MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2017

Prof. E.G. Schukat-Talamazzini

Stand: 25. August 2017

Teil V

Gruppierung von Objekten

Überwachungsszenarien

Etikettierung der Lerndatenobjekte nach Klassenzugehörigkeit ?

Überwachtes Lernen

Der Lehrer stellt Zielwert **aller** Lernobjekte bereit. { Klassifikation, Vorhersage }

Halbüberwachtes Lernen

Der Lehrer verrät Zielwert **weniger** Lernobjekte. { Bootstrap, Transduktion }

Reinforcement Lernen

Der Lehrer übt **Erfolgskontrolle** („*feed-back*“). { Spielstrategie, Aktionsplanung }

Unüberwachtes Lernen

Der Lehrer stellt **keinerlei** Zielwerte bereit. { Gruppierung, Assoziation }

Gruppierung a.k.a. Clusteranalyse

Partitionierung der Datenobjekte in Ballungs- oder Häufungsgebiete

Objektrepräsentation

Vektorraum · Attribute · Metrik

Zielgröße

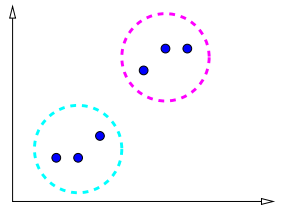
global · lokal · ad hoc

Zerlegungsstrategie

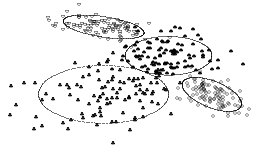
· top-down
· bottom-up
· Austausch (*K*-means, EM)
· split & merge

Gruppenrepräsentation

Mengen · Prototypen · Verteilungen
Formeln · Regeln



Gruppen ähnlicher Objekte



Unscharfe Gruppenzuordnung

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

Agglomerative Gruppierung

Generischer Bottom-up-Algorithmus

⟨Algorithmus⟩

Gegeben sind die Datenobjekte $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \boldsymbol{\Omega}$

1. INITIALISIERUNG
   Starte mit $K = T$ Gruppen $\omega_t = \{\boldsymbol{x}_t\}$, $t = 1..T$.
2. DISTANZBERECHNUNG
   Berechne für alle $1 \leq \kappa < \lambda \leq K$:
   $$D_{\kappa\lambda} \stackrel{\text{def}}{=} d(\omega_\kappa, \omega_\lambda)$$
3. VEREINIGUNG
   Vereinige die beiden Gruppen ω_{κ^*}, ω_{λ^*} mit
   $$(\kappa^*, \lambda^*) = \operatorname*{argmin}_{\lambda,\kappa} D_{\kappa\lambda}$$
4. TERMINIERUNG
   Wenn $K = 1$ dann ENDE, sonst $\rightsquigarrow$ 2.

⟨Algorithmus⟩

Mengendistanzfunktionen

$d : \mathfrak{P}\boldsymbol{\Omega} \times \mathfrak{P}\boldsymbol{\Omega} \to \mathbb{R}_0^+$

Single-Linkage

Kürzeste Brücke zwischen zwei Gruppen

$$d_{\mathrm{SL}}(A, B) \stackrel{\text{def}}{=} \min_{x \in A} \min_{y \in B} d(x, y)$$

Complete-Linkage

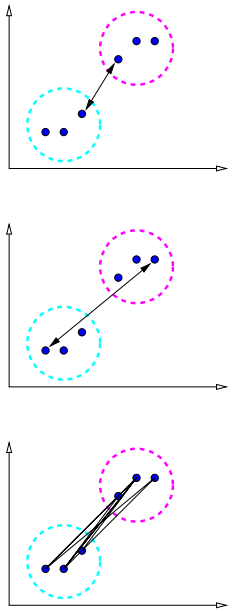
Durchmesser nach Vereinigung zweier Gruppen

$$d_{\mathrm{CL}}(A, B) \stackrel{\text{def}}{=} \max_{x \in A} \max_{y \in B} d(x, y)$$

Average-Linkage

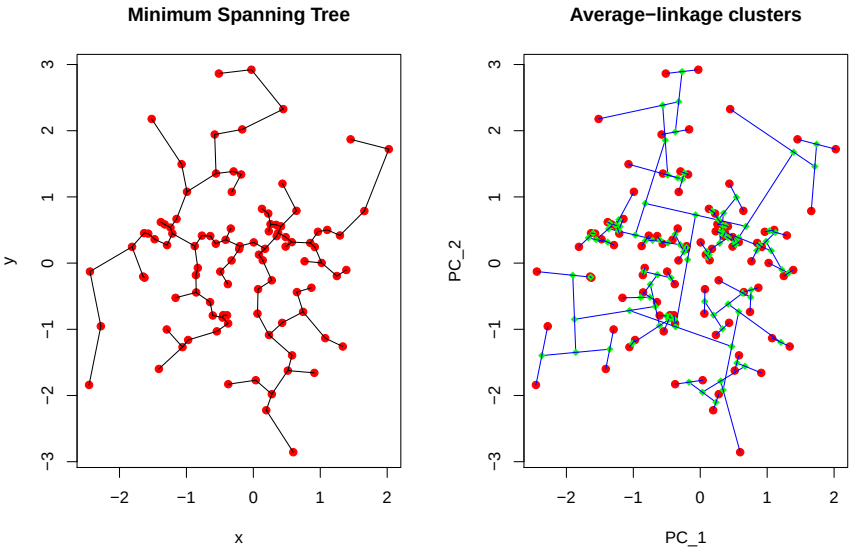
Mittlerer bipartiter Punkteabstand

$$d_{\mathrm{AL}}(A, B) \stackrel{\text{def}}{=} \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$$



Kettenbildung und Lassoeffekt

Beispiel mit $T = 100$ Objekten im $\mathbb{R}^2$



Mengendistanzfunktionen

Welche ist die beste ?

Single-Linkage

„Ketteneffekt“

- erzeugt minimalen Spannbaum
- ⊕ schwach monoton inv. monot. d-Transf.

Complete-Linkage

„Lassoeffekt“

- extrem anfällig gegen Ausreißer
- ⊕ schwach monoton inv. monot. d-Transf.

Average-Linkage

weder Ketten- noch Lassoeffekt

- bevorzugt sphärische Ballungsgebiete
- ⊕ schwach monoton Δ-invariant

Getrimmte Distanzen

Diese Effekte lassen sich abmildern, wenn in der Distanzformel jeweils $q > 1$ kleinste bzw. größte Distanzen eliminiert werden, wodurch die Einflußdramatik eventueller Ausreißer eingedämmt wird.

Lance-Williams-Rekursion

Effizient auswertbare Mengendistanzfunktionen

Definition

Gehorcht eine Distanzfunktion $d : \mathfrak{P}\Omega \times \mathfrak{P}\Omega \to \mathbb{R}$ in eindeutiger Weise dem Schema

$$\begin{aligned} d(\{\boldsymbol{x}\},\{\boldsymbol{y}\}) &= d(\boldsymbol{x},\boldsymbol{y}) \\ d(A_1 \uplus A_2, B) &= \alpha_1 \cdot d(A_1,B) + \alpha_2 \cdot d(A_2,B) + \beta \cdot d(A_1,A_2) \\ &\quad + \gamma \cdot |d(A_1,B) - d(A_2,B)| \end{aligned}$$

so heißt diese Vorschrift **Lance-Williams-Rekursion** mit den reellwertigen Parametern $\alpha_1 \geq 0$, $\alpha_2 \geq 0$, β und γ.

Bemerkung

Die drei X-Linkage-Funktionen besitzen alle die Lance-Williams-Gestalt:

1. Single-Linkage: $\alpha_1 = \alpha_2 = \frac{1}{2}$, $\beta = 0$, $\gamma = -\frac{1}{2}$
2. Complete-Linkage: $\alpha_1 = \alpha_2 = \frac{1}{2}$, $\beta = 0$, $\gamma = +\frac{1}{2}$
3. Average-Linkage: $\alpha_1 = \frac{|A_1|}{|A_1|+|A_2|}$, $\alpha_2 = \frac{|A_2|}{|A_1|+|A_2|}$, $\beta = 0$, $\gamma = 0$

Dreiecksungleichung und Monotonie

Satz (Lance & Williams, 1967)

Es sei eine rekursive Form der Mengendistanzfunktion vorausgesetzt.

1. *Gilt $\alpha_1 + \alpha_2 \geq 1$, $\beta \geq 0$ und $\gamma = 0$ und gilt die Dreiecksungleichung für alle Gruppendistanzen, so gilt sie auch noch nach der $d(\cdot,\cdot)$-optimalen Vereinigung:*

$$d(A_1 \uplus A_2, B) \;+\; d(A_1 \uplus A_2, C) \;\geq\; d(B,C)$$

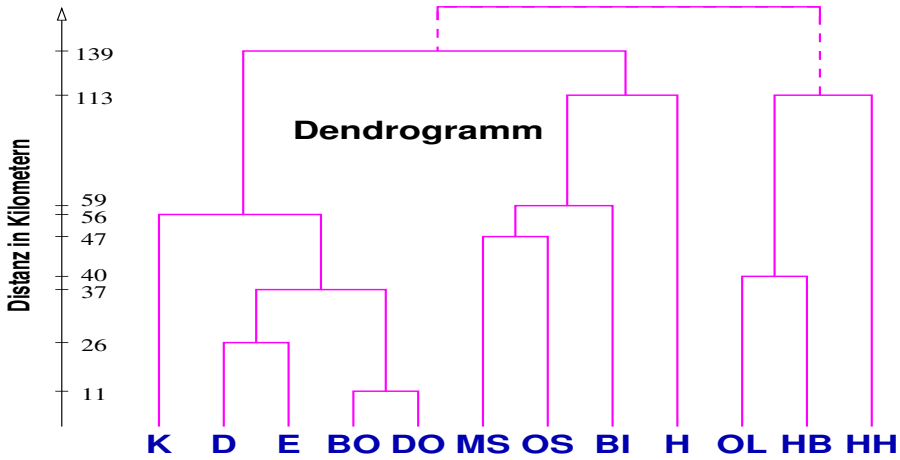
2. *Gilt $\alpha_1 + \alpha_2 + \beta \geq 1$ und $\gamma = 0$, so steigen die Gruppendistanzen schwach monoton an:*

$$d(A_1 \uplus A_2, B) \;\geq\; d(A_1, A_2)$$

3. *Erfüllt das Agglomerationsverfahren sogar die strenge Monotonie, so gelten für jede intermediäre Gruppenstruktur (mit Gruppe A) die **ultrametrischen** Ungleichungen:*

$$\forall \boldsymbol{x},\boldsymbol{y} \in A,\ \forall \boldsymbol{z} \notin A: \qquad d(\boldsymbol{x},\boldsymbol{y}) \;<\; d(\boldsymbol{x},\boldsymbol{z})$$

Beispiel — Dendrogramm für Städtedistanzen



Strenge Monotonie

Je später zwei Gruppen im agglomerativen Clusteralgorithmus vereinigt werden, desto größer ist ihre Mengendistanz.
(Nichtmonotonie ⇝ Inversionen des Dendrogramms)

Weitere Distanzfunktionen

Simple-Average

Keine globale Semantik, aber schwach monoton und Δ-invariant:

$$d_{\mathrm{SA}}(A_1 \uplus A_2, B) \stackrel{\mathrm{def}}{=} \frac{1}{2} \cdot d_{\mathrm{SA}}(A_1, B) + \frac{1}{2} \cdot d_{\mathrm{SA}}(A_2, B)$$

Lance-Williams-Parameter: $\alpha_1 = \alpha_2 = 1/2$, $\beta = 0$, $\gamma = 0$

Zentroid-Verfahren

Für numerische Attribute; weder schwach monoton noch Δ-invariant:

$$d_{\mathrm{ZEN}}(A, B) \stackrel{\mathrm{def}}{=} \|\mu(A) - \mu(B)\|^2$$

Lance-Williams-Parameter: $\alpha_1 = \frac{|A_1|}{|A_1|+|A_2|}$, $\alpha_2 = \frac{|A_2|}{|A_1|+|A_2|}$, $\beta = -\alpha_1\alpha_2$, $\gamma = 0$

Median/Gower-Verfahren

Wie Zentroid; ignoriert aber die relativen Größen vereinigter Gruppen:

$$\alpha_1 = \alpha_2 = 1/2\,, \qquad \beta = -1/4\,, \qquad \gamma = 0$$

Ward-Verfahren

Ähnelt der Zentroiddistanz · Garantiert aber Distanzmonotonie

Ward-Zielgröße

Das globale Clusterverzerrungsmaß

$$\varepsilon_{\mathrm{WARD}}(\{\omega_1, \ldots, \omega_K\}) \stackrel{\mathrm{def}}{=} \sum_{\lambda=1}^{K} \sum_{\boldsymbol{x} \in \omega_\lambda} \|\boldsymbol{x} - \boldsymbol{\mu}_\lambda\|^2\,, \qquad \boldsymbol{\mu}_\lambda = \mu(\omega_\lambda)$$

führt auf den **Heterogenitätszuwachs**

$$d_{\mathrm{WARD}}(A, B) = \varepsilon' - \varepsilon = \frac{|A| \cdot |B|}{|A| + |B|} \cdot \|\mu(A) - \mu(B)\|^2$$

bei Vereinigung der Gruppen $A = \omega_\kappa$ und $B = \omega_\lambda$ und diese Formel wiederum auf eine Lance-Williams-Darstellung:

$$d_{\mathrm{WARD}}(A_1{+}A_2, B) = \frac{(|A_1| + |B|) \cdot d(A_1, B) + (|A_2| + |B|) \cdot d(A_2, B) - |B| \cdot d(A_1, A_2)}{|A_1| + |A_2| + |B|}$$

Divisive Gruppierung

Generischer Top-down-Algorithmus

⟨Algorithmus⟩

Gegeben sind die Datenobjekte $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in \boldsymbol{\Omega}$.

