Clusteranalyse

Aufgabe 1: Clusteranalyse mit k-Means

In der folgenden Tabelle seien der *weight index* und der *pH index* von drei Medikamenten gegeben:

Objekt	weight index	pH index
Medizin A	1	1
Medizin B	2	1
Medizin C	4	3

a) Gruppieren Sie die Medikamente mithilfe des k-means Clusterverfahrens in zwei Cluster. Medikamente A und B sollen dabei die Startpartition für je eine Klasse sein. Als Distanzmaß soll die euklidische Distanz verwendet werden.

b) Gehen Sie nun davon aus, dass drei zusätzliche Medikamente in der Clusteranalyse miteinbezogen werden:

Objekt	weight index	pH index
Medizin D	5	4
Medizin E	3	2
Medizin F	6	4

Dabei ergeben sich für das k-means Verfahren sowie für das Single Linkage Verfahren folgende Clusterungen

Medikament	A	B	C	D	E	F
k-means	2	2	1	1	2	1
Single Linkage	1	1	1	2	1	2

Geben Sie eine geeignete Maßzahl für die Ähnlichkeit der beiden Clusterungen an und berechnen Sie diese.

c) **Zum Selbststudium:** Bestimmen Sie die Cluster, die in Teilaufgabe b) für das k-means Verfahren sowie das Single Linkage Verfahren angegeben sind, sowohl von Hand als auch mit R.

Lösung:

a) Gesucht: 2 Klassen / Cluster mit optimaler Zuordnung der Medikamente.

Distanz: euklidische Distanz

(1) Iteration 0: $\boldsymbol{x}_1^{(0)} = (1,1)^T,\ \boldsymbol{x}_2^{(0)} = (2,1)^T$

$$\boldsymbol{D}^{(0)} = \begin{pmatrix} \textcircled{0} & 1 & 3,61 \\ 1 & \textcircled{0} & \textcircled{2,83} \end{pmatrix} \begin{matrix} \\ \rightarrow \text{ Distanz zu } \boldsymbol{x}_1^{(0)} \\ \rightarrow \text{ Distanz zu } \boldsymbol{x}_2^{(0)} \end{matrix}$$

$A \quad B \quad C$ (umkringelt $\widehat{=}$ minimales Element pro Spalte)

z.B. $D(C, \boldsymbol{x}_1^{(0)}) = \sqrt{(4-1)^2 + (3-1)^2} = 3,61$

$\Rightarrow \mathcal{C}^{(0)} = \{\{A\}, \{B, C\}\}$

(2) Iteration 1:

$$\boldsymbol{x}_1^{(1)} = (1,1)^T \qquad \{A\}$$
$$\boldsymbol{x}_2^{(1)} = \left(\tfrac{2+4}{2}, \quad \tfrac{1+3}{2}\right)^T = \left(3, \quad 2\right)^T \qquad \{B, C\}$$

$$\boldsymbol{D}^{(1)} = \begin{pmatrix} \textcircled{0} & \textcircled{1} & 3,61 \\ 2,24 & 1,41 & \textcircled{1,41} \end{pmatrix} \begin{matrix} \\ \rightarrow \text{ Distanz zu } \boldsymbol{x}_1^{(1)} \\ \rightarrow \text{ Distanz zu } \boldsymbol{x}_2^{(1)} \end{matrix}$$

$A \quad B \quad C$ (umkringelt $\widehat{=}$ minimales Element pro Spalte)

z.B. $D(C, \boldsymbol{x}_2^{(1)}) = \sqrt{(4-3)^2 + (3-2)^2} = 1,41$

$\Rightarrow \mathcal{C}^{(1)} = \{\{A, B\}, \{C\}\}$

(3) Iteration 2:

$$\boldsymbol{x}_1^{(2)} = \left(\tfrac{1+2}{2}, \quad \tfrac{1+1}{2}\right)^T = \left(\tfrac{3}{2}, \quad 1\right)^T \qquad \{A, B\}$$
$$\boldsymbol{x}_2^{(2)} = \left(4, \quad 3\right)^T \qquad \{C\}$$

$$\boldsymbol{D}^{(2)} = \begin{pmatrix} \textcircled{0,5} & \textcircled{0,5} & 3,20 \\ 3,61 & 2,83 & \textcircled{0} \end{pmatrix} \begin{matrix} \\ \rightarrow \text{ Distanz zu } \boldsymbol{x}_1^{(2)} \\ \rightarrow \text{ Distanz zu } \boldsymbol{x}_2^{(2)} \end{matrix}$$

