34		
$\{4\}$	9	4	38	

Single-Linkage Verfahren: Beispiel

3. Schritt: Fusioniere die Klassen 1 und 4
Es gilt:

- $D_{SL}(\{1,4\},\{2,5\}) = \min\{d_{125}, d_{425}\} = \min\{5,9\} = 5$
- $D_{SL}(\{1,4\},\{3,6\}) = \min\{d_{136}, d_{436}\} = \min\{34,38\} = 34$

Single-Linkage Verfahren: Beispiel

3. Schritt: Fusioniere die Klassen 1 und 4

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	5		
$\{3,6\}$	28	34	

Single-Linkage Verfahren: Beispiel

4. Schritt: Fusioniere die Klassen $\{1, 4\}$ und $\{2, 5\}$
Es gilt:

- $D_{SL}(\{1, 2, 4, 5\}, \{3, 6\}) = \min\{d_{1436}, d_{2536}\} = \min\{34, 38\} = 34$

Single-Linkage Verfahren: Beispiel

4. Schritt: Fusioniere die Klassen $\{1,4\}$ und $\{2,5\}$

	$\{1,2,4,5\}$	$\{3,6\}$
$\{1,2,4,5\}$		
$\{3,6\}$	34	

Complete-Linkage Verfahren

$$D_{CL}(C_r, C_s) = \max_{\substack{a_i \in C_r \\ a_\ell \in C_s}} \{d(a_i, a_\ell)\}$$

$\rightarrow$ Fusion bestimmt sich anhand der entferntesten Nachbarn.

Average-Linkage Verfahren

$$D_{AL}(C_r, C_s) = \frac{1}{n_r n_s} \sum_{a_i \in C_r} \sum_{a_\ell \in C_s} d(a_i, a_\ell)\,,$$

wobei $\quad n_k = |C_k|, \quad k \in \{r, s\}.$

$\rightarrow$ Kompromiss zwischen Single-Linkage und Complete-Linkage.

Average-Linkage Verfahren: Beispiel

Erinnerung: Distanzmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}$$

Average-Linkage Verfahren: Beispiel

1. Schritt: Fusioniere die Klassen 2 und 5
Es gilt:

- $D_{AL}(\{2,5\},\{1\}) = \frac{d_{21}+d_{51}}{2} = \frac{5+6}{2} = 5.5$
- $D_{AL}(\{2,5\},\{3\}) = \frac{d_{23}+d_{53}}{2} = \frac{32+31}{2} = 31.5$
- $D_{AL}(\{2,5\},\{4\}) = \frac{d_{24}+d_{54}}{2} = \frac{9+10}{2} = 9.5$
- $D_{AL}(\{2,5\},\{6\}) = \frac{d_{26}+d_{56}}{2} = \frac{29+28}{2} = 28.5$

Average-Linkage Verfahren: Beispiel

1. Schritt: Fusioniere die Klassen 2 und 5

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$	5.5				
$\{3\}$	31.5	37			
$\{4\}$	9.5	4	41		
$\{6\}$	28.5	34	3	38	

Average-Linkage Verfahren: Beispiel

2. Schritt: Fusioniere die Klassen 3 und 6
Es gilt:

- $D_{AL}(\{3,6\},\{2,5\}) = \frac{d_{32}+d_{35}+d_{62}+d_{65}}{4} = \frac{32+31+29+28}{4} = 30$
- $D_{AL}(\{3,6\},\{1\}) = \frac{d_{31}+d_{61}}{2} = \frac{37+34}{2} = 35.5$
- $D_{AL}(\{3,6\},\{4\}) = \frac{d_{34}+d_{64}}{2} = \frac{41+38}{2} = 39.5$

Average-Linkage Verfahren: Beispiel

2. Schritt: Fusioniere die Klassen 3 und 6

	$\{2,5\}$	$\{1\}$	$\{3,6\}$	$\{4\}$
$\{2,5\}$				
$\{1\}$	5.5			
$\{3,6\}$	30	35.5		
$\{4\}$	9.5	4	39.5	

Average-Linkage Verfahren: Beispiel

3. Schritt: Fusioniere die Klassen 1 und 4
Es gilt:

- $D_{AL}(\{1,4\},\{2,5\}) = \frac{d_{12}+d_{15}+d_{42}+d_{45}}{4} = \frac{5+6+9+10}{4} = 7.5$
- $D_{AL}(\{1,4\},\{3,6\}) = \frac{d_{13}+d_{16}+d_{43}+d_{46}}{4} = \frac{37+34+41+38}{4} = 37.5$

Average-Linkage Verfahren: Beispiel

3. Schritt: Fusioniere die Klassen 1 und 4

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	7.5		
$\{3,6\}$	30	37.5	

Average-Linkage Verfahren: Beispiel

4. Schritt: Fusioniere die Klassen $\{1,4\}$ und $\{2,5\}$
Es gilt:

$$
\begin{aligned}
D_{AL}(\{1,2,4,5\},\{3,6\}) &= \frac{d_{13}+d_{16}+d_{43}+d_{46}+d_{23}+d_{26}+d_{53}+d_{56}}{8} \\
&= \frac{37+34+41+38+32+29+31+28}{8} \\
&= 33.75
\end{aligned}
$$

Average-Linkage Verfahren: Beispiel

4. Schritt: Fusioniere die Klassen $\{1, 4\}$ und $\{2, 5\}$

	$\{1, 2, 4, 5\}$	$\{3, 6\}$
$\{1, 2, 4, 5\}$		
$\{3, 6\}$	33.75	

Zentroid-Verfahren

$$D_Z(C_r, C_s) = d_2(\bar{\mathbf{x}}_r, \bar{\mathbf{x}}_s) = \|\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\|^2$$

$$\text{wobei} \quad \bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i, \quad k \in \{r, s\}.$$

$\rightarrow$ Fusion bestimmt sich anhand des Schwerpunkts der Daten.