1. INITIALISIERUNG
   Starte mit $K = 1$ Gruppe(n) $\omega_1 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$.
2. HETEROGENITÄTSKRITERIUM
   Berechne für alle $1 \le \kappa \le K$ die Gruppenheterogenität
   $$H_\kappa \stackrel{\mathrm{def}}{=} d(\omega_\kappa)\,.$$
3. AUFSPALTUNG
   Zerlege diejenige Gruppe ω_{κ^*} mit
   $$\kappa^* = \operatorname*{argmax}_\kappa H_\kappa$$
   in zwei disjunkte Teilgruppen (z.B. via Austauschverfahren).
4. TERMINIERUNG
   Wenn $K = T$, dann Ende, sonst $\rightsquigarrow$ 2.

⟨Algorithmus⟩

Heterogenitätskriterien

Gruppendurchmesser

$$d_{\mathrm{DIAM}}(\omega) \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{x}, \boldsymbol{y} \in \omega} d(\boldsymbol{x}, \boldsymbol{y})$$

Mittlere Innergruppenspanne

$$d_{\mathrm{AD}}(\omega) \stackrel{\mathrm{def}}{=} \frac{1}{|\omega|^2 - |\omega|} \sum_{\boldsymbol{x}, \boldsymbol{y} \in \omega} d(\boldsymbol{x}, \boldsymbol{y})$$

Empirische Gruppenvarianz

$$d_{\mathrm{VAR}}(\omega) \stackrel{\mathrm{def}}{=} \frac{1}{|\omega|} \sum_{\boldsymbol{x} \in \omega} \|\boldsymbol{x} - \boldsymbol{\mu}(\omega)\|^2 = \mathrm{spur}\,(\boldsymbol{S}(\omega))$$

Gaußäquivalente Entropie

$$d_{\mathrm{CE}}(\omega) \stackrel{\mathrm{def}}{=} -\frac{2}{|\omega|} \cdot \log \mathcal{N}(\omega \mid \boldsymbol{\mu}(\omega), \boldsymbol{S}(\omega)) = \mathrm{const} + \log\det \boldsymbol{S}(\omega)$$

Hierarchische Gruppierung

Divisive Gruppierung $\hat{=}$ Top-down-Induktion

Kontrollflußregelung durch Heterogenitätsmaß; $\mathrm{O}(T \cdot n_{\text{split}})$
Polythetische Verzweigungsfragen · extensionale Zerlegung

Blinde Gruppierung

Keine Heterogenitätsprüfung
Balancierte Aufspaltung in 2^b Gruppen

Agglomerative Gruppierung $\hat{=}$ Bottom-up-Iteration

Gierige Verschmelzung mit Aufwand $\mathrm{O}(T \cdot T^2)$

ISODATA-Algorithmus

„*Split+merge*"-Strategie
Pulsierende Folge von Teilungen & Verschmelzungen

Welches ist die beste Gruppierungsstufe ?

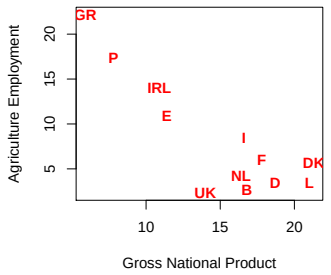
Wähle die „*richtige*" Clusteranzahl $K \in \{1, 2, \ldots, T\}$

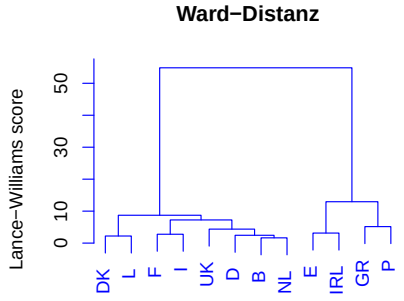
Beispiel — Agrarnationen der EU (1993)

Datensatz 'agriculture' (cluster)

12 europäische Länder
Attribut x_1 = Bruttosozialprodukt der Hauptstadt
Attribut x_2 = Bevölkerungsanteil (%) in landwirtschaftlicher Anstellung

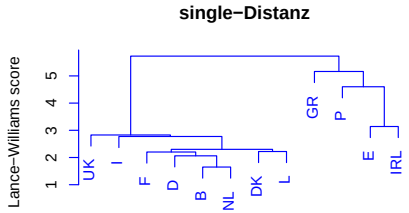
	B	DK	D	GR	E	F	IRL	I	L	NL	P	UK
x_1	16.8	21.3	18.7	5.9	11.4	17.8	10.9	16.6	21.0	16.4	7.8	14.0
x_2	2.7	5.7	3.5	22.2	10.9	6.0	14.0	8.5	3.5	4.3	17.4	2.3

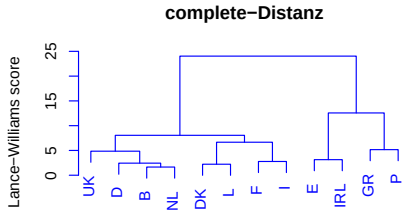


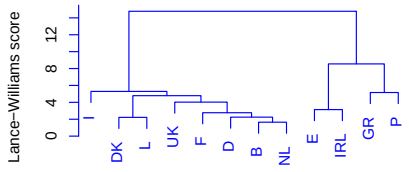


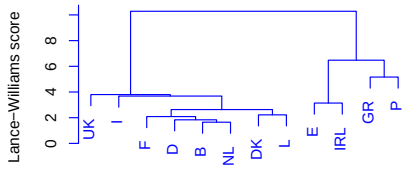
Beispiel — Agrarnationen der EU (1993)

Vergleich unterschiedlicher Lance-Williams-Distanzen



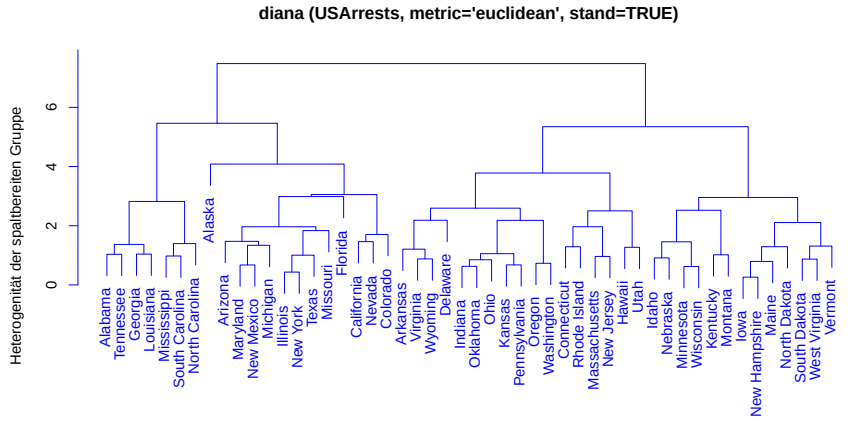






Beispiel — Verbrechensstatistik

Divisive Gruppierung mit 'diana'/R



Datensatz 'USArrests' (datasets)

50 Objekte: Kriminalstatistiken aller US-Bundesstaaten (1973)
3 Attribute: „*Murder*", „*Assault*", „*Rape*" (Anzahl je 10^5 Einwohner) und
1 Attribut: „*UrbanPop*" (Prozentsatz Stadtbevölkerung)

Scharfe Gruppierung

bei vorgegebener Gruppenanzahl $K \in \mathbb{N}$

GESUCHT

ist eine K-Partition des Datensatzes $\omega \subset \mathbf{\Omega}$.

- **extensional:** Teilmengensystem $\omega_1 \uplus \omega_2 \uplus \ldots \uplus \omega_K = \omega$
- **intensional:** Gruppenprototypen $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_K \in \mathbf{\Omega}$

Verzerrung einer Gruppe

hinsichtlich einer Objektraummetrik $d : \mathbf{\Omega} \times \mathbf{\Omega} \to \mathbb{R}$:

$$\varepsilon(\omega_\kappa) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x} \in \omega_\kappa} d(\boldsymbol{x}, \boldsymbol{\mu}(\omega_\kappa)) \ , \qquad \kappa = 1, \ldots, K$$

Verzerrung einer Partition

$$\varepsilon(\{\omega_1, \ldots, \omega_K\}) \stackrel{\text{def}}{=} \sum_{\kappa=1}^{K} \varepsilon(\omega_\kappa)$$

⇨ Kombinatorische Optimierungsaufgabe

Permutationsverfahren

Gieriges Suchverfahren · extensional · alle Metriken

⟨Algorithmus⟩

1. INITIALISIERUNG
   Wähle eine Startpartition $\omega_1, \ldots, \omega_K$ mit vorgegebenen $|\omega_\kappa| = T_\kappa$.
2. VERZERRUNGSDIFFERENZEN
   Berechne für alle $\boldsymbol{x} \in \omega_\kappa$ und $\boldsymbol{y} \in \omega_\lambda$ mit $\kappa \neq \lambda$

$$\Delta\varepsilon(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\text{def}}{=} \varepsilon(\{\ldots, \underbrace{\omega'_\kappa, \omega'_\lambda}_{\boldsymbol{x} \Leftrightarrow \boldsymbol{y}}, \ldots\}) \ - \ \varepsilon(\{\omega_1, \ldots, \omega_K\}) \ .$$

3. VERTAUSCHUNG
   Vertausche innerhalb der aktuellen Partition das Datenvektorpaar

$$(\boldsymbol{x}^*, \boldsymbol{y}^*) \ = \ \operatorname{argmin} \Delta\varepsilon(\boldsymbol{x}, \boldsymbol{y}) \ .$$

4. TERMINIERUNG
   Wenn $\varepsilon \leq \theta$ dann Ende sonst $\rightsquigarrow$ 2.

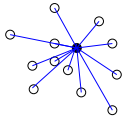
⟨Algorithmus⟩

Intensionale Gruppierung

Gruppierung mit Prototypen — *„Vektorquantisierung“*

Lemma

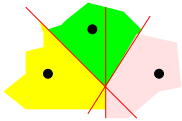
Es sei $\omega_1, \ldots, \omega_K$ eine Gruppierung der Elemente $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ des metrischen Raumes $(\mathbf{\Omega}, d)$, welche die globale Verzerrung minimiert. Dann gibt es ***Gruppenprototypen*** *$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \in \mathbf{\Omega}$ mit*



Klassenzentroid

1. *Jedes $\boldsymbol{z}_\kappa$ ist Zentroid seiner Gruppe ω_κ:*

$$\boldsymbol{z}_\kappa \ = \ \operatorname*{argmin}_{\boldsymbol{y} \in \mathbf{\Omega}} \sum_{\boldsymbol{x} \in \omega_\kappa} d(\boldsymbol{x}, \boldsymbol{y})$$



Voronoi–Partition

2. *Jeder Datenvektor $\boldsymbol{x}_t$, $t = 1, \ldots, T$ gehört zu der Gruppe des nächstliegenden Prototypen:*

$$\boldsymbol{x}_t \in \omega_\kappa \quad \Rightarrow \quad d(\boldsymbol{x}_t, \boldsymbol{z}_\kappa) \ = \ \min_\lambda d(\boldsymbol{x}_t, \boldsymbol{z}_\lambda)$$

Für die euklidische Distanz gilt natürlich $\boldsymbol{z}_\kappa = \boldsymbol{\mu}(\omega_\kappa)$ für alle $\kappa = 1, \ldots, K$.