$A \quad B \quad C$ (umkringelt $\widehat{=}$ minimales Element pro Spalte)

$\Rightarrow \mathcal{C}^{(2)} = \mathcal{C}^{(1)} \Rightarrow$ Ende

Wiederholung:

(Adjustierter) Rand-Index:

Ziel: Quantifizierung der Ähnlichkeit der Ergebnisse zweier Clusterungen

Grundlage: Aufstellen einer Vierfeldertafel bzgl. der Einordnung aller Objekt**paare**

		Clusterung B	
		1	0
Clusterung A	1	a	b
	0	c	d

- a entspricht der Anzahl aller Objektpaare, die sowohl für Clusterung A als auch für Clusterung B im gleichen Cluster sind.
- b entspricht der Anzahl aller Objektpaare, die für Clusterung A im gleichen Cluster aber für Clusterung B in verschiedenen Clustern sind.
- c entspricht der Anzahl aller Objektpaare, die für Clusterung A in verschiedenen Clustern sind aber für Clusterung B gleichen Cluster sind.
- d entspricht der Anzahl aller Objektpaare, die sowohl für Clusterung A als auch für Clusterung B in zwei verschiedenen Clustern sind

Dieses Verfahren kann auf beliebige Clusteranzahlen verallgemeinert werden. Dabei brauchen die Clusterungen A und B *nicht* die gleiche Clusteranzahl aufweisen.

Rand-Index: $R = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}}$

Problematik: Zufällige Übereinstimmungen

$\Rightarrow$ **Adjustierter Rand-Index** korrigiert dafür: $R_{adj} = \frac{\binom{n}{2}(a+d)-[(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2-[(a+b)(a+c)+(c+d)(b+d)]}$

(Quelle: Santos & Embrechts (2009))

Invertiere zunächst eine der Clusterungen $\rightarrow$ größere Übereinstimmung der Cluster $\rightarrow$ bessere Übersichtlichkeit:

Medikament	A	B	C	D	E	F
k-means	1	1	2	2	1	2
Single Linkage	1	1	1	2	1	2

Vierfeldertafeln:

Vergleiche zunächst Medikament A mit Medikament B – F:

	1	0
1	2	–
0	1	2

Medikament B verglichen mit Medikament C – F:

	1	0
1	1	–
0	1	2

...

Insgesamt ergibt sich:

		Single Linkage: 1	Single Linkage: 0
k-means	1	4	2
	0	3	6

$\Rightarrow$ Adj. Rand-Index:

$$R_{adj} = \frac{\binom{6}{2}(4+6) - [(4+2)(4+3) + (3+6)(2+6)]}{\binom{6}{2}^2 - [(4+2)(4+3) + (3+6)(2+6)]}$$

$$= \frac{15 \cdot 10 - [6 \cdot 7 + 9 \cdot 8]}{225 - [6 \cdot 7 + 9 \cdot 8]}$$

$$= \frac{150 - [42 + 72]}{225 - [42 + 72]} = \frac{36}{111} = 0,3243$$

Zusätzliche Ergänzung: nicht-adj. Rand-Index:

$$R = \frac{4+6}{\binom{6}{2}} = \frac{10}{15} \approx 0,67$$

$\rightarrow$ Die Clusterungen stimmen nur mäßig überein. Zwar wurde nur Medikament C anders eingeteilt, bei insgesamt lediglich 6 Medikamenten scheint dies aber bereits einen großen Unterschied zu machen (bereits viele Übereinstimmungen durch Zufälligkeiten zu erwarten).