Zentroid-Verfahren: Beispiel

- Alter von 4 Personen: 19, 25, 20, 23
- Beginn: jede Beobachtung eine Klasse
- Zentroid: Beobachtung selbst

	{1}	{2}	{3}	{4}
{1}				
{2}	36			
{3}	1	25		
{4}	16	4	9	

Zentroid-Verfahren: Beispiel

1. Schritt: Fusioniere die Klassen 1 und 3
Es gilt:

- Mittelwert der Klasse $\{1,3\} = (19+20)/2 = 19.5$
- $D_Z(\{1,3\},\{2\}) = (19.5-25)^2 = 30.25$
- $D_Z(\{1,3\},\{4\}) = (19.5-23)^2 = 12.25$

Zentroid-Verfahren: Beispiel

1. Schritt: Fusioniere die Klassen 1 und 3

	$\{1,3\}$	$\{2\}$	$\{4\}$
$\{1,3\}$			
$\{2\}$	30.25		
$\{4\}$	12.25	4	

Zentroid-Verfahren: Beispiel

2. Schritt: Fusioniere die Klassen 2 und 4
Es gilt:

- Mittelwert der Klasse $\{2,4\} = (25 + 23)/2 = 24$
- $D_Z(\{1,3\},\{2,4\}) = (19.5 - 24)^2 = 20.25$

Zentroid-Verfahren: Beispiel

2. Schritt: Fusioniere die Klassen 2 und 4

	$\{1,3\}$	$\{2,4\}$
$\{1,3\}$		
$\{2,4\}$	20.25	

Vergleich von Zentroid und Average-Linkage

$$
\begin{aligned}
D_{AL}(C_r, C_s) &= \frac{1}{n_r n_s} \sum_{a_i \in C_r} \sum_{a_\ell \in C_s} \{d(a_i, a_\ell)\} \\
&= \frac{1}{n_r n_s} \sum_{\mathbf{x}_i \in C_r} \sum_{\mathbf{x}_\ell \in C_s} \|\mathbf{x}_i - \mathbf{x}_\ell\|^2 \\
&= \|\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\|^2 + \frac{1}{n_r} \sum_{\mathbf{x}_i \in C_r} \|\mathbf{x}_i - \bar{\mathbf{x}}_r\|^2 + \frac{1}{n_s} \sum_{\mathbf{x}_\ell \in C_s} \|\mathbf{x}_\ell - \bar{\mathbf{x}}_s\|^2 \\
&= D_Z(C_r, C_s) + s_r^2 + s_s^2
\end{aligned}
$$

$\rightarrow$ Average-Linkage berücksichtigt die Distanz zwischen den Schwerpunkten und die Streuung.

Verfahren nach Ward

- Motivation: Fusionieren die beiden Cluster, die die minimalste Erhöhung der Varianz (Heterogenität) im neuen Cluster erzeugen
- $$H(\mathbb{C}) = \sum_{r=1}^{g} \sum_{\mathbf{x}_i \in C_r} \|\mathbf{x}_i - \bar{\mathbf{x}}_r\|^2$$

  mit

  $$\bar{\mathbf{x}}_r = \frac{1}{n_r} \sum_{\mathbf{x}_i \in C_r} \mathbf{x}_i$$

Verfahren nach Ward: Beispiel

- Alter von 4 Personen: 19, 25, 20, 23
- Beginn: jedes Objekt eine Klasse $\rightarrow H(\mathbb{C}) = 0$

Verfahren nach Ward: Beispiel

1. Schritt: Betrachte alle Möglichkeiten, zwei Objekte zu einem Cluster zu verschmelzen

- $\mathbb{C} = \{\{1,2\},\{3\},\{4\}\}$:

$$H(\mathbb{C}) = (19-22)^2 + (25-22)^2 + (20-20)^2 + (23-23)^2 = 18$$

- $\mathbb{C} = \{\{1,3\},\{2\},\{4\}\}$:

$$H(\mathbb{C}) = (19-19.5)^2 + (25-25)^2 + (20-19.5)^2 + (23-23)^2 = 0.5$$

- $\mathbb{C} = \{\{1,4\},\{2\},\{3\}\}$:

$$H(\mathbb{C}) = (19-21)^2 + (25-25)^2 + (20-20)^2 + (23-21)^2 = 8$$

- $\mathbb{C} = \{\{2,3\},\{1\},\{4\}\}$:

$$H(\mathbb{C}) = (19-19)^2 + (25-22.5)^2 + (20-22.5)^2 + (23-23)^2 = 12.5$$

- $\mathbb{C} = \{\{2,4\},\{1\},\{3\}\}$:

$$H(\mathbb{C}) = (19-19)^2 + (25-24)^2 + (20-20)^2 + (23-24)^2 = 2$$

- $\mathbb{C} = \{\{3,4\},\{1\},\{2\}\}$:

$$H(\mathbb{C}) = (19-19)^2 + (25-25)^2 + (20-21.5)^2 + (23-21.5)^2 = 4.5$$

Verfahren nach Ward: Beispiel

2. Schritt: Betrachte alle Möglichkeiten, $\mathbb{C} = \{\{1,3\},\{2\},\{4\}\}$ zu fusionieren

- $\mathbb{C} = \{\{1,2,3\},\{4\}\}$:

  $H(\mathbb{C}) = (19-21.33)^2+(25-21.33)^2+(20-21.33)^2+(23-23)^2 = 20.67$

- $\mathbb{C} = \{\{1,3,4\},\{2\}\}$:

  $H(\mathbb{C}) = (19-20.67)^2+(25-25)^2+(20-20.67)^2+(23-20.67)^2 = 8.67$

- $\mathbb{C} = \{\{1,3\},\{2,4\}\}$:

  $H(\mathbb{C}) = (19-19.5)^2+(25-24)^2+(20-19.5)^2+(23-24)^2 = 2.5$

Verfahren nach Ward: Beispiel

3. Schritt: Betrachte alle Möglichkeiten, $\mathbb{C} = \{\{1,3\},\{2,4\}\}$ zu fusionieren

- $\mathbb{C} = \{\{1,2,3,4\}\}$:

$$\begin{aligned} H(\mathbb{C}) &= (19-21.75)^2 + (25-21.75)^2 \\ &+ (20-21.75)^2 + (23-21.75)^2 \\ &= 22.75 \end{aligned}$$

Rekursionsbeziehung von Lance und Williams

Betrachte die Rekursionsformel für die Distanz zwischen Cluster C_r und $C = C_s \cup C_{\tilde{s}}$:

$$D(C_r, C) = \alpha_s\, D(C_s, C_r) + \alpha_{\tilde{s}}\, D(C_{\tilde{s}}, C_r) + \beta\, D(C_s, C_{\tilde{s}}) + \gamma\, |D(C_s, C_r) - D(C_{\tilde{s}}, C_r)|$$

Verfahren	α_s	$\alpha_{\tilde{s}}$	β	γ
Single-Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete-Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average-Linkage	$\frac{n_s}{n_s+n_{\tilde{s}}}$	$\frac{n_{\tilde{s}}}{n_s+n_{\tilde{s}}}$	0	0
Zentroid	$\frac{n_s}{n_s+n_{\tilde{s}}}$	$\frac{n_{\tilde{s}}}{n_s+n_{\tilde{s}}}$	$-\frac{n_s n_{\tilde{s}}}{(n_s+n_{\tilde{s}})^2}$	0
Ward	$\frac{n_s+n_r}{n_s+n_{\tilde{s}}+n_r}$	$\frac{n_{\tilde{s}}+n_r}{n_s+n_{\tilde{s}}+n_r}$	$-\frac{n_r}{n_s+n_{\tilde{s}}+n_r}$	0

Beispiel: Ward-Verfahren nach Rekursionsbeziehung

- Distanzmatrix mit quadrierten euklidischen Distanzen:

$$\mathbf{D} = \begin{pmatrix} 0 & 36 & 1 & 16 \\ 36 & 0 & 25 & 4 \\ 1 & 25 & 0 & 9 \\ 16 & 4 & 9 & 0 \end{pmatrix}$$