Stapelweiser K-means-Algorithmus

Lloyd 1957 · Forgy 1965

⟨Algorithmus⟩

1. INITIALISIERUNG
   Wähle eine zufällige Startpartition
   $$\omega_1 \uplus \omega_2 \uplus \ldots \uplus \omega_K \;=\; \omega$$
2. REPRÄSENTATION
   Berechne alle neuen Prototypen
   $$\boldsymbol{z}_\kappa \;=\; \boldsymbol{\mu}_{\mathrm{ZEN}}(\omega_\kappa) \overset{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{y}\in\boldsymbol{\Omega}} \sum_{\boldsymbol{x}\in\omega_\kappa} d(\boldsymbol{x},\boldsymbol{y})$$
3. REKLASSIFIKATION
   Berechne alle neuen Gruppen
   $$\omega_\kappa \;=\; \left\{ \boldsymbol{x}_t \in \omega \mid \operatorname*{argmin}_\lambda d(\boldsymbol{x}_t, \boldsymbol{z}_\lambda) = \kappa \right\}$$
4. TERMINIERUNG
   Wenn $\varepsilon(\{\omega_1,\ldots,\omega_K\}) \le \theta$ dann Ende sonst $\rightsquigarrow$ 2.

⟨Terminierung⟩

Inkrementeller K-means-Algorithmus

MacQueen 1967

⟨Algorithmus⟩

1. INITIALISIERUNG
   Wähle zufällige Startprototypen $\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_K\}$, setze $t\leftarrow 1$.
2. REKLASSIFIKATION
   Wähle $\boldsymbol{y} = \boldsymbol{x}_{t \bmod T}$ und setze $\kappa = \operatorname{argmin}_\lambda d(\boldsymbol{y}, \boldsymbol{z}_\lambda)$.
3. REPRÄSENTATION
   Verschiebe $\boldsymbol{z}_\kappa \leftarrow \alpha_t \cdot \boldsymbol{y} + (1-\alpha_t)\cdot \boldsymbol{z}_\kappa$.
4. TERMINIERUNG
   Wenn $\varepsilon(\cdot) \le \theta$ dann Ende sonst $t\leftarrow t+1$ und $\rightsquigarrow$ 2.

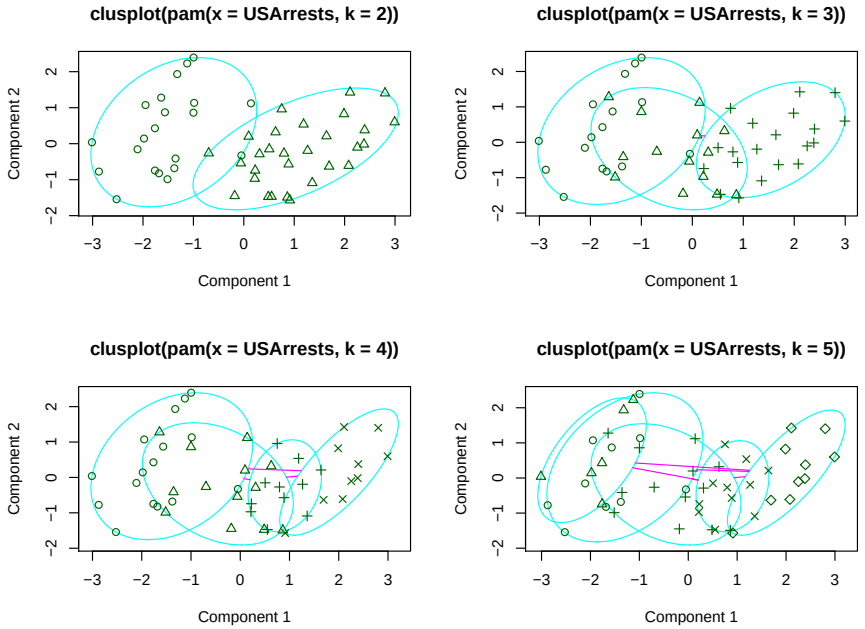
⟨Terminierung⟩

Bemerkungen

1. Die Gewinnerprototypen $\boldsymbol{z}_\kappa$ werden nach jedem Einzelschritt aktualisiert.
2. Die Datenprobe wird zyklisch oder randomisiert durchlaufen.
3. Distanz $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|^2$ ➪ Zentroid ≙ Mittelwert.
4. Mittelungsgewichte *exponentiell* ($\alpha_t \equiv \alpha_0$) oder *kumulativ* ($\alpha_t = 1/|\omega_\kappa|$).
5. Schnellerer Abstieg — aber Oszillationsgefahr!

Beispiel — 'USArrests'-Datensatz

K-medoids-Algorithmus minimiert $\|\cdot\|^1$-Summe · robuster als K-means



Unscharfe Gruppierung

GESUCHT

ist eine **Zugehörigkeitsfunktion** für den Datensatz $\omega \subset \boldsymbol{\Omega}$:

$$\boldsymbol{u} : \left\{ \begin{array}{ccc} \omega & \to & [0,1]^K \\ \boldsymbol{x}_t & \mapsto & \{u_{\kappa,t}\}_{\kappa=1}^K \end{array} \right. , \qquad \sum_\kappa u_{\kappa,t} = 1 \;(\forall t)$$

Fuzzy K-means Zielgröße

Distanzfunktion ist (hier) der quadrierte euklidische Abstand:

$$\varepsilon(\{u_\kappa\},\{\boldsymbol{z}_\kappa\}) \;=\; \sum_{\kappa=1}^K \sum_{t=1}^T (u_\kappa(\boldsymbol{x}_t))^\alpha \cdot \|\boldsymbol{x}_t - \boldsymbol{z}_\kappa\|^2 , \qquad \alpha \ge 1$$

Opt. Prototypen/Zugehörigk.

Normierung $\rightsquigarrow$ Lagrangemultiplikatoren

$$\sum_{\boldsymbol{x}\in\omega} \beta_{\boldsymbol{x}} \cdot \left(\sum_{\kappa=1}^K u_\kappa(\boldsymbol{x}) - 1 \right)$$

Spezialfälle

$\left\{ \begin{array}{l} \alpha = 1\text{: scharfe Datenmengen} \\ \alpha = 2\text{: unscharfe Datenmengen} \\ \alpha = \infty\text{: identische Gruppen} \end{array} \right\}$

Fuzzy K-means-Algorithmus

⟨Algorithmus⟩

1. INITIALISIERUNG
   Wähle zufällige Startzugehörigkeiten $u_{\kappa,t} \in [0,1]$.
2. PROTOTYPEN
   Für alle $1 \leq \kappa \leq K$ berechne

$$\boldsymbol{z}_\kappa \;=\; \sum_{\boldsymbol{x}\in\omega} (u_\kappa(\boldsymbol{x}))^\alpha \cdot \boldsymbol{x} \;\Big/\; \sum_{\boldsymbol{x}\in\omega} (u_\kappa(\boldsymbol{x}))^\alpha$$

3. ZUGEHÖRIGKEITEN
   Für alle $1 \leq \kappa \leq K$ und $\boldsymbol{x} \in \omega$ berechne neue unscharfe Gruppen:

$$u_\kappa(\boldsymbol{x}) \;=\; 1 \;\Big/\; \sum_{\lambda=1}^{K} \left(\frac{\|\boldsymbol{x}-\boldsymbol{z}_\kappa\|^2}{\|\boldsymbol{x}-\boldsymbol{z}_\lambda\|^2} \right)^{\frac{1}{\alpha-1}}$$

4. TERMINIERUNG
   Wenn $\varepsilon \leq \theta$ dann Ende sonst $\rightsquigarrow$ 2.

⟨zumAnfang⟩

Zugehörigkeitsfunktionen

Fuzzy Clustering mit steileren Abklingraten

Harmonische Zugehörigkeitsfunktion

(Standardverfahren: Fuzzy K-means)

Cauchy Zugehörigkeitsfunktion

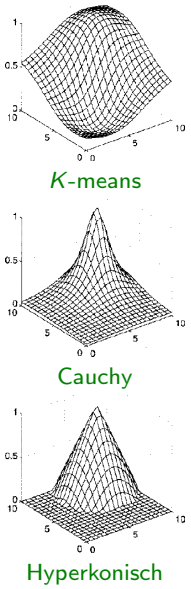
mit den Halbwertsbreiten $\eta_\kappa > 0$:

$$u_\kappa(\boldsymbol{x}) \stackrel{\text{def}}{=} 1 \;\Big/\; \left(1 + \left(\frac{\|\boldsymbol{x}-\boldsymbol{z}_\kappa\|^2}{\eta_\kappa} \right)^{\frac{1}{\alpha-1}} \right)$$

Hyperkonische Zugehörigkeitsfunktion

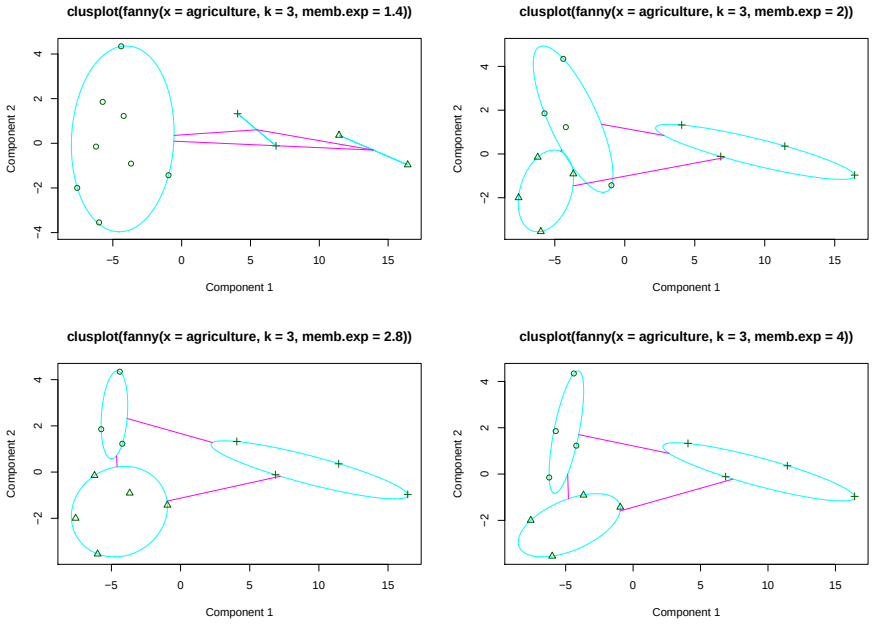
mit den Radien $r_\kappa > 0$:

$$u_\kappa(\boldsymbol{x}) \stackrel{\text{def}}{=} \begin{cases} 1 - \|\boldsymbol{x}-\boldsymbol{z}_\kappa\|/r_\kappa & \text{falls } \|\boldsymbol{x}-\boldsymbol{z}_\kappa\| \leq r_\kappa \\ 0 & \text{sonst} \end{cases}$$



Beispiel — 'agriculture'-Datensatz

Fuzzy K-means-Algorithmus ($\alpha \in \left\{ \sqrt{2}^i \mid i = 1,2,3,4 \right\}$)