Aufgabe 2: Modellbasiertes Clustering

Gehen Sie davon aus, dass Sie eine modellbasierte Clusteranalyse auf bivariaten Daten durchführen. Dazu liegt Ihnen das Ergebnis eines entsprechenden Ansatzes in Form der resultierenden Mischverteilung vor. Konkret sei diese für eine Clusteranzahl von 2 gegeben durch

$$\hat{f}(\boldsymbol{x}) = \sum_{k=1}^{2} \hat{p}(k) f(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k),$$

wobei f der Dichte einer bivariaten Normalverteilung entspricht mit den Parametern

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} -2 \\ 2 \end{pmatrix}, \; \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} \sigma^2_{1;\,1} & \sigma_{12;\,1} \\ \sigma_{12;\,1} & \sigma^2_{2;\,1} \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \; \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} \sigma^2_{1;\,2} & \sigma_{12;\,2} \\ \sigma_{12;\,2} & \sigma^2_{2;\,2} \end{pmatrix}$$

und mit $\hat{p}(1) = \hat{p}(2) = 0,5$. Außerdem seien drei in dem Datensatz enthaltene Datenpunkte konkret angegeben mit $\boldsymbol{x}_1 = (0,\, 0)^T$, $\boldsymbol{x}_2 = (-1,\, 1)^T$ und $\boldsymbol{x}_3 = (1,\, 3)^T$ (welche nach Annahme ebenfalls für die Berechnung der Mischverteilung berücksichtigt wurden).

a) Machen Sie sich mit dem Konzept des (computergestützten) modellbasierten Clusterns vertraut. Betrachten Sie hierbei auch die `R`-Funktion `Mclust()` aus dem Paket `mclust`. Lesen Sie sich dazu insbesondere die Hilfe zu `mclustModelNames` durch und machen Sie sich die einzelnen Modelle verständlich. Wie erhält man generell aus der Mischverteilung eine Partitionierung der Daten?

b) Welchen der beiden Cluster werden die Datenpunkte $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ und $\boldsymbol{x}_3$ für das Modell *spherical, equal volume* zugeordnet? Veranschaulichen Sie dies in einer Skizze.

c) Überlegen Sie sich jeweils konkrete Kovarianzmatrizen $\hat{\boldsymbol{\Sigma}}_1$ und $\hat{\boldsymbol{\Sigma}}_2$, für die unter den folgenden Modellen die jeweils vorgegebene Zuordnung getroffen wird:

	Modell	Zuordnung
i)	*spherical, unequal volume*	$\boldsymbol{x}_1, \boldsymbol{x}_2 \longmapsto C_1, \boldsymbol{x}_3 \longmapsto C_2$
ii)	*diagonal, equal volume and shape*	$\boldsymbol{x}_2, \boldsymbol{x}_3 \longmapsto C_1, \boldsymbol{x}_1 \longmapsto C_2$
iii)	*ellipsoidal, varying volume, shape and orientation*	$\boldsymbol{x}_3 \longmapsto C_1, \boldsymbol{x}_1, \boldsymbol{x}_2 \longmapsto C_2$

Hinweise:

i) Die Dichte der multivariaten Normalverteilung lässt sich mittels der Funktion `dmvnorm()` aus dem Paket `mvtnorm` berechnen.

ii) In dem `R`-Code zu Aufgabe 3 (b) auf dem 2. Übungsblatt können Sie sich noch einmal die Zusammenhänge zwischen den Parametern und der Form der multivariaten Normalverteilung veranschaulichen.

a) Details

The following models are available in package mclust:

univariate mixture

"E" = equal variance (one-dimensional)

"V" = variable variance (one-dimensional)