- Kleinstes Element 1 $\rightarrow$ fusioniere $\{1\}$ und $\{3\}$

$$\begin{aligned} D_W(\{1,3\},\{2\}) &= \frac{2}{3}D(\{1\},\{2\}) + \frac{2}{3}D(\{3\},\{2\}) - \frac{1}{3}D(\{1\},\{3\}) \\ &= \frac{2}{3}\cdot 36 + \frac{2}{3}\cdot 25 - \frac{1}{3}\cdot 1 = \frac{121}{3} \end{aligned}$$

$$\begin{aligned} D_W(\{1,3\},\{4\}) &= \frac{2}{3}D(\{1\},\{4\}) + \frac{2}{3}D(\{3\},\{4\}) - \frac{1}{3}D(\{1\},\{3\}) \\ &= \frac{2}{3}\cdot 16 + \frac{2}{3}\cdot 9 - \frac{1}{3}\cdot 1 = \frac{49}{3} \end{aligned}$$

Beispiel: Ward-Verfahren nach Rekursionsbeziehung

Erster Schritt des Ward-Verfahrens:

	$\{1,3\}$	$\{2\}$	$\{4\}$
$\{1,3\}$			
$\{2\}$	121/3		
$\{4\}$	49/3	4	

$$
\begin{aligned}
D_W(\{2,4\},\{1,3\}) &= \frac{3}{4}D(\{2\},\{1,3\}) + \frac{3}{4}D(\{4\},\{1,3\}) - \frac{2}{4}D(\{2\},\{4\}) \\
&= \frac{3}{4}\cdot\frac{121}{3} + \frac{3}{4}\cdot\frac{49}{3} - \frac{2}{4}\cdot 4 = \frac{162}{4} = 40.5
\end{aligned}
$$

Beispiel: Ward-Verfahren nach Rekursionsbeziehung

Zweiter Schritt des Ward-Verfahrens:

	$\{1,3\}$	$\{2,4\}$
$\{1,3\}$		
$\{2,4\}$	40.5	

$\rightarrow$ fusioniere $\{1,3\}$ und $\{2,4\}$

Eigenschaften agglomerativer Clusterverfahren

- Single-Linkage neigt zur Kettenbildung (eignet sich zur Identifikation von Ausreißern).
- Average-Linkage und Ward führen zu sehr homogenen Clustern.
- Complete-Linkage ist empfindlicher gegenüber kleinen Änderungen in den Daten als Single-Linkage.
- Zentroid und Ward sind nur für metrische Merkmale anwendbar.
- Zentroid und Ward können zu Inversion führen (Distanzmaß verringert sich im Vergleich zum vorherigen Schritt)

Praktische Aspekte - Güte der Lösung

- Kophenetischer Korrelationskoeffizient
- Bestimmt Korrelation zwischen Distanzen der Distanzmatrix und der kophenetischen Matrix (Distanzmatrix, die sich aus dem Dendrogramm ergibt)

$\rightarrow$ Entscheidung für das Verfahren mit dem größten Korrelationskoeffizienten

Praktische Aspekte - Güte der Lösung

- Gamma-Koeffizient
- Idee: Bestimme die konkordanten (C) und diskordanten (D) Paare der Distanzen zwischen Distanzmatrix und kophenetischer Matrix
- Berechne $\gamma = \frac{C-D}{C+D}$
- Beurteilung der Clusterlösung:

Wert von γ	Beurteilung
$0.9 \leq \gamma \leq 1.0$	sehr gut
$0.8 \leq \gamma < 0.9$	gut
$0.7 \leq \gamma < 0.8$	befriedigend
$0.6 \leq \gamma < 0.7$	noch ausreichend
$0.0 \leq \gamma < 0.6$	nicht ausreichend

Distanzmaße für binäre Merkmale

Gegeben sind

$$\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})^\top, \text{ mit } x_{ij} \in \{0, 1\},$$
$$\mathbf{x}_\ell = (\mathbf{x}_{\ell 1}, \ldots, \mathbf{x}_{\ell p})^\top, \text{ mit } x_{\ell j} \in \{0, 1\}.$$

Betrachte die Anzahlen

$$n_{00} = \sum_{j=1}^{p} \mathbf{1}(x_{ij} = 0, x_{\ell j} = 0), \quad n_{01} = \sum_{j=1}^{p} \mathbf{1}(x_{ij} = 0, x_{\ell j} = 1),$$
$$n_{10} = \sum_{j=1}^{p} \mathbf{1}(x_{ij} = 1, x_{\ell j} = 0), \quad n_{11} = \sum_{j=1}^{p} \mathbf{1}(x_{ij} = 1, x_{\ell j} = 1),$$

mit $p = n_{00} + n_{01} + n_{10} + n_{11}$.

Distanzmaße für binäre Merkmale

1. Simple Matching (M-Koeffizient)

$$d_M(a_i, a_\ell) = \frac{n_{00} + n_{11}}{p}$$

2. Jaccard (S-Koeffizient)

$$d_S(a_i, a_\ell) = \frac{n_{11}}{n_{01} + n_{10} + n_{11}}$$

Insbesondere gilt:

$$d_S(a_i, a_\ell) = \phi^2 \quad (\phi\text{-Koeffizient der } 2x2\text{-Tafel})$$

Distanzmaße für binäre Merkmale: Beispiel

Student	Geschlecht	Raucher	Auto	MatheLK
1	0	0	1	1
2	0	0	1	0

$$d_M = \frac{2+1}{4} = 0.75$$

$$d_S = \frac{1}{0+1+1} = 0.5$$

Optimale Partitionen

Optimale Partitionen

- Messe Qualität einer Partition $\mathbb{C}$ anhand eines Gütekriteriums
- Gesucht: Partition, die hinsichtlich des Gütekriteriums optimal ist, nämlich

$$H(\mathbb{C}_{\text{opt}}) = \min_{\mathbb{C}} H(\mathbb{C})$$

- Betrachtung aller zulässigen Partitionen (”Try all, keep best”) scheitert an Dimension
- Anzahl der Möglichkeiten n Objekte in g Gruppen aufzuteilen:

$$\frac{1}{g!} \sum_{r=0}^{g-1} (-1)^r \binom{g}{r} (g-r)^n$$

- Beispiel: 20 Objekte in 2 Gruppen aufteilen $\rightarrow$ 524287 Partitionen

Numerische Lösung durch Austauschverfahren

Idee:

- Bestimme Ausgangspartition
- Verschiebe Elemente zwischen den Gruppen gemäß eines Gütekriteriums iterativ bis zu einem Optimum
- Verschiedene Verfahren resultieren aus unterschiedlichen Gütekriterien

Numerische Lösung durch Austauschverfahren

1. Wähle eine Ausgangspartition $\mathbb{C}^{(0)}$.
2. Prüfe in der Partition $\mathbb{C}^{\nu}$ $(\nu \geq 0)$ für jedes Objekt, ob die Zuordnung in einen anderen Cluster das Gütekriterium H verbessert.
3. Teile jenes Objekt, das die größte Verbesserung in H ergibt, dem entsprechenden Cluster zu. Damit erhält man die neue Partition $\mathbb{C}^{\nu+1}$.
4. Iteriere Schritt 2 und 3 bis keine Verbesserung mehr entritt.