Geometrische Clusterformen

Perpendikulare Linien im $\mathbb{R}^N$

Punktförmiges Zentrum im $\mathbb{R}^2$

$$d^2(\boldsymbol{x},\boldsymbol{z}) \;=\; \|\boldsymbol{x}-\boldsymbol{z}\|^2 \;=\; (x_1-z_1)^2 + (x_2-z_2)^2$$

Vertikal linienförmig im $\mathbb{R}^2$

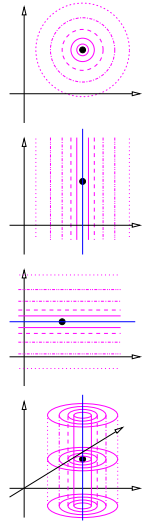
$$d^2(\boldsymbol{x},\boldsymbol{z}) \;=\; (x_1-z_1)^2$$

Horizontal linienförmig im $\mathbb{R}^2$

$$d^2(\boldsymbol{x},\boldsymbol{z}) \;=\; (x_2-z_2)^2$$

Vertikal linienförmig im $\mathbb{R}^3$

$$\begin{aligned} d^2(\boldsymbol{x},\boldsymbol{z}) &= (x_1-z_1)^2 + (x_2-z_2)^2 \\ &= \|\boldsymbol{x}-\boldsymbol{z}\|^2 - (x_3-z_3)^2 \end{aligned}$$



Perpendikulare Linienzentren

Unendlich lange, koordinatenachsenparallele Cluster

Linienförmiges Klassenzentrum

des $\mathbb{R}^N$ in Richtung der x_n-Achse, $n \in \{1, \ldots, N\}$:

$$\begin{aligned} d^2(\boldsymbol{x} \mid \boldsymbol{z}, n) &= \|\boldsymbol{x}-\boldsymbol{z}\|^2 - (x_n - z_n)^2 \\ &= \|\boldsymbol{x}-\boldsymbol{z}\|^2 - (\boldsymbol{e}_n^\top \cdot (\boldsymbol{x}-\boldsymbol{z}))^2 \\ &= \|\boldsymbol{x}-\boldsymbol{z}\|^2 - \left\|\boldsymbol{e}_n^\top \cdot (\boldsymbol{x}-\boldsymbol{z})\right\|^2 = d^2(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{e}_n) \end{aligned}$$

Vom euklidischen Abstand wird also die Norm einer Achsenprojektion subtrahiert.

Verallgemeinerung auf „*schräge*“ Cluster ?

Es ist naheliegend, daß dieser Zusammenhang auch für den nichtperpendikularen Fall gilt.

Achsenrotation

Datentransformation $\phi : \boldsymbol{x} \mapsto \boldsymbol{U}^\top \boldsymbol{x}$, $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{E}$

Rotationen sind distanzinvariant

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{E} \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x} = \left\|\boldsymbol{U}^\top \boldsymbol{x}\right\|^2 = \sum_{n=1}^{N} (\boldsymbol{u}_n^\top \boldsymbol{x})^2$$

Summendarstellung mit den Spaltenvektoren $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ von $\boldsymbol{U}$.

Linienbezogener Abstand in $\boldsymbol{u}_n$-Richtung

$$\begin{aligned} d^2(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{u}_n) &= \|\phi\boldsymbol{x} - \phi\boldsymbol{z}\|^2 - ((\phi\boldsymbol{x})_n - (\phi\boldsymbol{z})_n)^2 \\ &= \|\phi(\boldsymbol{x}-\boldsymbol{z})\|^2 - (\boldsymbol{u}_n^\top \boldsymbol{x} - \boldsymbol{u}_n^\top \boldsymbol{z})^2 \\ &= \|\boldsymbol{x}-\boldsymbol{z}\|^2 - (\boldsymbol{u}_n^\top (\boldsymbol{x}-\boldsymbol{z}))^2 \end{aligned}$$

Flächenbezogener Abstand in $\boldsymbol{u}, \boldsymbol{v}$-Richtung

(Richtungsvektoren $\boldsymbol{u}, \boldsymbol{v}$ normiert und senkrecht zueinander)

$$d^2(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{x}-\boldsymbol{z}\|^2 - (\boldsymbol{u}^\top (\boldsymbol{x}-\boldsymbol{z}))^2 - (\boldsymbol{v}^\top (\boldsymbol{x}-\boldsymbol{z}))^2$$

M-dimensionale Hyperflächenzentren

Satz (Pythagoras)

Sei $0 \leq M \leq N$. *Für alle* $\boldsymbol{x} \in \mathbb{R}^N$ *berechnet sich der lotrechte Abstand zwischen* $\boldsymbol{x}$ *und der* M*-dimensionalen Hyperfläche mit dem* ***Aufpunktvektor*** $\boldsymbol{z}$ *und den orthonormalen* ***Richtungsvektoren*** $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M \in \mathbb{R}^N$ *gemäß*

$$\min_{a_1, \ldots, a_M} \left\| \boldsymbol{x} - \left(\boldsymbol{z} + \sum_{m=1}^{M} a_m \boldsymbol{u}_m\right) \right\|^2 = \|\boldsymbol{x}-\boldsymbol{z}\|^2 - \left\|\boldsymbol{U}^\top (\boldsymbol{x}-\boldsymbol{z})\right\|^2,$$

wenn $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$ *ist.*

Beweis.

Der Abstand für ein M-dimensionales Zentrum, das durch die orthonormalen Vektoren $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M) \in \mathbb{R}^{N \times M}$ aufgespannt wird:

$$d^2(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{U}) = \|\boldsymbol{x}-\boldsymbol{z}\|^2 - \sum_{m=1}^{M} (\boldsymbol{u}_m^\top (\boldsymbol{x}-\boldsymbol{z}))^2 = \|\boldsymbol{x}-\boldsymbol{z}\|^2 - \left\|\boldsymbol{U}^\top (\boldsymbol{x}-\boldsymbol{z})\right\|^2$$

□

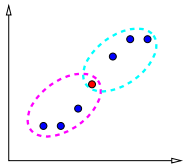
Fuzzy K-Varieties

Definition

Das (unscharfe) Gruppierungsverfahren mit der Zielgröße

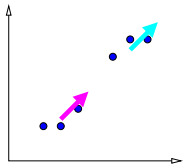
$$\varepsilon(\{\omega_\kappa\}) = \sum_{\kappa=1}^{K} \sum_{\boldsymbol{x} \in \omega} u_\kappa(\boldsymbol{x})^\alpha \cdot d^2(\boldsymbol{x} \mid \boldsymbol{z}_\kappa, \boldsymbol{U}_\kappa)$$

heißt **fuzzy K-varieties**-Algorithmus; im Spezialfall $M = 1$ heißt es **fuzzy K-lines**-Algorithmus.



Elliptotypzentren („*fuzzy K-elliptotypes*“)

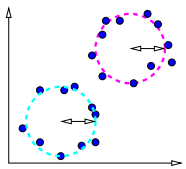
$$d^2(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{U}) = \|\boldsymbol{x}-\boldsymbol{z}\|^2 - \rho \cdot \left\|\boldsymbol{U}^\top (\boldsymbol{x}-\boldsymbol{z})\right\|^2$$



Spezialfälle: $\rho = 0$ Punktzentrum · $\rho = 1$ Hyperflächenzentrum

Hyperkugelschalen („*fuzzy K-shells*“)

$$d^2(\boldsymbol{x} \mid \boldsymbol{z}, r) = (\|\boldsymbol{x}-\boldsymbol{z}\| - r)^2$$



Gradientenabstieg für Fuzzy K-Varieties

Lemma

Die Minimierung der Zielgröße mit Lagrangemultiplikatoren für die Normierungsbedingungen liefert die Bestimmungsgleichungen

$$\boldsymbol{z}_\kappa \;=\; \cdot \sum_{\boldsymbol{x}\in\omega} (u_\kappa(\boldsymbol{x}))^\alpha \cdot \boldsymbol{x} \;\Big/\; \sum_{\boldsymbol{x}\in\omega} (u_\kappa(\boldsymbol{x}))^\alpha$$

für die ***Aufpunktvektoren****,*

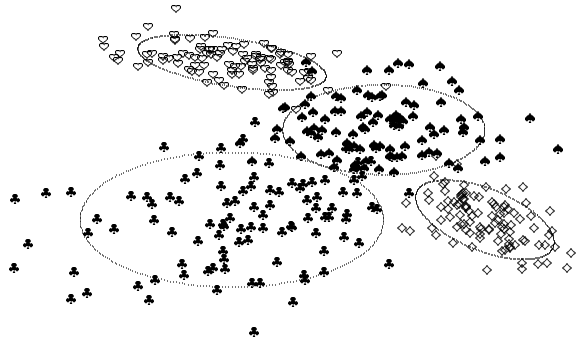
$$u_\kappa(\boldsymbol{x}) \;=\; 1 \;\Big/\; \sum_{\lambda=1}^{K} \left(\frac{d^2(\boldsymbol{x} \mid \boldsymbol{z}_\kappa, \boldsymbol{U}_\kappa)}{d^2(\boldsymbol{x} \mid \boldsymbol{z}_\lambda, \boldsymbol{U}_\lambda)} \right)^{\frac{1}{\alpha-1}}$$

für die ***Gruppenzugehörigkeiten*** *und für die* ***Gruppenkovarianzen***

$$\boldsymbol{S}_\kappa \;=\; \sum_{\boldsymbol{x}\in\omega} u_\kappa(\boldsymbol{x})^\alpha (\boldsymbol{x}-\boldsymbol{z}_\kappa)(\boldsymbol{x}-\boldsymbol{z}_\kappa)^\top \;.$$

Die m-te Spalte $\boldsymbol{u}_{\kappa,m}$ von $\boldsymbol{U}_\kappa$ schließlich ergibt sich als Eigenvektor zum m-größten Eigenwert von $\boldsymbol{S}_\kappa$.

Identifikation von Mischverteilungen



Problem

Angenommen, obige Daten sind gemäß $f(\boldsymbol{x}) = \sum_{\kappa=1}^{K} \pi_\kappa \cdot g(\boldsymbol{x}|\boldsymbol{\theta}_\kappa)$ *mischverteilt. Wie lauten die* ***bestpassenden*** *(ML) Verteilungsparameter* $\hat{\pi}_\kappa$, $\hat{\boldsymbol{\theta}}_\kappa$, $\kappa = 1, \ldots, K$ *des Modells?*

Lösung

Im Normalverteilungsfall $g(\boldsymbol{x}|\boldsymbol{\theta}_\kappa) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_\kappa, \boldsymbol{S}_\kappa)$ *existiert eine* ***asymptotisch*** *eindeutige Lösung sowie ein* ***lokales*** *Optimierungsverfahren (EM).*

EM-Algorithmus

zur Identifikation gaußscher Mischverteilungen

⟨Algorithmus⟩

1. INITIALISIERUNG
   Wähle zufällige Startparameter $(\pi_\kappa, \boldsymbol{\mu}_\kappa, \boldsymbol{S}_\kappa)$, $\kappa = 1, \ldots, K$.
2. ERWARTUNGSWERTE
   Berechne für $\kappa = 1..K$ und $t = 1..T$ die a posteriori Wahrsch'keiten

$$\gamma_{\kappa,t} \overset{\text{def}}{=} \mathrm{P}(\Omega_\kappa|\boldsymbol{x}_t) \;=\; \frac{\mathrm{P}(\Omega_\kappa)\cdot\mathrm{P}(\boldsymbol{x}_t|\Omega_\kappa)}{\mathrm{P}(\boldsymbol{x}_t)} \;\propto\; \pi_\kappa \cdot \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_\kappa, \boldsymbol{S}_\kappa)$$

3. MAXIMIERUNG

$$\pi_\kappa \leftarrow \frac{\sum_t \gamma_{\kappa,t}}{\sum_\lambda \sum_t \gamma_{\lambda,t}}\,, \quad \boldsymbol{\mu}_\kappa \leftarrow \frac{\sum_t \gamma_{\kappa,t}\boldsymbol{x}_t}{\sum_t \gamma_{\kappa,t}}\,, \quad \boldsymbol{S}_\kappa \leftarrow \frac{\sum_t \gamma_{\kappa,t}\boldsymbol{x}_t\boldsymbol{x}_t^\top}{\sum_t \gamma_{\kappa,t}} - \boldsymbol{\mu}_\kappa\boldsymbol{\mu}_\kappa^\top$$