multivariate mixture

"EII" = spherical, equal volume

"VII" = spherical, unequal volume

"EEI" = diagonal, equal volume and shape

"VEI" = diagonal, varying volume, equal shape

"EVI" = diagonal, equal volume, varying shape

"VVI" = diagonal, varying volume and shape

"EEE" = ellipsoidal, equal volume, shape, and orientation

"EVE" = ellipsoidal, equal volume and orientation

"VEE" = ellipsoidal, equal shape and orientation

"VVE" = ellipsoidal, equal orientation

"EEV" = ellipsoidal, equal volume and equal shape

"VEV" = ellipsoidal, equal shape

"EVV" = ellipsoidal, equal volume

"VVV" = ellipsoidal, varying volume, shape, and orientation

single component

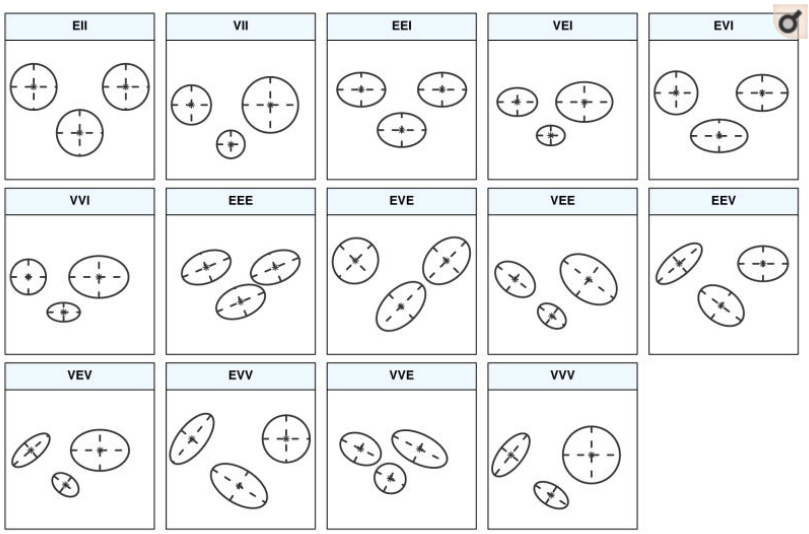
"X" = univariate normal

"XII" = spherical multivariate normal

"XXI" = diagonal multivariate normal

"XXX" = ellipsoidal multivariate normal

- Die Gruppeneinteilung erfolgt anhand dessen, wie plausibel es für das jeweilige Objekt ist einer bestimmten Gruppe anzugehören.
- Kennzahl: Posteriori-Dichte
- Diese ergibt sich für das Objekt i mit Merkmalsvektor x_i aus dem Produkt der Priori-Wahrscheinlichkeit und der Dichte der jeweiligen Gruppenverteilung im Punkt x_i.
- Ein Objekt wird derjenigen Gruppe zugeordnet, für die es die höchste Posteriori-Dichte aufweist.



Ellipses of isodensity for each of the 14 Gaussian models obtained by eigen-decomposition in case of three groups in two dimensions.

Quelle: mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. Scrucca, L. et al.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/

b) *spherical, equal volume* $\Rightarrow$ $\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$

Damit gilt:

$$\det\left(\hat{\boldsymbol{\Sigma}}_k\right) = \alpha^2 \qquad \text{und} \qquad \hat{\boldsymbol{\Sigma}}_k^{-1} = \begin{pmatrix} 1/\alpha & 0 \\ 0 & 1/\alpha \end{pmatrix}$$

Dichtefunktion:

$$f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) = \frac{1}{\sqrt{(2\,\pi)^2 \det\left(\hat{\boldsymbol{\Sigma}}_k\right)}} \exp\left\{-\frac{1}{2}\left((\boldsymbol{x}-\hat{\boldsymbol{\mu}}_k)^\top \hat{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_k)\right)\right\}$$

$$= \frac{1}{2\,\pi\alpha} \exp\left\{-\frac{1}{2\,\alpha}\left(\underbrace{(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_k)^\top(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_k)}_{=||\boldsymbol{x}-\hat{\boldsymbol{\mu}}_k||^2}\right)\right\}$$

Somit:

$$\hat{p}(1) \cdot f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \lessgtr \hat{p}(2) \cdot f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2)$$