Numerische Lösung durch Austauschverfahren

Beachte:
Je nach Wahl der Ausgangspartition können sich (suboptimale) lokale Lösungen ergeben.

- Verwende mehrere Startpartionen
- Wähle als geschätzte, optimale Partition die Partition mit dem größten bzw. kleinsten Wert des Gütekriteriums

Im Folgenden werden verschiedene Gütekriterien definiert, die sich hinsichtlich ihrer Separationseigenschaften unterscheiden.

Varianzkriterium (k-Means-Algorithmus)

- Ausgangspunkt: Zentren
- Minimiere die Gesamt-intra-Cluster-Varianz bzw. Gesamtsumme der quadrierten Abweichungen

$\rightarrow$ Minimiere das Gütekriterium

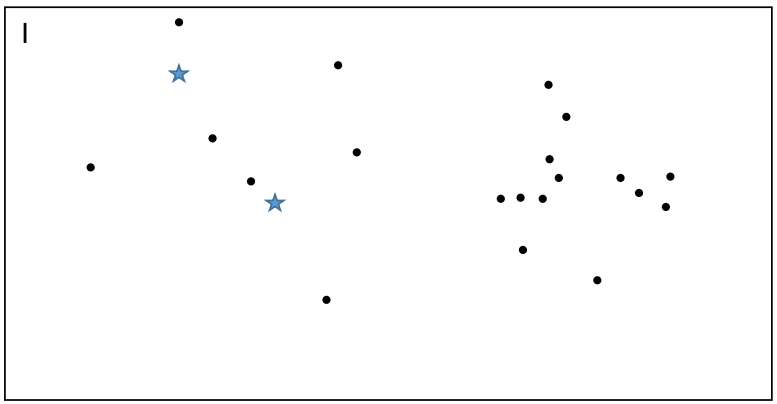
$$H(\mathbb{C}) = \sum_{r=1}^{g} \sum_{\mathbf{x}_i \in C_r} \|\mathbf{x}_i - \bar{\mathbf{x}}_r\|^2,$$

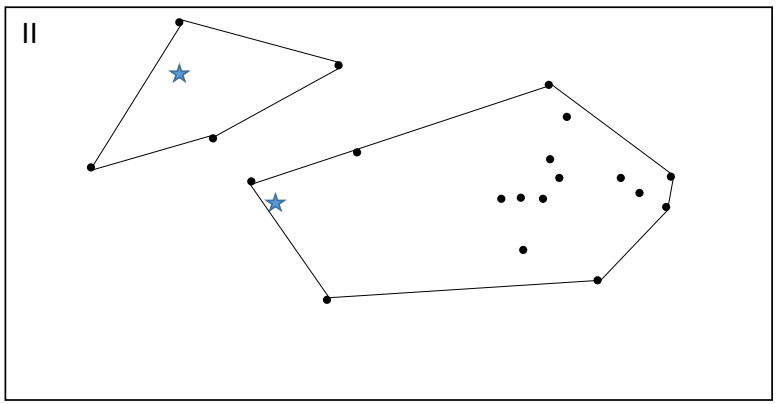
wobei es g Gruppen/ Cluster gibt und die $\bar{\mathbf{x}}_r$ die Schwerpunkte der Cluster sind

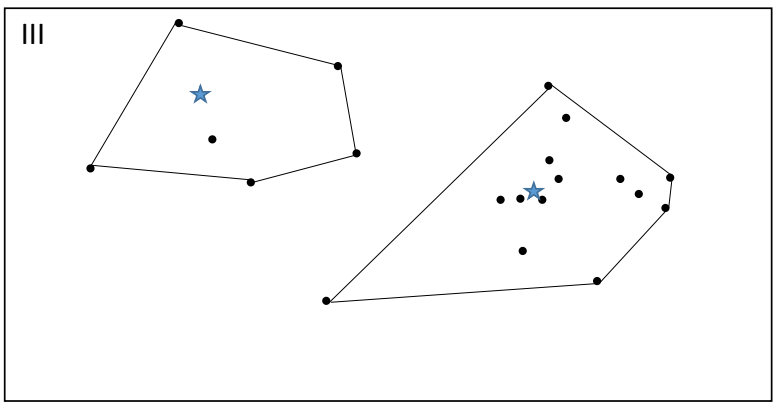
k-Means-Algorithmus

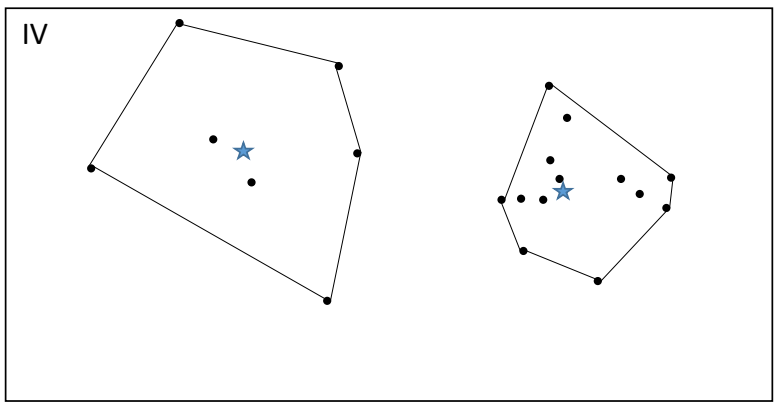
1. Es wird eine Anfangspartition vorgegeben.
2. Zur vorliegenden Partition werden die Gruppenschwerpunkte berechnet.
3. Jedes Element wird in die Gruppe verschoben, die den im Sinn der euklidischen Distanz am nächsten liegenden Schwerpunkt besitzt.
4. Es wird zu (2) zurückgegangen, falls mindestens ein Element die Gruppe gewechselt hat. Andernfalls ist die Gruppenbildung abgeschlossen.

k-Means-Algorithmus









Beispiel: Evaluation von Fachbereichen

Bewertung von Lehr- und Forschungseinheiten (LFE) nach

- wissenschaftlichen Veröffentlichungen (X_1, Punkte)
- eingworbenen Drittmitteln (X_2, TDM)
- Anzahl der Promotionen (X_3)
- Anzahl der Absolventen (X_4)
- Anzahl der Studierenden (X_5)

Beispiel: Evaluation von Fachbereichen

Ergebnis des k-means-Verfahrens:

	C-Nr.	$(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5)$
2 Cluster	1	(13.724, 27.944, 1.142, 23.299, 92.818)
	2	(20.802, 179.020, 3.460, 31.962, 67.637)
3 Cluster	1	(12.893, 11.851, 1.222, 30.808, 111.773)
	2	(14.462, 42.249, 1.072, 16.625, 75.968)
	3	(20.802, 179.020, 3.460, 31.962, 67.637)

- 2-Cluster-Lösung: vier LFE mit besonders vielen Drittmitteln, sowie im Spitzenbereich bei Veröffentlichungen und Promotionen in Cluster 2 abspalten
- 3-Cluster-Lösung: Abspaltung des anderen Clusters, sodass LFE mit hoher Anzahl an Studierenden und Absolventen eigenständigen Cluster bilden, Drittmittel am geringsten

Varianzkriterium: Eigenschaften

- Liegt optimale Partition $\mathbb{C}_{\text{opt}} = \left\{ C_1^{(\text{opt})}, \ldots, C_g^{(\text{opt})} \right\}$ vor, so gilt für alle $i \in C_k^{(\text{opt})}$:

$$\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2 \leq \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2, \quad \forall j \neq k$$

  (Minimal-Distanz-Eigenschaft)
- Varianzkriterium ergibt kugelförmige Cluster mit etwa gleich vielen Objekten (Separationseigenschaft)

Varianzkriterium: Bemerkungen

- Es kann in jedem Schritt vorkommen, dass Objekte zu mehreren Klassenmittelpunkten die minimale Distanz aufweisen
- Durch Austauschverfahren sukzessiv erzeugte Partitionen sind hinsichtlich des Varianzkriteriums nie schlechter:

$$H(\mathbb{C}^0) \geq H(\mathbb{C}^1) \geq \ldots \geq H(\mathbb{C}^\nu) \geq H(\mathbb{C}^{\nu+1}) \,.$$

- Das Austauschverfahren kann zu leeren Mengen führen, wenn ein Klassenmittelpunkt für kein Objekt der nächstgelegene Klassenmittelpunkt ist

Determinantenkriterium

- Erinnerung: Varianzzerlegung (Kap. 2): Gesamtstreuung = Inner-Gruppen-Streuung + Zwischen-Gruppen-Streuung

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

- Determinantenkriterium: Minimiere das Gütekriterium

$$H(\mathbb{C}) = |\mathbf{W}(\mathbb{C})|$$

- Interpretation: Durch Verwendung der Determinante werden die Kovarianzen zwischen den Merkmalen mit einbezogen.

Zusammenhang zur MANOVA

- Gegeben: Grundgesamtheit, die in g Gruppen partitioniert ist, d.h. die Daten

$$\mathbf{x}^r_{(1)}, \ldots, \mathbf{x}^r_{(n_r)} \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}), \quad r = 1, \ldots, g$$

- Hypothesen:
  $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g \quad \text{vs.} \quad H_1 : \exists k, \ell : \boldsymbol{\mu}_k \neq \boldsymbol{\mu}_\ell$
- Teststatistik: $\Lambda = \frac{|\mathbf{W}(\mathbb{C})|}{|T|}$
- Interpretation: Je kleiner $|\mathbf{W}(\mathbb{C})|$ ausfällt, desto mehr unterscheiden sich die Erwartungswertvektoren und somit die Klassen.

Determinantenkriterium: Eigenschaften

- Liegt die optimale Partition $\mathbb{C}_{\text{opt}} = \left\{ C_1^{(\text{opt})}, \ldots, C_g^{(\text{opt})} \right\}$ vor, so gilt für alle $i \in C_k^{(\text{opt})}$:

  $$(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top \mathbf{W}(\mathbb{C})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \leq (\mathbf{x}_i - \bar{\mathbf{x}}_j)^\top \mathbf{W}(\mathbb{C})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_j), \quad \forall j \neq k$$

  (Minimal-Distanz-Eigenschaft)
- Determinantenkriterium ergibt ellipsenförmige Cluster mit gleich langen, parallel verlaufenden Achsen, d.h. mit gleicher Ausrichtung (Separationseigenschaft).

Verallgemeinertes Determinantenkriterium

Minimiere das Gütekriterium

$$H(\mathbb{C}) = \sum_{r=1}^{g} n_r \log \left| \frac{1}{n_r} \mathbf{W}(C_r) \right|,$$

wobei $\mathbf{W}(\mathbb{C}) = \sum_{r=1}^{g} \mathbf{W}(C_r)$.

$\rightarrow$ Betrachtet wird jeweils die Streuung innerhalb eines Clusters r.

Verallgemeinertes Determinantenkriterium: Eigenschaften

- Liegt die optimale Partition $\mathbb{C}_{\text{opt}} = \left\{ C_1^{(\text{opt})}, \ldots, C_g^{(\text{opt})} \right\}$ vor, so gilt für alle $i \in C_k^{(\text{opt})}$:

  $$(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top \mathbf{W}(C_k)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \leq (\mathbf{x}_i - \bar{\mathbf{x}}_j)^\top \mathbf{W}(C_j)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_j), \quad \forall j \neq k$$

  (Minimal-Distanz-Eigenschaft)
- Verallgemeinertes Determinantenkriterium ergibt ellipsenförmige Cluster mit durchaus unterschiedlicher Ausrichtung (Separationseigenschaft).

Mischverteilungsansätze

Mischverteilungsansätze - Idee

- Beobachtungen sind Realisierungen eines Zufallsvektors
- Zufallsvektor hat in jeder Klasse eine andere Verteilung

$\rightarrow$ Beobachtungen stammen aus Mischverteilung mit unbekannten Parametern

Mischverteilungsansätze

- Annahme: Grundgesamtheit in g Gruppen $C_1, \ldots, C_g$ partitioniert
- Betrachten Zufallsvektoren $\mathbf{X}_r,\ r = 1, \ldots, g$
- Dahinter: g Grundgesamtheiten $Y \in \{1, \ldots, g\}$, wobei
  - $p(r) = P(Y = r)$: Mischanteil
  - $f(\mathbf{x}|\theta_r)$: Dichte des Zufallsvektors $\mathbf{X}$ in der r-ten Population mit Parametervektor θ_r.
- In Grundgesamtheit folgt $\mathbf{x}$ der Mischverteilung:

$$f(\mathbf{x}) = \sum_{r=1}^{g} p(r) f(\mathbf{x}|\theta_r)$$

Identifizierbarkeit

Identifizierbarkeit

Die zur Verteilungsfamilie $\{f(\mathbf{x}|\theta), \theta \in \Theta\}$ gehörige Familie aller endlichen Mischungen, d.h. die Familie aller Verteilungen, die sich in der Form

$$f(\mathbf{x}) = \sum_{r=1}^{g} p(r) f(\mathbf{x}|\theta_r)$$

schreiben lassen, heißt identifizierbar, wenn die Klassenzahl g und $p(r), \theta_r, r = 1, \ldots, g$ bis auf Umnummerierung eindeutig bestimmt sind.