4. TERMINIERUNG
   Wenn die ML-Zielgröße $\ell(\ldots)$ stagniert dann Ende sonst $\rightsquigarrow$ 2.

⟨zumAlgorithmus⟩

Konvergenzeigenschaften

des EM-Algorithmus für Gaußsche Mischverteilungsmodelle (GMM)

1. **Schwache Monotonie**
   Verfahren erreicht stationären Punkt

$$\ell(\theta_0) \leq \ell(\theta_1) \leq \ell(\theta_2) \leq \ell(\theta_3) \leq \ldots \leq \ell(\theta_j) \leq \ldots \leq \ldots$$

2. **Beschränktheit**
   pathologische Aufgabenstellung („*ill-posed problem*“)

$$\mathcal{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{S}_\lambda) = \mathcal{N}(\boldsymbol{x}_t, \mathbf{0})$$

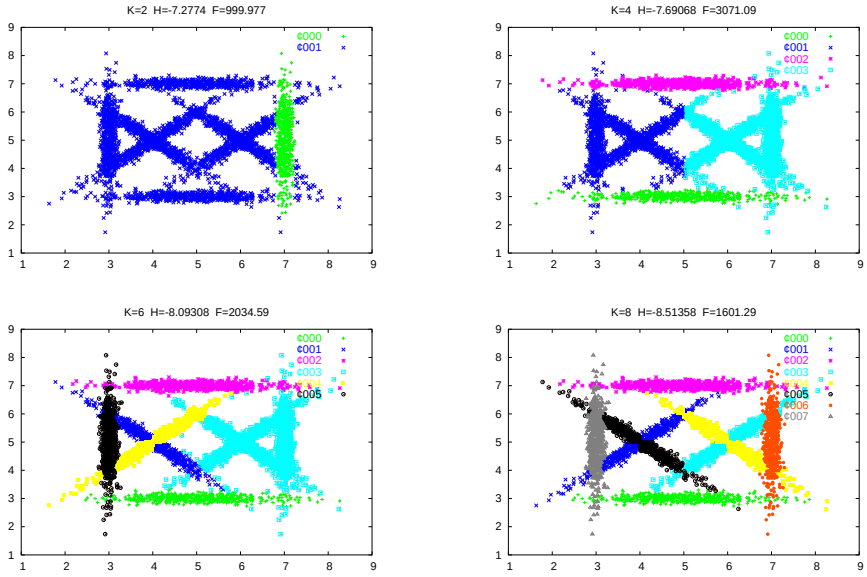
3. **Lokale Maxima**
   viele relative Maxima mit $\ell(\theta) < \infty$ und großem Einzugsbereich

4. **Zyklischer Iterationsverlauf**
   Kraterrandphänomen

$$\theta_1 \neq \theta_2 \neq \ldots \neq \theta_m \quad \text{mit} \quad \ell(\theta_1) = \ell(\theta_2) = \ldots = \ell(\theta_m)$$

Verhalten in unkritischen Fällen

Was kann EM, das *K*-mean nicht kann ?



Problematik des Rangdefizits

$\text{rg}(\boldsymbol{S}_\kappa) < N \Rightarrow \det(\boldsymbol{S}_\kappa) = 0 \Rightarrow \boldsymbol{S}_\kappa^{-1} = ?$

Gratregularisierung

Anisotropes Aufblasen der Konzentrationsellipse („*Speckschicht*“)

$$\boldsymbol{S}^{(\delta)} \stackrel{\text{def}}{=} \boldsymbol{S} + \delta \cdot \boldsymbol{E} = \begin{pmatrix} s_{11}+\delta & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22}+\delta & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN}+\delta \end{pmatrix}$$

Fixierung & Verklebung

Alle Eigenschaften des EM-Algorithmus bleiben erhalten:

· Kovarianzmatrizen fixieren ($\forall \kappa : \boldsymbol{S}_\kappa \stackrel{!}{=} \boldsymbol{S}^*$) $\rightsquigarrow$ keine pathologische Lösung

· Kovarianzmatrizen verkleben ($\forall \kappa, \lambda : \boldsymbol{S}_\kappa \stackrel{!}{=} \boldsymbol{S}_\lambda$) $\rightsquigarrow$ mehr Robustheit

Hintergrundkomponente

Streuungsintensives Rückweisungscluster zur Ausreißerbehandlung

$$f_0(\cdot) = \mathcal{N}(\cdot \mid \boldsymbol{\mu}(\omega), \boldsymbol{S}_0) \quad \text{mit } \boldsymbol{S}_0 = \boldsymbol{S}(\omega) \text{ oder } \boldsymbol{S}_0 = \sigma_0^2 \cdot \boldsymbol{E}$$

Probabilistische PCA

Zerlegung des $\mathbb{R}^N$ in systematisch und in zufällig streuende Komponenten

Normalverteilungsmodelle für rangdefizite Daten

Das homogene Faktoranalysemodell

$$\mathbb{X} = \boldsymbol{\mu} + \mathbb{E} + \boldsymbol{A} \cdot \mathbb{V} \quad \text{mit} \quad \begin{cases} \boldsymbol{\mu} \in \mathbb{R}^N \\ \boldsymbol{A} \in \mathbb{R}^{N \times M} \\ \mathbb{E} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \boldsymbol{E}_N) \\ \mathbb{V} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{E}_M) \end{cases}$$

(Dimension $M \leq N$; PCA-Annahme identischer Störvarianzen)

- besitzt als Ladungsvektoren die bereits hinlänglich bekannten M führenden Hauptachsen der Verteilungsellipse,
- definiert aber gleichzeitig eine explizite Wahrscheinlichkeitsverteilung für die Daten.

PPCA-Schätzung bei bekanntem Modellrang M

Lemma (Tipping & Bishop 1999)

Der Zufallsvektor des homogenen FA-Modells ist gemäß $\mathbb{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ normalverteilt mit der Kovarianzmatrix

$$\boldsymbol{S} \;=\; \boldsymbol{A}\boldsymbol{A}^\top + \sigma^2 \cdot \boldsymbol{E}_N \;.$$

*Der **ML-Schätzer** für $\boldsymbol{S}$ ergibt sich durch Einsetzen der Schätzwerte*

$$\begin{aligned} \hat{\boldsymbol{A}} &= \boldsymbol{U}_M \cdot \left(\boldsymbol{D}_M - \sigma^2 \cdot \boldsymbol{E}_M\right)^{1/2} \\ \hat{\sigma}^2 &= \frac{1}{N-M} \cdot \sum_{j=M+1}^{N} \lambda_j \end{aligned}$$

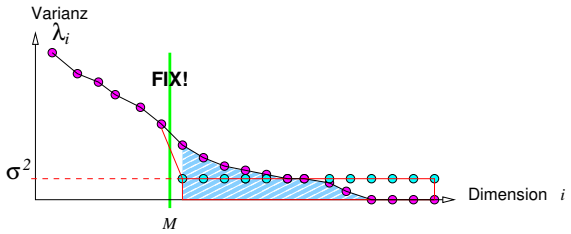
mit der $(M:N)$-eigenzerlegten Datenkovarianzmatrix

$$\hat{\boldsymbol{S}}(\omega) \stackrel{!}{=} (\boldsymbol{U}_M, \boldsymbol{U}'_M) \cdot \begin{pmatrix} \boldsymbol{D}_M & \boldsymbol{0} \\ \boldsymbol{0}^\top & \boldsymbol{D}'_M \end{pmatrix} \cdot (\boldsymbol{U}_M, \boldsymbol{U}'_M)^\top \;, \quad \begin{pmatrix} \boldsymbol{D}_M & \boldsymbol{0} \\ \boldsymbol{0}^\top & \boldsymbol{D}'_M \end{pmatrix} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N) \;.$$

Schätzung der mittleren Reststreuung

unter ausschließlicher Verwendung der $M \ll N$ Hauptachsen

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{N-M} \cdot \operatorname{spur}\left(\boldsymbol{D}'_M\right) \\ &= \frac{1}{N-M} \left\{ \operatorname{spur}\begin{pmatrix} \boldsymbol{D}_M & \boldsymbol{0} \\ \boldsymbol{0}^\top & \boldsymbol{D}'_M \end{pmatrix} - \operatorname{spur}\left(\boldsymbol{D}_M\right) \right\} \\ &= \frac{1}{N-M} \left\{ \operatorname{spur}\left(\hat{\boldsymbol{S}}(\omega)\right) - \operatorname{spur}\left(\boldsymbol{D}_M\right) \right\} \\ &= \frac{1}{N-M} \left\{ \frac{1}{T} \sum_{t=1}^{T} \|\boldsymbol{x}_t - \boldsymbol{\mu}\|^2 - \sum_{j=1}^{M} \lambda_j \right\} \end{aligned}$$



Invertierung der PPCA-Kovarianzmatrix

Kovarianzstruktur

Die ursprüngliche Darstellung zeigt eine rangverminderte Darstellung mit einem additiv aufgeprägtem Fehlergrat von σ^2 auf der Diagonalen.

$$\boldsymbol{S} \;=\; \boldsymbol{A}\boldsymbol{A}^\top + \sigma^2 \cdot \boldsymbol{E}_N$$

Die gleichwertige alternative Darstellung präsentiert eine vollständige Eigenzerlegung mit den kanonischen Eigenvektoren und -werten; nur die letzten $(N-M)$ Eigenwerte wurden gemittelt.

$$\boldsymbol{S} \;=\; \boldsymbol{U}_M \boldsymbol{D}_M \boldsymbol{U}_M^\top + \sigma^2 \cdot \boldsymbol{U}'_M \boldsymbol{U}'^\top_M$$

Inverse Kovarianz

Diese Inversionsformel verwendet ausschließlich die M führenden Eigenvektoren sowie die führenden reziproken Eigenwerte.

$$\boldsymbol{S}^{-1} \;=\; \frac{1}{\sigma^2} \cdot \left\{ \boldsymbol{E}_N - \boldsymbol{U}_M \cdot \left(\boldsymbol{E}_M - \sigma^2 \cdot \boldsymbol{D}_M^{-1}\right) \cdot \boldsymbol{U}_M^\top \right\}$$

PPCA-Schätzung bei bekannter Störvarianz σ^2

Lemma (Meinicke & Ritter 2000)

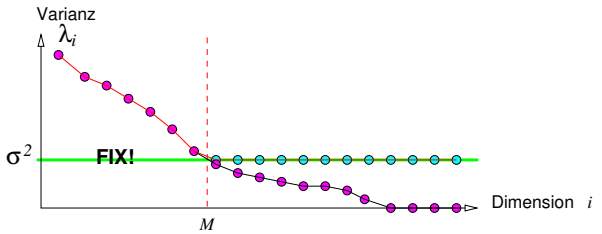
Ein Zufallsvektor sei gemäß $\mathbb{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ normalverteilt mit der Kovarianzmatrix

$$\boldsymbol{S} \;=\; \boldsymbol{\Psi} + \sigma^2 \cdot \boldsymbol{E}_N$$

*mit **bekannter** Störvarianz σ^2 und positiv-semidefinitem $\boldsymbol{\Psi}$ mit **unbekanntem** Rang $\nu = \operatorname{ran}\boldsymbol{\Psi}$, $\nu \leq N$.*

Mit der empirischen Datenkovarianz $\hat{\boldsymbol{S}}(\omega)$ und ihrer Eigenzerlegung $\hat{\boldsymbol{S}}(\omega) = \boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{U}^\top$, $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ ergeben sich die ML-Schätzer

$$\hat{\nu} \;=\; \left|\left\{\lambda_i \mid \lambda_i > \sigma^2\right\}\right| \quad \textit{und} \quad \hat{\boldsymbol{\Psi}} \;=\; \boldsymbol{U}_\nu \cdot \left(\boldsymbol{D}_\nu - \sigma^2 \cdot \boldsymbol{E}_\nu\right) \cdot \boldsymbol{U}_\nu^\top \;.$$