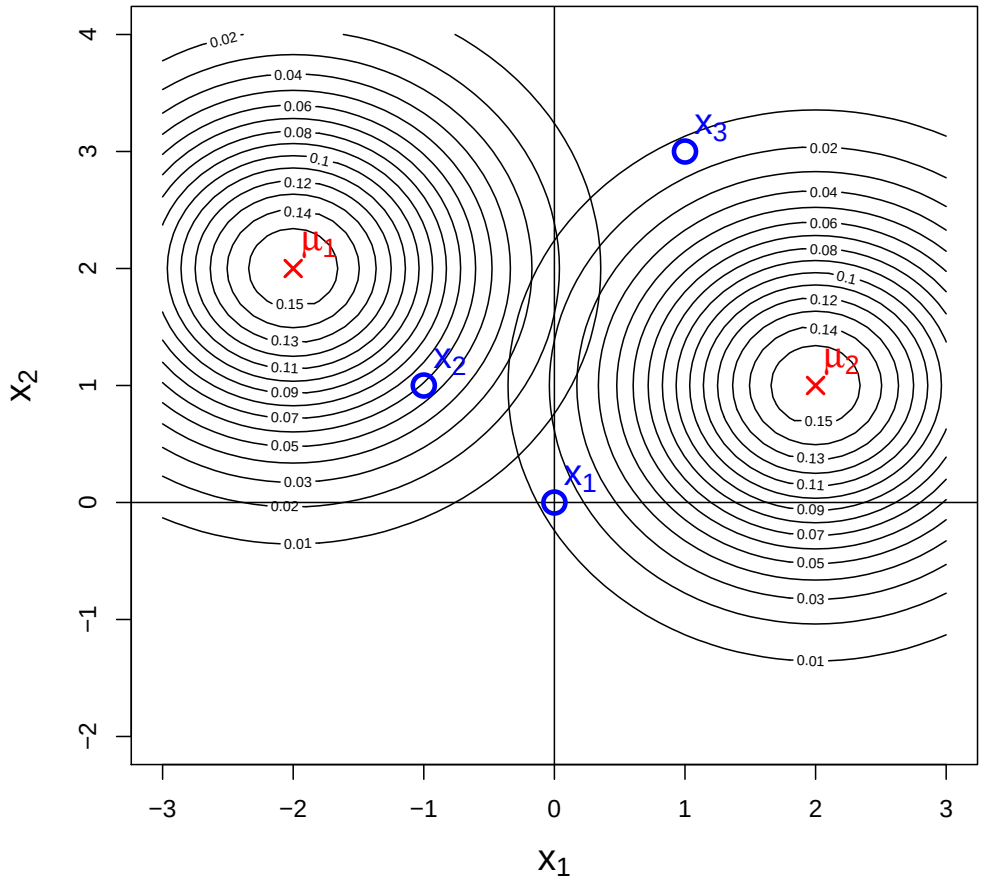
$$\overset{\hat{p}(1)=\hat{p}(2)}{\Longleftrightarrow} \qquad f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \lessgtr f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2)$$

$$\Longleftrightarrow \qquad ||\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1||^2 \gtrless ||\boldsymbol{x}-\hat{\boldsymbol{\mu}}_2||^2$$

Setze $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ ein: z.B.: $\left\| \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -2 \\ 2 \end{pmatrix} \right\|^2 = 2^2 + (-2)^2 = 8.$

	Distanzen			
	$\boldsymbol{\mu}_1$		$\boldsymbol{\mu}_2$	$\Rightarrow$ Cluster
$\boldsymbol{x}_1$	8	$>$	5	2
$\boldsymbol{x}_2$	2	$<$	9	1
$\boldsymbol{x}_3$	10	$>$	5	2

Abbildung des Modells *spherical, equal volume*:

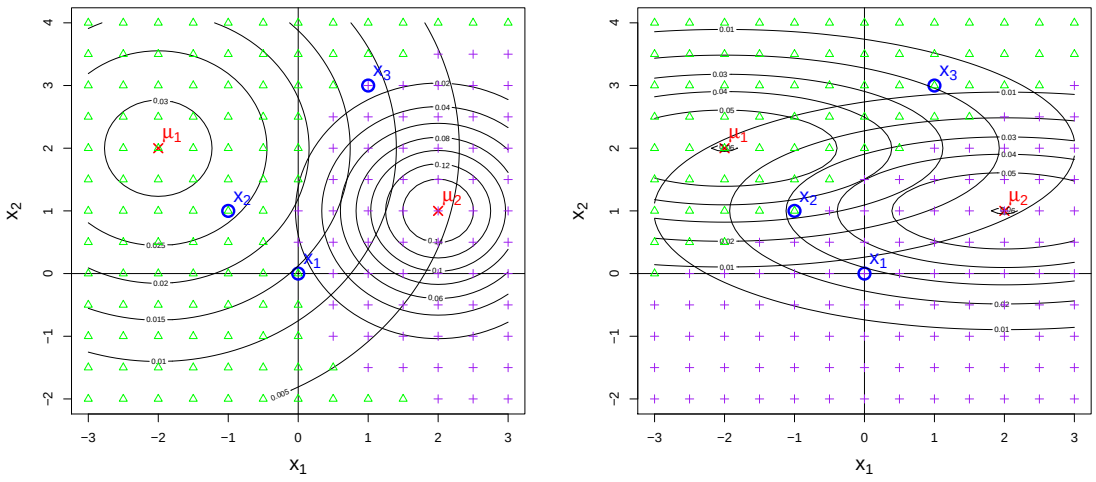


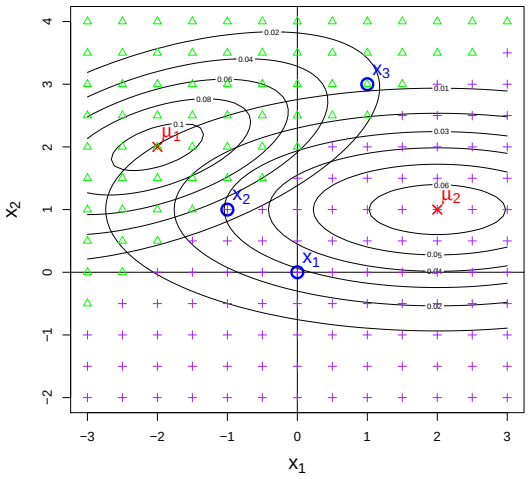
c) Mögliche Matrizen:

(i) $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$ und $\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

(ii) $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 7 & 0 \\ 0 & 1 \end{pmatrix}$ und $\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 7 & 0 \\ 0 & 1 \end{pmatrix}$

(iii) $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 3 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ und $\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 6 & 0 \\ 0 & 1 \end{pmatrix}$





- Grüne Dreiecke = Posteriori-Dichte des Clusters 1 ist größer
- Violette Kreuze = Posteriori-Dichte des Clusters 2 ist größer

Aufgabe 3: Clusteranalyse in R II

Der Datensatz `geyser` aus dem R-Paket `MASS` beinhaltet für 299 Eruptionen des berühmten *Old Faithful* Geysirs im Yellowstone Nationalpark die Wartezeit seit der vorangegangenen Eruption (in Minuten) sowie die Eruptionsdauer (in Minuten). Im Folgenden soll dieser Datensatz mit Hilfe eines Mischmodellansatzes untersucht werden. Für die Verteilung in den Klassen wird eine bivariate Normalverteilung angenommen.

a) Skizzieren Sie kurz die verwendeten Modellannahmen des hier betrachteten Mischmodellansatzes. Wie kann das Mischmodell geschätzt werden?

b) Plotten Sie die Daten und bestimmen Sie visuell eine geeignete Anzahl an Klassen für den Mischmodellansatz.

c) Führen Sie das modellbasierte Clustering durch:

i) Clustern Sie die Geysir-Daten mit Hilfe der Funktion `Mclust()` aus dem Package `mclust`. Verwenden Sie unterschiedliche Annahmen (z.B. „EII“, „VVI“, „VVV“) für die Kovarianzstruktur in den Klassen. Vergleichen Sie die sich daraus ergebenden Modelle und evaluieren Sie die Güte der finalen Partitionierung.

ii) Nutzen Sie die `Mclust()` Funktion, um unter allen möglichen modellbasierten Clusterverfahren für die aktuellsten Daten die bzgl. dem BIC optimalsten Annahmen und Anzahl der Klassen zu bestimmen.