Identifizierbarkeit - Beispiel: Bernoulli-Verteilung

Betrachte zwei Münzen, wobei die Wahrscheinlichkeit für Kopf gleich π_1 bzw. π_2 ist. Damit gilt:

$$
\begin{aligned}
f(x) &= p(1)\pi_1^x(1-\pi_1)^{(1-x)} + p(2)\pi_2^x(1-\pi_2)^{(1-x)} \\
&= \begin{cases} p(1)\pi_1 + p(2)\pi_2, & \text{falls } x = 1 \\ p(1)(1-\pi_1) + p(2)(1-\pi_2), & \text{falls } x = 0 \end{cases}
\end{aligned}
$$

Definiere: $\pi = p(1)\pi_1 + p(2)\pi_2$
$\rightarrow x \sim B(1, \pi)$ ist nicht identifizierbar!

Multivariate Normalverteilung

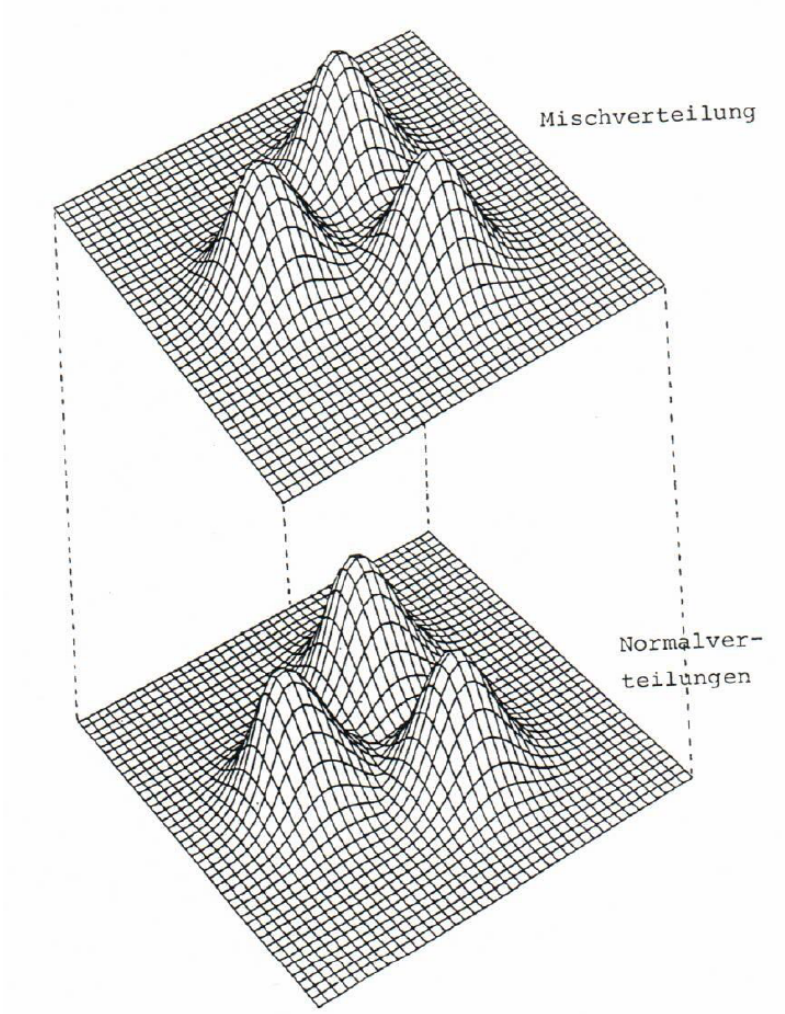
Für $p > 1$ erhält man eine identifizierbare, endliche Mischung für die p-dimensionale Normalverteilung, d.h. $\mathbf{X}_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ mit Dichte:

$$f(\mathbf{x}_i|\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma_r} \mid^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_r)^\top \boldsymbol{\Sigma}_r^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_r) \right\}.$$

Mit der Eigenvektorzerlegung der Kovarianzmatrix ergibt sich:
$\boldsymbol{\Sigma}_r = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$

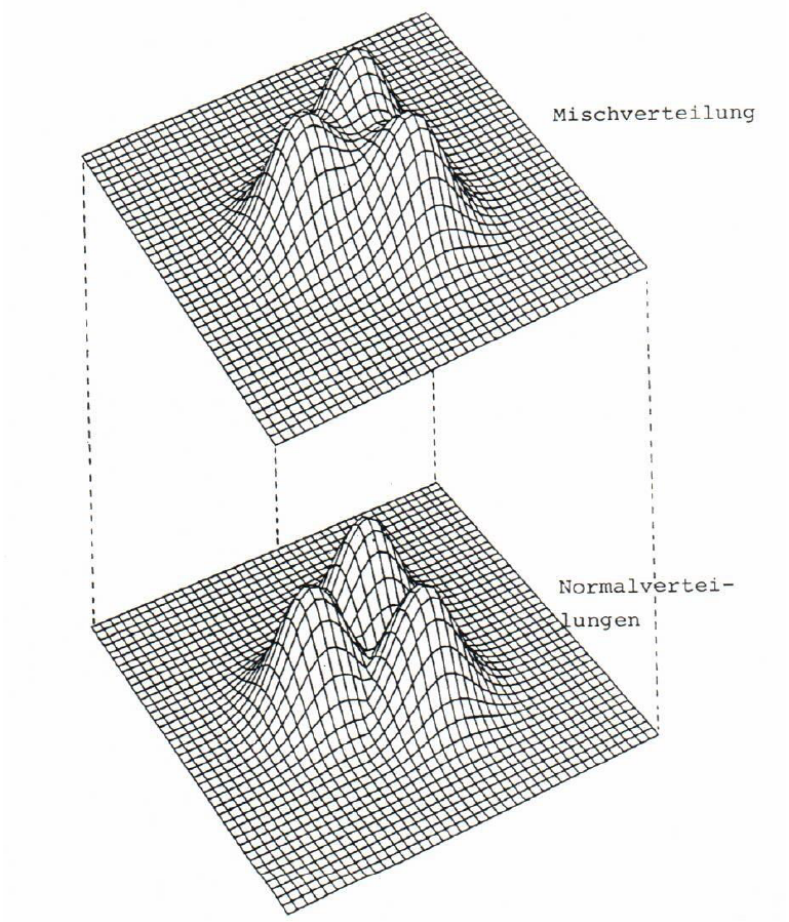
Beispiel I

Drei zweidimensionale Normalverteilungen und deren Mischverteilung für p(1)=p(2)=p(3)=1/3. Überlappung: 5%