Effiziente Berechnung der PPCA-Dichtewerte

auch in extrem hochdimensionalen ($N \gg M$) Vektorräumen

Determinante $\det(\boldsymbol{S})$

Determinanten sind rotationsinvariant ($\det(\boldsymbol{S}) = \det(\boldsymbol{U}^\top \boldsymbol{S} \boldsymbol{U})$); also gilt:

$$\det(\boldsymbol{S}) \;=\; \prod_{i=1}^{N} \tilde{\lambda}_i \;=\; \sigma^{2\cdot(N-M)} \cdot \prod_{i=1}^{M} \lambda_i \;.$$

Quadratische Form $(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$

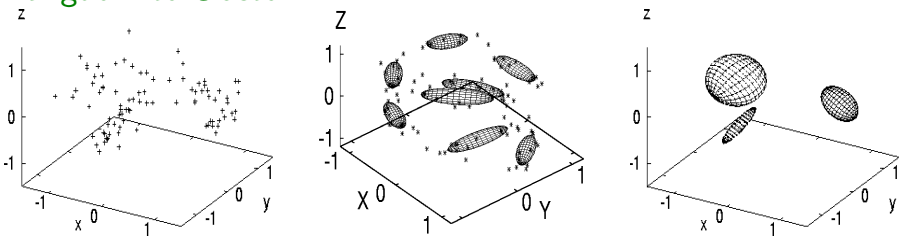
Wegen der Darstellung von $\boldsymbol{S}^{-1}$ gilt für die quadratische Form

$$\begin{aligned} \boldsymbol{y}^\top \boldsymbol{S}^{-1} \boldsymbol{y} &= \frac{1}{\sigma^2} \cdot \left\{ \boldsymbol{y}^\top \boldsymbol{E}_N \boldsymbol{y} \;-\; \boldsymbol{y}^\top \boldsymbol{U}_M \cdot \left(\boldsymbol{E}_M - \sigma^2 \boldsymbol{D}_M^{-1}\right) \cdot \boldsymbol{U}_M^\top \boldsymbol{y} \right\} \\ &= \frac{1}{\sigma^2} \cdot \left\{ \|\boldsymbol{y}\|^2 \;-\; \|\tilde{\boldsymbol{y}}\|^2 \;+\; \sigma^2 \cdot \tilde{\boldsymbol{y}}^\top \boldsymbol{D}_M^{-1} \tilde{\boldsymbol{y}} \right\} \\ &= \frac{\|\boldsymbol{y}\|^2 - \|\tilde{\boldsymbol{y}}\|^2}{\sigma^2} \;+\; \sum_{i=1}^{M} \frac{\tilde{y}_i^2}{\lambda_i} \end{aligned}$$

unter Verwendung des Hauptachsenprojektionsvektors
$\tilde{\boldsymbol{y}} = \boldsymbol{U}_M^\top \boldsymbol{y} = \boldsymbol{U}_M^\top \cdot (\boldsymbol{x} - \boldsymbol{\mu})$.

PPCA-Mischverteilungsidentifikation

Rangdefizite Cluster



Punktmenge im $\mathbb{R}^3$ | unterschiedliche Richtungen | unterschiedliche Ränge

Sphärische Gruppierung

Alle Gaußkomponenten sphärisch (kugelförmige Konzentration; konstante Streuung σ_m^2)

Klassenmittelwertvektoren $\boldsymbol{\mu}_\kappa$

Ende sobald Anzahl K_{eff} Gruppenprototypen gleich Sollgruppenzahl K ist.

Lokaladaptive Gruppierung

PPCA-Gaußkomponenten (zeppelinförmige Konzentration; variable Effektivdimension ν_κ)

Rangdefiziente Matrizen $\boldsymbol{\Psi}_\kappa$

- Reststreuung
- Ränge (Parameterkomplexität)

Zweistufiges EM-Abkühlverfahren

⟨Algorithmus⟩

1. Vorwahl von $\sigma^2_{\max} > 0$, $\sigma^2_{\min} > 0$ und $\alpha \in (0,1)$.
2. Setze $\theta \leftarrow (\boldsymbol{\mu}, \ldots, \boldsymbol{\mu})$, $m \leftarrow 0$ und $\sigma_m^2 \leftarrow \sigma^2_{\max}$.
3. SPHÄRISCHE GRUPPIERUNG (EM)

$$\ell(\boldsymbol{\theta} \mid \omega, \sigma_m^2) \;=\; \sum_t \sum_\kappa \gamma_{\kappa,t} \cdot \log\left(\pi_\kappa \cdot \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_\kappa, \sigma_m^2 \boldsymbol{E})\right) \stackrel{!}{\to} \max$$

4. Setze $m \leftarrow m+1$ und $\sigma_m^2 \leftarrow \alpha \cdot \sigma^2_{m-1}$.
5. Wenn $K_{\text{eff}} < K$ dann $\rightsquigarrow$ 3.
6. LOKALADAPTIVE PPCA-GRUPPIERUNG (EM)

$$\mathcal{L}(\boldsymbol{\Theta} \mid \omega, \sigma_m^2) \;=\; \sum_t \sum_\kappa \gamma_{\kappa,t} \cdot \log\left(\pi_\kappa \cdot \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_\kappa, \boldsymbol{\Psi}_\kappa + \sigma_m^2 \boldsymbol{E})\right) \stackrel{!}{\to} \max$$

7. Setze $m \leftarrow m+1$ und $\sigma_m^2 \leftarrow \alpha \cdot \sigma^2_{m-1}$.
8. Wenn $\sigma_m^2 > \sigma^2_{\min}$ dann $\rightsquigarrow$ 6 sonst Ende.

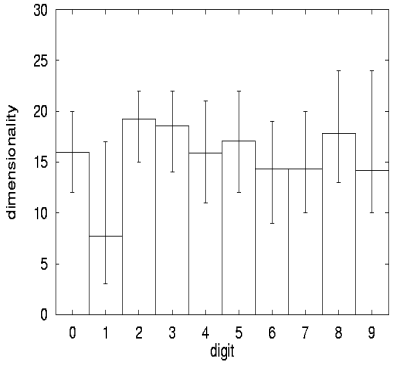
⟨Algorithmus⟩

Beispiel — Handgeschriebene Ziffern

MNIST Datensammlung, LeCun 1998

Datensatz

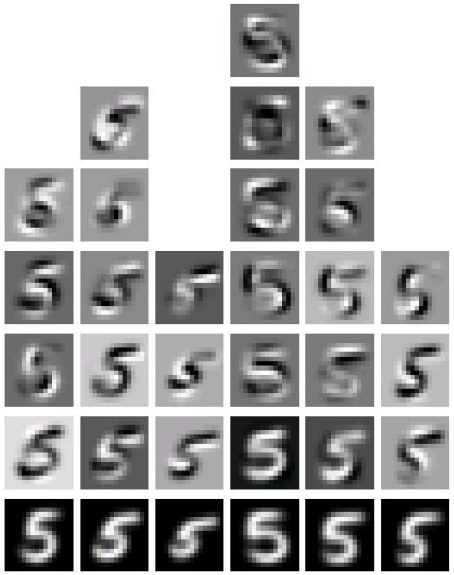
60 000 (10 000) Lern- und Testmuster
Originalziffern 28×28 Pixel zu 8 Bit $\rightsquigarrow$ Umrasterung 8×8
Verschiedene Gruppenstärken $K \in \{1, 2, 4, 8, 16\}$ getestet



PPCA-Dimensionen

alle zehn Ziffernklassen
ν-Durchschnitt und min/max
$K = 16$ Mischungskomponenten
im „*Gefrierpunkt*“ $\sigma^2_{\min}$

Beispiel — die Ziffernklasse „5“



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

PPCA-Mischung
für Ziffernklasse Ω_5

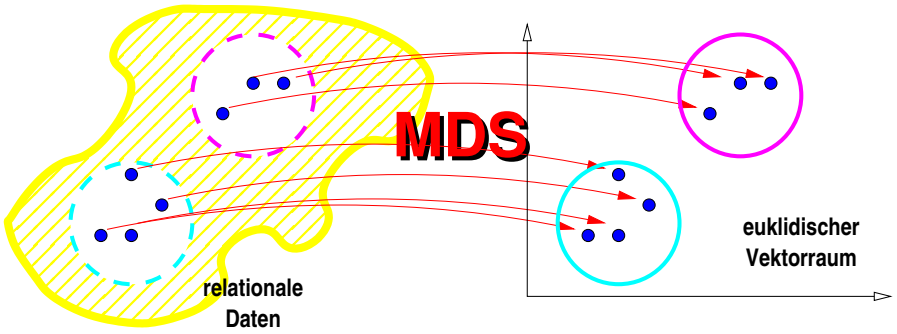
Gruppenanzahl
$K = 6$ gewählt

Modellrang
$M \in \{4, 5, 6, 7\}$ ➪ $M \ll 64$

Hauptachsen
von unten nach oben aufgetürmt
als 8×8-Grauwertbild

Relationale Gruppierung

Datenobjekte mit wechselseitiger Distanz — ohne Attribute



MDS-Gruppierung

1. Mehrdimensionale Skalierung von ω nach $\omega' = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\} \subseteq \mathbb{R}^N$
2. K-means Gruppierung des Datensatzes ω'
3. Aufprägung der ω'-Gruppierung auf die Urbilder ω

Fuzzy K-medoids für Metriken

Datensatz

Objektmenge $\omega = \{o_1, \ldots, o_T\}$ mit der Abstandsmatrix

$$\boldsymbol{R} = [r_{s,t}] \in \mathbb{R}^{T \times T}, \qquad r_{s,t} = d(o_s, o_t)$$

Insbesondere gelte die **Symmetrie** $\boldsymbol{R}^\top = \boldsymbol{R}$ und die **Definitheit** $\operatorname{diag}(\boldsymbol{R}) = \boldsymbol{0}$.

Zugehörigkeitsfunktionen

- **Harmonisch:** $$u_\kappa(o_t) = 1 \Big/ \sum_{\lambda=1}^{K} \left(\frac{r_{t_\kappa,t}}{r_{t_\lambda,t}}\right)^{\frac{2}{\alpha-1}}$$
- **Cauchy / possibilistisch:** $$u_\kappa(o_t) = \left(1 + \left(\frac{r_{t_\kappa,t}}{\sqrt{\eta_\kappa}}\right)^{\frac{2}{\alpha-1}}\right)^{-1}$$
- **Hyperkonisch:** $$u_\kappa(o_t) = \begin{cases} 1 - r_{t_\kappa,t}/\rho_\kappa & \text{falls } r_{t_\kappa,t} \le \rho_\kappa \\ 0 & \text{sonst} \end{cases}$$

Für *harmonische* Zugehörigkeiten gilt die Normierungseigenschaft $\sum_\lambda u_\lambda(o_t) = 1$.

RACE — Relationaler Austauschalgorithmus

⟨Algorithmus⟩

GEGEBEN
Datenrelation $\boldsymbol{R} \in \mathbb{R}^{T\times T}$, Gruppenzahl $K \in \mathbb{N}$, Iterationen $I \in \mathbb{N}$.

1. INITIALISIERUNG
   Setze $i \leftarrow 1$.
   Wähle zufällige Prototypenindizes $\{t_1, \ldots, t_K\} \subseteq \{1, \ldots, T\}$.
2. ITERATIONSSCHRITT
   1. Bestimme alle Zugehörigkeiten $u_\kappa(o_t)$
   2. Bestimme die „*Restenergien*“
   $$e_{\kappa,t} \;=\; \sum_{\lambda\neq\kappa} u_\lambda(o_t)$$
   3. Bestimme die neuen Prototypen
   $$t_\kappa \;\leftarrow\; \operatorname*{argmin}_{t=1..T} e_{\kappa,t}$$
3. TERMINIERUNG
   Wenn $i = I$ dann $\rightsquigarrow$ Ende sonst $i \leftarrow i+1$, $\rightsquigarrow$ 2.

⟨zumdrnoglA⟩

Defuzzifizierung

Finales Schärfen (Aushärten) der Gruppenzugehörigkeiten

$$u_\kappa(o_t) \overset{\text{def}}{=} \begin{cases} 1 & \text{falls } u_\kappa(o_t) = \max_\lambda u_\lambda(o_t) \\ 0 & \text{sonst} \end{cases}, \qquad \kappa = 1, \ldots, K$$

Diese Aushärtungsregel ergibt Partitionen mit variablen Gruppenstärken.