d) Möglichkeiten zur Evaluierung der Güte einzelner Clusterungen sowie zum Vergleich verschiedener Clustering-Ergebnisse:

i) Schränken Sie den `geyser`-Datensatz für eine übersichtlichere Darstellung auf die ersten 50 Zeilen ein und führen Sie ein k-Means Clustering mit 4 Clustern auf dem reduzierten Datensatz durch.

ii) Evaluieren Sie die Ergebnisse des k-Means Clusterings auf Basis verschiedener graphischer Darstellungen:

- Silhouette-Plot (Funktion `factoextra::fviz_silhouette`)
- Stripes-Plot (`flexclust::stripes`)
- Heatmap der Distanzmatrix (`lattice::levelplot` oder `factoextra::fviz_dist`)

iii) Nutzen Sie den Rand-Index (Funktion `mclust::adjustedRandIndex`), um das k-Means Clustering mit dem in c) ii) erhaltenen BIC-optimalen modellbasierten Clustering zu vergleichen. Schätzen Sie hierfür das modellbasierte Clustering ebenfalls nur auf den ersten 50 Zeilen der Daten.

Lösung:

a)

- Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ mit $\boldsymbol{x}_i \in \mathbb{R}^p$, hier: $p = 2$
- Item-Werte bekannt, aber unbekannte Gruppenzugehörigkeit $r_1, \ldots, r_n$
- bei gegebener Gruppenzugehörigkeit ist $\boldsymbol{x}_i$ normalverteilt:

- $\boldsymbol{x}_i|r \sim \mathcal{N}(\boldsymbol{\mu}_r,\, \boldsymbol{\Sigma}_r),\ r \in \{1, \ldots, g\}$
- $f_r(\boldsymbol{x}_i) = f(\boldsymbol{x}_i|r) = f(\boldsymbol{x}_i|\boldsymbol{\mu}_r,\, \Sigma_r)$

- Priori-Wahrscheinlichkeit der Gruppenzugehörigkeit:

$$p(r),\ r \in \{1, \ldots, g\}$$

- Annahme: Mischverteilung:

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

- Posterior-Wahrscheinlichkeit für die Gruppenzugehörigkeit:

$$\hat{p}(r|\boldsymbol{x}_i) = \frac{\hat{p}(r)\hat{f}(\boldsymbol{x}_i|r)}{\hat{f}(\boldsymbol{x}_i)} =: \hat{p}_{ir} \tag{1}$$

- Gruppeneinteilung über marginale, geschätzte Posteriori-Wahrscheinlichkeit:

$$\mathcal{C}_r = \{\boldsymbol{x}_i|\hat{p}_{ir} \geq \hat{p}_{is},\ r \neq s\},\ r \in \{1, \ldots, g\}$$

- Schätzung mit *EM-Algorithmus*, Iteration bis zur Konvergenz:

  (1) E-Schritt:

  * Gegeben $\hat{p}(r),\ \hat{f}(\boldsymbol{x}|r),\ \hat{f}(\boldsymbol{x})$
  * Berechne $\hat{p}_{ir}$ gemäß (**??**)

  (2) M-Schritt:

  * Gegeben $\hat{p}_{ir}$, update $\hat{p}(r) = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ir}$
  * $(\hat{\boldsymbol{\mu}}_r,\, \hat{\boldsymbol{\Sigma}}_r) = \arg\max_{\boldsymbol{\mu}_r,\, \boldsymbol{\Sigma}_r} \sum_{i=1}^{n} \hat{p}_{ir} \cdot \log\left(f_r(\boldsymbol{x}_i|\boldsymbol{\mu}_r,\, \boldsymbol{\Sigma}_r)\right),\ r \in \{1, \ldots, g\}$ (gewichteter ML-Schätzer)

  – häufig zwecks Parameterökonomie: weniger flexible Annahmen für die Kovarianzen $\boldsymbol{\Sigma}_r$, z.B.

  * $\boldsymbol{\Sigma}_r = \sigma^2 \cdot I,\ r \in \{1, \ldots, g\}$
  * $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}$
  * $\boldsymbol{\Sigma}_r = \sigma_r^2 \cdot I,\ r \in \{1, \ldots, g\}$

b) `R`-Code

c) `R`-Code

d) `R`-Code