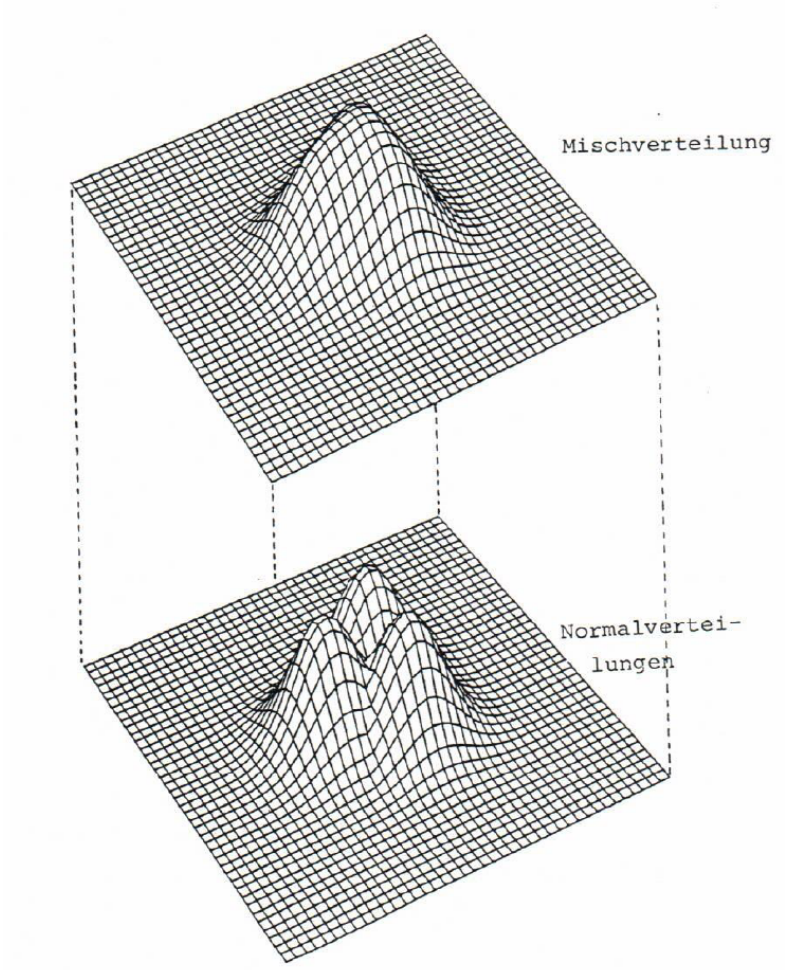
Beispiel II

Drei zweidimensionale Normalverteilungen und deren Mischverteilung für p(1)=p(2)=p(3)=1/3. Überlappung: 15%



Beispiel III

Drei zweidimensionale Normalverteilungen und deren Mischverteilung für p(1)=p(2)=p(3)=1/3. Überlappung: 25%



Modellbasierte Clusteranalyse

Verschiedene Annahme für die Varianzstruktur der Cluster

Modell	Σ_k	Verteilung	Volumen	Form	Ausrichtung
EII	λI	Sphärisch	Identisch	Identisch	-
VII	$\lambda_k I$	Sphärisch	Variabel	Identisch	-
EEI	λA	Diagonal	Identisch	Identisch	Koordinatenachsen
VEI	$\lambda_k A$	Diagonal	Variabel	Identisch	Koordinatenachsen
EVI	λA_k	Diagonal	Identisch	Variabel	Koordinatenachsen
VVI	$\lambda_k A_k$	Diagonal	Variabel	Variabel	Koordinatenachsen
EEE	λDAD^T	Elliptisch	Identisch	Identisch	Identisch
EVE	$\lambda DA_k D^T$	Elliptisch	Identisch	Variabel	Identisch
VEE	$\lambda_k DAD^T$	Elliptisch	Variabel	Identisch	Identisch
VVE	$\lambda_k DA_k D^T$	Elliptisch	Variabel	Variabel	Identisch
EEV	$\lambda D_k AD_k^T$	Elliptisch	Identisch	Identisch	Variabel
VEV	$\lambda_k D_k AD_k^T$	Elliptisch	Variabel	Identisch	Variabel
EVV	$\lambda D_k A_k D_k^T$	Elliptisch	Identisch	Variabel	Variabel
VVV	$\lambda_k D_k A_k D_k^T$	Elliptisch	Variabel	Variabel	Variabel

Verwende BIC-Kriterium $BIC = 2ln(x, \hat{\theta}) - p \cdot ln(n)$

Schätzung der Parameter

Seien $\mathbf{x}_1, \ldots, \mathbf{x}_n$ unabhängig identisch verteilte Beobachtungen eines Merkmals $\mathbf{x}$ aus einer endlichen Mischung, so ergibt sich die log-Likelihood-Funktion

$$\ell = \sum_{i=1}^{n} \log \left(\sum_{r=1}^{g} p(r) f(\mathbf{x}_i | \boldsymbol{\theta}_r) \right)$$

Zur Bestimmung der Schätzer $\hat{p}_r$ und $\hat{\theta}_r,\ r = 1, \ldots, g$, gelten folgende Nebenbedingungen:

$$\sum_{r=1}^{g} p(r) = 1 \quad \text{und} \quad p(r) > 0 \ \forall r$$

Schätzung der Parameter

Damit ergeben sich die folgenden Schätzgleichungen

$$(1) \quad \hat{p}_{ri} = \hat{p}(r|\mathbf{x}_i) = \frac{\hat{p}(r) f(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_r)}{\sum_{j=1}^{g} \hat{p}(j) f(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_j)}$$

$$(2) \quad \hat{p}(r) = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{ri}$$

$$(3) \quad \sum_{i=1}^{n} \hat{p}_{ri} \frac{\partial \log(f(\mathbf{x}_i|\hat{\boldsymbol{\theta}}_r))}{\partial \boldsymbol{\theta}_r} = 0$$

Interpretation: Die Wahrscheinlichkeiten $\hat{p}_{ri}$ sind die geschätzten a-posteriori Wahrscheinlichkeiten, dass die Beobachtung i mit Merkmalsvektor $\mathbf{x}_i$ aus der Klasse r stammt.

Schätzung der Parameter

Lösung der Schätzgleichungen mithilfe des EM-Algorithmus: Generelle Methode zur Maximum-Likelihood-Schätzung mit fehlenden Daten.

1. Wähle Startwerte $\hat{\theta}_r^{(0)}$, $r = 1, \ldots, g$.
2. E-Schritt: Bestimme im ν-ten Schritt die a-posteriori Wahrscheinlichkeiten $\hat{p}_{ri}^{(\nu)}$ für geschätzte Parameter $\hat{\theta}_r^{(\nu-1)}$.
3. M-Schritt: Bestimme im ν-ten Schritt die Parameterschätzer $\hat{\theta}_r^{(\nu)}$ durch Maximierung der gewichteten log-Likelihood für gegebene a-posteriori Wahrscheinlichkeiten $\hat{p}_{ri}^{(\nu)}$.
4. Iteriere Schritt 2 und 3 bis zur Konvergenz.