ESS-Defuzzyfizierung („*equal size subset*“)

⟨Algorithmus⟩

1. INITIALISIERUNG
   Setze $\mathcal{I} \leftarrow \{1, 2, \ldots, T\}$ (Indexmenge)
   Setze $\mathcal{C}_\kappa \leftarrow \varnothing$ (Gruppenindexmenge) für alle $\kappa = 1..K$
2. ITERATION (für alle $t = 1, \ldots, T$)
   1. Setze Gruppenindex $\lambda = t \bmod K + 1$.
   2. Bestimme bestpassenden Restindex $t^* = \operatorname{argmax}_{t\in\mathcal{I}} u_\lambda(o_t)$.
   3. Verschiebe Index t^* von $\mathcal{I}$ nach $\mathcal{C}_\lambda$.
3. TERMINIERUNG
   Jede Gruppe enthält entweder $\lfloor T/K \rfloor$ oder $\lceil T/K \rceil$ Datenelemente.

⟨zumdrnoglA⟩

Beispiel — Text Mining

Automatische Erstellung eines Stichwortinventars

Datensammlung

Kapitel 2 aus dem Buch „*Information Mining*“ (Th. Runkler)
Alle Formeln und Sonderzeichen wurden entfernt.
Großbuchstaben $\mapsto$ Kleinbuchstaben

Objektmenge und Metrik

1605 Wortvorkommen, davon $T = 564$ verschieden
Matrix $\boldsymbol{R} \in \mathbb{R}^{564\times 564}$ der Levenshteinabstände

Verarbeitung

RACE-Algorithmus mit $K = 20$ Gruppen und $I = KT = 11280$ Schritten
ESS-Defuzzifizierung auf 28 (bzw. 29) Wörter/Gruppe

Die 20 häufigsten Wörter des Textes

die	*der*	*und*	*für*	*in*	*als*	*werden*	*ist*	*sich*	*mit*
oder	*den*	*sind*	*ein*	*auch*	*daten*	*läßt*	*können*	*abstand*	*wird*

Beispiel — Text Mining

Gruppenprototypen $o_{t_1}, o_{t_2}, \ldots, o_{t_{20}}$

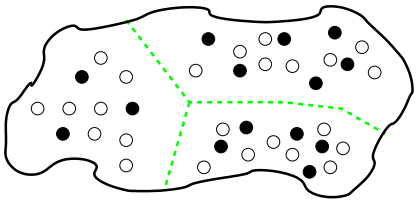
originalsignal	*mengenschreibweise*	*bzw*
inkompatibilität	*quantisierungsschritte*	*wertkontinuierlich*
intervallskalierten	*unterschiedlichen*	*übereinstimmungen*
matrixdarstellung	*mindestabtastrate*	*abtastzeitpunkten*
datencharakteristika	*ordnungsrelation*	*objektdatensatz*
quantisierungsfehler	*polygonzug*	*kovarianzmatrix*
speicherplatzes	*kaufmännisches*	

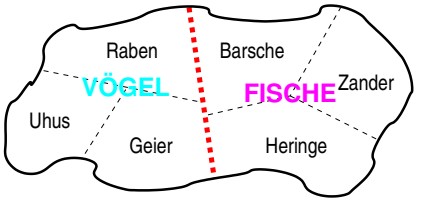
Gruppen Ω_1, Ω_2 und Ω_6

(die Wörter mit den höchsten Zugehörigkeitsbewertungen)

Gruppe 1	**Gruppe 2**	**Gruppe 6**
originalsignal	*mengenschreibweise*	*wertkontinuierlich*
zeitsignal	*matrixschreibweise*	*zeitkontinuierlich*
ordinal	*beschreiben*	*kontinuierliche*
signal	*schrittweise*	*wertebereich*
signals	*schreiben*	*rekonstruieren*
digitalen	*beschreibt*	*nichtnumerisch*
zeitsignalen	*geschrieben*	*willkürlich*
proportional	*beschrieben*	*konstruierten*

Konzeptuelle Gruppierung





Gegeben

Objektbereich $\mathbf{\Omega}$ · Hypothesenraum $\mathcal{H}$ · Beispielmenge $\omega \subseteq \mathbf{\Omega}$

Gesucht

eine **intensionale Partition** von ω, d.h.:

> *Eine Folge von Hypothesen $h_1, \ldots, h_K$,*
> *welche die beobachteten Beispiele aus ω*
> *sowie auch neue Objekte aus $\mathbf{\Omega} \setminus \omega$*
> *überschneidungsfrei gegeneinander abgrenzen.*

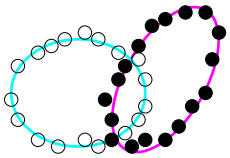
Und welches **Gütekriterium** optimiert die Partition ? ... ?

Lokale *versus* globale Objektähnlichkeit

Traditionelle Clusteranalyse

Lokaler Ähnlichkeitsbegriff (Hamming-Distanz)

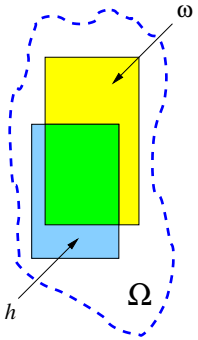
$$d(\boldsymbol{x}, \boldsymbol{y}) \;=\; \#\{\ell \mid x_\ell \neq y_\ell\}$$



Konzeptuelle Gruppierung

Globale Ähnlichkeit (engste $\mathcal{H}$-Umfassung)

$$d(\boldsymbol{x}, \boldsymbol{y}) \;=\; \min\{\sigma(h, \omega) \mid h \models \boldsymbol{x}, \boldsymbol{y}\}$$



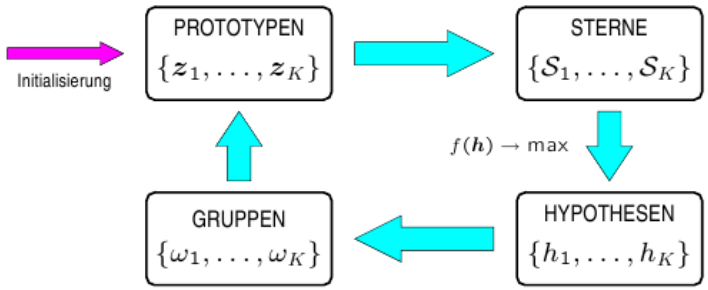
Überdeckungsgrad

der Hypothese h durch die Daten ω:

$$\sigma(h, \omega) \;=\; \frac{\#\{\boldsymbol{x} \in \omega \mid h \models \boldsymbol{x}\}}{\#\{\boldsymbol{x} \in \mathbf{\Omega} \mid h \models \boldsymbol{x}\}} \;=\; \frac{\blacksquare}{\blacksquare}$$

Wiederholtes Austauschen von Gruppenelementen

„conceptual K-means“



Gütekriterium

Kumulativer Überdeckungsgrad

$$f(\boldsymbol{h}) \;=\; \frac{1}{K} \cdot \sum_{k=1}^{K} \sigma(h_k, \omega)$$

Gruppenprototypen

müssen unbedingt ω angehören!
(Sterne, Überdeckung)

- Medoide
- Pseudomediane

Wahl der Gruppenzentren

Medoid

Dasjenige Gruppenelement mit minimaler **Exzentrizität**:

$$\boldsymbol{\mu}_{\text{med}}(\omega_\kappa) \;=\; \operatorname*{argmin}_{\boldsymbol{x}\in\omega_\kappa} \varepsilon(\boldsymbol{x},\omega_\kappa) \;=\; \operatorname*{argmin}_{\boldsymbol{x}\in\omega_\kappa} \sum_{\boldsymbol{y}\in\omega_\kappa} d(\boldsymbol{x},\boldsymbol{y})$$

Pseudomedian

Kombiniere je nach Skalentyp der $\mathcal{X}_n$, $n = 1,\ldots,N$ komponentenweise Mittelwerte, Zentroide, Mediane:

$$\boldsymbol{\mu}_{\text{pseudo}}(\omega_\kappa) \;\stackrel{\text{def}}{=}\; (\mu_{\kappa,1},\ldots,\mu_{\kappa,N})^\top\,, \qquad \mu_{\kappa,n} \;\stackrel{\text{def}}{=}\; \mu(\{x_n \mid \boldsymbol{x}\in\omega_\kappa\})$$

Ergibt eine hocheffizient berechenbare Näherung $\boldsymbol{\mu}_{\text{pseudo}} \approx \boldsymbol{\mu}_{\text{med}}$.

Reintegration

Wegen $\boldsymbol{\mu}_{\text{pseudo}}(\omega_\kappa) \notin \omega_\kappa$ verwende den nächsten ω_κ-Nachbarn:

$$\tilde{\boldsymbol{\mu}}_{\text{pseudo}}(\omega_\kappa) \;=\; \operatorname*{argmin}_{\boldsymbol{x}\in\omega_\kappa} d(\boldsymbol{\mu}_{\text{pseudo}}(\omega_\kappa),\boldsymbol{x})$$

Konzeptueller Austauschalgorithmus

⟨Algorithmus⟩

1. INITIALISIERUNG
   Wähle Klassenzahl K und wähle $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_K$ zufällig aus ω.
2. PROTOTYPEN ➪ STERNE
   $$\mathcal{S}_\kappa \;=\; \mathcal{S}(\boldsymbol{z}_\kappa \mid \{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_K\}\setminus\{\boldsymbol{z}_\kappa\})$$
3. STERNE ➪ HYPOTHESEN
   Bestimme ω-**eindeutigen** Hypothesensatz $\boldsymbol{h}\in\mathcal{S}_1\times\ldots\times\mathcal{S}_K$ mit
   $$f(\boldsymbol{h}) \;=\; \text{MAX}$$
4. HYPOTHESEN ➪ GRUPPEN
   $$\omega_\kappa \;=\; \{x\in\omega \mid h_\kappa \models x\}$$
5. GRUPPEN ➪ PROTOTYPEN
   $$\boldsymbol{z}_\kappa \;=\; \boldsymbol{\mu}_{\text{med}}(\omega_\kappa)$$
6. TERMINIERUNG Wenn $f(\boldsymbol{h}) \geq \theta$ dann $\rightsquigarrow$ Ende sonst $\underset{\text{⟨endlosIA⟩}}{\rightsquigarrow}$ 2.

Beispiel — Gebrauchtwagenhandel

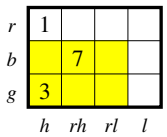
Datensammlung

Objekt:	$\boldsymbol{o}_1$	$\boldsymbol{o}_2$	$\boldsymbol{o}_3$	$\boldsymbol{o}_4$	$\boldsymbol{o}_5$	$\boldsymbol{o}_6$	$\boldsymbol{o}_7$	$\boldsymbol{o}_8$
x_1 Geschwindigkeit	h	m	h	l	m	l	h	m
x_2 Farbe	r	r	g	b	b	g	b	r
x_3 Preis	h	l	h	rh	rl	l	rh	rh

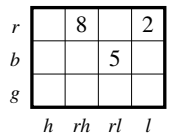
$x_1 \in \mathcal{X}_1 = \{high, medium, low\}$
$x_2 \in \mathcal{X}_2 = \{red, blue, green\}$
$x_3 \in \mathcal{X}_3 = \{high, rel\_high, rel\_low, low\}$

Hypothesen als Attributkomplexe

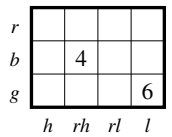
zum Beispiel: $x_1 \in \{h\} \wedge x_2 \in \{b,g\} \wedge x_3 \in \{h, rh, rl\}$



Ebene $x_1 = h$



Ebene $x_1 = m$



Ebene $x_1 = l$

Überdeckung

$\sigma(h,\omega) \;=\; 2/6 \;=\; 0.\bar{3}$

Beispiel — Gebrauchtwagenhandel

Erster Iterationsschritt

P1 Wähle als initiale Prototypen $\boldsymbol{z}_1 = \boldsymbol{o}_1$ und $\boldsymbol{z}_2 = \boldsymbol{o}_2$ aus