Mischverteilungsansätze: R-Pakete

1. **mclust: Gaussian Mixture Modelling for Model-Based Clustering**

   Entscheidend für die Modellierung ist die Annahme über die Kovarianzstruktur $\mathbf{\Sigma}_r$ der Cluster. Diese kann man über die Charakteristiken (Verteilung, Volumen, Form und Ausrichtung) spezifizieren.
   Beispiele: `EII` (sphärisch, identisch, identisch, identisch) und `VVE` (elliptisch, variabel, variabel, identisch).
2. **flexmix: A General Framework for Finite Mixtures**

Stochastische Partitionsverfahren

Stochastische Partitionsverfahren

- Annahme: Grundgesamtheit ist in g Gruppen $C_1, \ldots, C_g$ partitioniert
- Wahre Klassenzugehörigkeit: fest aber unbeobachtet
- Zu jedem Merkmalsvektor $\mathbf{x}_1, \ldots, \mathbf{x}_n$ gehört feste unbekannte Klasse $k_1, \ldots, k_n \in \{1, \ldots, g\}$
- Innerhalb der Klasse r sind Beobachtungen identisch verteilt mit $f(\mathbf{x}|\theta_r)$

Maximum-Likelihood Schätzung

- Log-Likelihood:

$$\ell(k_1, \ldots, k_n, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g) = \sum_{i=1}^{n} \log\left(f(\mathbf{x}_i|\boldsymbol{\theta}_{k_i})\right)$$

- Alternativ mit $C_r = \{i|k_i = r\}$:

$$\ell(C_1, \ldots, C_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g) = \sum_{r=1}^{g} \sum_{i=1}^{n} \log\left(f(\mathbf{x}_i|\boldsymbol{\theta}_r)\right)$$

Maximum-Likelihood Schätzung

- Für feste Partition $\mathbb{C} = \{C_1, \ldots, C_g\}$ Schätzung von $\boldsymbol{\theta}_r$ aus Maximierung von

$$\ell_r = \sum_{\mathbf{x}_i \in C_r} \log(f(\mathbf{x}_i|\boldsymbol{\theta}_r)),$$

$\rightarrow$ $\hat{\boldsymbol{\theta}}_r$ kann für jeden Cluster $r = 1 \ldots, g$ mit klassischer Maximum-Likelihood Schätzung getrennt geschätzt werden

Profile Likelihood

- Eisetzen der Schätzungen $\hat{\theta}_1, \ldots, \hat{\theta}_g$ in log-Likelihood ℓ liefert konzentrierte log-Likelihood Funktion oder auch profile likelihood

$$\ell^* = \sum_{r=1}^{g} \sum_{\mathbf{x}_i \in C_r} \log(f(\mathbf{x}_i|\hat{\theta}_r)) \,,$$

- Hängt nur von $k_1, \ldots, k_n$ bzw. $C_1, \ldots, C_g$

Bestimmung der optimalen Partition

- Liegt optimale Partition $\mathbb{C}_{\text{opt}} = \left\{ C_1^{(\text{opt})}, \ldots, C_g^{(\text{opt})} \right\}$ mit Klassenzugehörigkeiten $\hat{k}_1, \ldots, \hat{k}_n$ und Parameterwerten $\hat{\theta}_1, \ldots, \hat{\theta}_g$ vor, so gilt

$$f(\mathbf{x}_i | \hat{\theta}_{\hat{k}_i}) \geq f(\mathbf{x}_i | \hat{\theta}_r) \,,$$

  falls der Tausch des Objekts i in die Klasse r wieder eine zulässige Partition ergibt

Numerische Lösung durch Austauschverfahren

1. Wähle Startpartition $k_1, \ldots, k_n$ und berechne $\hat{\theta}_1, \ldots, \hat{\theta}_g$ durch Maximierung von $\ell_1, \ldots, \ell_g$.
2. Für alle $i = 1, \ldots, n$ und $r = 1, \ldots, g$:
   Falls $f(\mathbf{x}_i|\hat{\theta}_r) > f(\mathbf{x}_i|\hat{\theta}_{k_i}),\ k_i \neq r$, ändere die Zuteilung k_i zu r und berechne $\hat{\theta}_1, \ldots, \hat{\theta}_g$ neu.
3. Wiederhole Schritt 2 bis $f(\mathbf{x}_i|\hat{\theta}_{k_i}) \geq f(\mathbf{x}_i|\hat{\theta}_r)$ für alle $\mathbf{x}_i$ und alle Cluster r.

Multivariate Normalverteilung

- Dichte der p-dimensionalen Normalverteilung, d.h. $\mathbf{X}_r \sim N_p(\boldsymbol{\mu}_r, \sigma^2 \boldsymbol{I})$:

$$f(\mathbf{x}_i | \boldsymbol{\mu}_r, \sigma^2 \boldsymbol{I}) = (2\pi\sigma^2)^{-p/2} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{x}_i - \boldsymbol{\mu}_r)^\top (\mathbf{x}_i - \boldsymbol{\mu}_r) \right\},$$

- Log-Likelihood:

$$\ell = -\frac{np}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{r=1}^{g} \sum_{\mathbf{x}_i \in C_r} (\mathbf{x}_i - \boldsymbol{\mu}_r)^\top (\mathbf{x}_i - \boldsymbol{\mu}_r)$$

Multivariate Normalverteilung

- Schätzer für feste Partition $\mathbb{C} = \{C_1, \ldots, C_g\}$:

$$\hat{\boldsymbol{\mu}}_r = \bar{\mathbf{x}}_r = \frac{1}{n_r} \sum_{\mathbf{x}_i \in C_r} \mathbf{x}_i, \quad r = 1, \ldots, g\,,$$

$$\hat{\sigma}^2 = \frac{1}{np} \sum_{r=1}^{g} \sum_{\mathbf{x}_i \in C_r} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)^\top (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)$$

- Profile likelihood:

$$\ell^* = -\frac{np}{2} \ln \left(\frac{2\pi}{np} \sum_{r=1}^{g} \sum_{\mathbf{x}_i \in C_r} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r)^\top (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_r) \right) - \frac{np}{2}$$

Zusammenhang zum Varianzkriterium

- Äquivalenz aus profile likelihood:

$$\ell^* \to \max \quad \Leftrightarrow \quad \hat{\sigma}^2 \to \min \ .$$

- Entspricht bis auf konstanten Faktor $\frac{1}{np}$ Varianzkriterium

$\to$ Varianzkriterium ist stochastisch motiviert als Mischung spezieller multivariater Normalverteilungen

Weitere Spezialfälle

1. $\mathbf{X}_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$:
   Maximierung von $\ell^* \Leftrightarrow$ Determinantenkriterium
2. $\mathbf{X}_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$:
   Maximierung von $\ell^* \Leftrightarrow$ verallgemeinertes Determinantenkriterium