S1 Stern von $\boldsymbol{z}_1$:

$$h_1^1 \;=\; (x_2 = r,b) \wedge (x_3 = h,rh,rl) \quad \text{und} \quad h_1^2 \;=\; (x_1 = h,l)$$

Stern von $\boldsymbol{z}_2$:

$$h_2^1 \;=\; (x_1 = h) \wedge (x_2 = g) \vee (x_3 = l) \quad \text{und} \quad h_2^2 \;=\; (x_1 = m)$$

G1 Dann gilt

$h_1^1 \models \boldsymbol{o}_1, \boldsymbol{o}_4, \boldsymbol{o}_5, \boldsymbol{o}_7, \boldsymbol{o}_8$ $\qquad$ $h_2^1 \models \boldsymbol{o}_2, \boldsymbol{o}_3, \boldsymbol{o}_6$
$h_1^2 \models \boldsymbol{o}_1, \boldsymbol{o}_3, \boldsymbol{o}_4, \boldsymbol{o}_6, \boldsymbol{o}_7$ $\qquad$ $h_2^2 \models \boldsymbol{o}_2, \boldsymbol{o}_5, \boldsymbol{o}_8$

H1 Nur die Kombinationen (h_1^1, h_2^1) und (h_1^2, h_2^2) bilden konzeptuelle Partitionen

$$f(h_1^1) + f(h_2^1) \;=\; 5/18 + 3/12 \;=\; 38/72$$
$$f(h_1^2) + f(h_2^2) \;=\; 5/24 + 3/12 \;=\; 33/72$$

Beispiel — Gebrauchtwagenhandel

Zweiter Iterationsschritt

P2 Objekte und ihr Median in h_1^1

Attribut	o_1	o_4	o_5	o_7	o_8	Modus
x_1	h	l	m	h	m	h, m
x_2	r	b	b	b	r	b
x_3	h	rh	rl	rh	rh	rh

⇨ Zentroid ist o_7

Die Hypothese h_2^1 hat Pseudomedian (m, g, l) und Median o_6

⇨ neue Gruppenprototypen sind $z_1 = o_7, z_2 = o_6$

S2 Stern von o_7:

$$h_1^1 = (x_3 = h, rh)$$
$$h_1^2 = (x_1 = h, m) \wedge (x_2 = r, b) \wedge (x_3 = r, rh)$$

Stern von o_6:

$$h_2^1 = (x_1 = m, l) \wedge (x_3 = rl, l)$$

G2 Dann gilt

$h_1^1 \models o_1, o_3, o_4, o_7, o_8$ $\qquad$ $h_2^1 \models o_2, o_5, o_6$

$h_1^2 \models o_1, o_7, o_8$

H2 Nur die Kombination (h_1^1, h_2^1) bildet eine konzeptuelle Partition

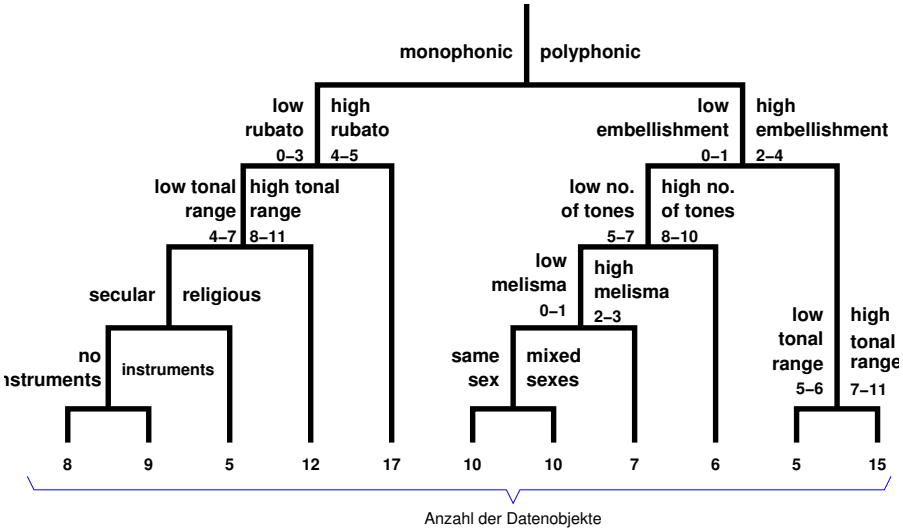
$$f(h_1^1) + f(h_2^1) = 5/18 + 3/12 = 38/72$$

Beispiel — Taxonomie spanischer Volkslieder

Gattungsdendrogramm nach konzeptueller Division

100 Lernbeispiele spanischer Volkslieder

22 Attribute mit nominalen / ordinalen Wertebereichen



Google Page Rank

„Gute Webseiten werden von guten Webseiten erwähnt."

Relevanz und Qualität

Seitenbewertung = Anfragepassung + Seriositätsmaß

$$\text{score}_q^{\text{Google}}(\text{doc}) = \text{Rel}_q(\text{doc}) + \text{rank}(\text{doc})$$

Worldwide Web als gerichteter Graph

Adjazenzmatrix $\boldsymbol{A} \in \{1, 0\}^{T \times T}$ mit $a_{st} = 1$ ⇔ $\text{doc}_i \mapsto \text{doc}_j$

Irrfahrtmodell

Der „*Random Surfer*" besucht Webseiten mit W'keit p_j und der Politik

$$p_j = (1 - \beta) \cdot \frac{1}{T} + \beta \cdot \sum_i p_i \cdot a_{ij} \cdot {}^{1}\!/\!_{\sum_k a_{ik}}$$

Die **Gleichgewichtsverteilung** gehorcht einer Eigenwertaufgabe ($\lambda = 1$):

$$\boldsymbol{B} \cdot \boldsymbol{p} = \left((1 - \beta) \cdot \frac{\mathbf{1}}{T} + \beta \cdot \tilde{\boldsymbol{A}}\right) \cdot \boldsymbol{p} = \boldsymbol{p} = \lambda \cdot \boldsymbol{p}\,, \qquad \tilde{a}_{ij} \stackrel{\text{def}}{=} {}^{a_{ij}}\!/\!_{\sum_k a_{ik}}$$

Schnitte in gewichteten Graphen

Definition

Sei $(\mathcal{K}, \mathcal{E}, \boldsymbol{A})$ ein ungerichteter, gewichteter Graph mit nicht-negativer, symmetrischer **Affinitätsmatrix** $\boldsymbol{A}$. Für zwei Knotenmengen $\mathcal{B}$, $\mathcal{C}$ sei

$$\ell_{\text{aff}}(\mathcal{B}, \mathcal{C}) \;=\; \sum_{s \in \mathcal{B}} \sum_{t \in \mathcal{C}} A_{st} \;, \qquad \tilde{\ell}_{\text{aff}}(\mathcal{B}, \mathcal{C}) \;=\; \frac{\ell_{\text{aff}}(\mathcal{B}, \mathcal{C})}{\ell_{\text{aff}}(\mathcal{B}, \mathcal{K})}$$

definiert. Eine Menge $\mathcal{C} \subset \mathcal{K}$ mit minimalem $\ell_{\text{aff}}(\mathcal{C}, \mathcal{K}\backslash\mathcal{C})$ bzw. mit minimalem $\tilde{\ell}_{\text{aff}}(\mathcal{C}, \mathcal{K}\backslash\mathcal{C})$ heißt **Schnitt** oder **normierter Schnitt**. Eine Partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ von $\mathcal{K}$ mit minimalem

$$\tilde{\ell}_{\text{aff}}(\{\mathcal{C}_\kappa\}_{\kappa=1}^{K}) \;\stackrel{\text{def}}{=}\; \frac{1}{K} \cdot \sum_{\kappa=1}^{K} \tilde{\ell}_{\text{aff}}(\mathcal{C}_\kappa, \mathcal{K}\backslash\mathcal{C}_\kappa)$$

heißt **normierter** K**-Schnitt**.

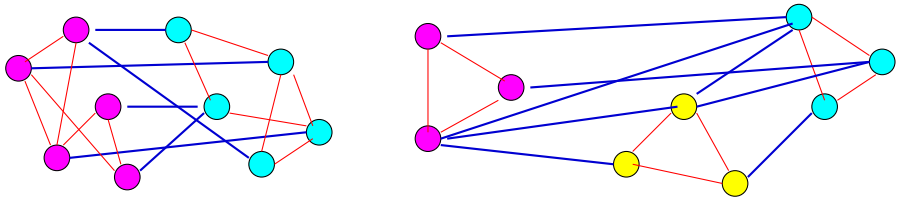
Bemerkung

Für die Affinitätsmatrix $\boldsymbol{A}$ gilt $A_{ss} = 0$ und $A_{st} = A_{ts}$ für alle $s, t \in \{1, \ldots, T\}$.

Gewöhnliche & normierte Schnitte

Fakt

Die Berechnung eines (normierten) 2-Schnittes ist beweisbar NP-hart.



Gewöhnlicher Schnitt

Dichotome Partition von $\mathcal{K}$ mit **minimaler Mengenaffinität**:

Summe der Querverbindungsgewichte zwischen $\mathcal{C}$ und $\mathcal{K}\backslash\mathcal{C}$

Normierter Schnitt

Minimale relative Mengenaffinität:

Proportion der Querverbindungsgewichte zu den Gewichten aller $\mathcal{C}$ verlassenden Kanten

Normierter K-Schnitt

Minimale Summe aller relativen Affinitäten zwischen den Schnittmengen $\mathcal{C}_\kappa$ und ihren **Komplementen** $\mathcal{K}\backslash\mathcal{C}_\kappa$

K-NC als Spurmaximierung

„K-way normalized cut“

Matrixalgebraische Formulierung

Die **Indikatormatrix** $\boldsymbol{C} \in \{1, 0\}^{T \times K}$ beschreibt die Zugehörigkeit der Knoten v_t zu den Gruppen $\mathcal{C}_\kappa$:

$$C_{t\kappa} \;\stackrel{\text{def}}{=}\; \begin{cases} 1 & t \in \mathcal{C}_\kappa \\ 0 & t \notin \mathcal{C}_\kappa \end{cases}$$

Die **Diagonalmatrix** $\boldsymbol{D} \in \mathbb{R}^{T \times T}$ enthält je Knoten die Summe seiner ausgehenden (einlaufenden) Kantengewichte:

$$\boldsymbol{D} \;=\; \text{diag}(\{d_s\}) \;, \qquad d_s \;\stackrel{\text{def}}{=}\; \sum_{t=1}^{T} A_{st}$$

Lemma

Das K-NC Kriterium ist äquivalent zur Maximierung der Größe

$$1/K \cdot \text{spur}\left(\boldsymbol{Z}^\top \boldsymbol{A} \boldsymbol{Z}\right) \qquad \textit{mit} \qquad \boldsymbol{Z} \;\stackrel{\text{def}}{=}\; \boldsymbol{C} \cdot (\boldsymbol{C}^\top \boldsymbol{D} \boldsymbol{C})^{-1/2} \;.$$

Relaxationslösung

F. Chung: *Spectral Graph Theory*, AMS 1997

Skalierung

Die umskalierte Matrix $\tilde{\boldsymbol{Z}} := \boldsymbol{D}^{1/2} \boldsymbol{Z} \in \mathbb{R}^{T \times K}$ besitzt offenbar orthonormale Spalten:

$$\tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}} \;=\; \boldsymbol{E} \;\in\; \mathbb{R}^{K \times K}$$

Gelockerte Bedingung

Ersetze die komplizierte Strukturforderung für $\boldsymbol{Z}$ durch Orthonormalitätsbedingung (*links*).

➪ Spurmaximierung durch Berechnung der K ersten Eigenvektoren

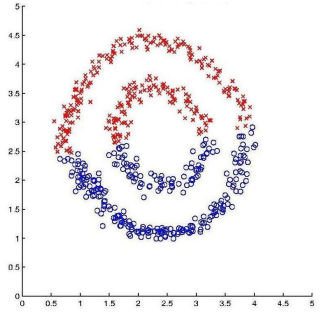
Lemma

Die Matrix $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{T \times K}$ mit den K oberen Eigenvektoren von $\tilde{\boldsymbol{A}} := \boldsymbol{D}^{-1/2} \boldsymbol{A} \boldsymbol{D}^{-1/2}$ als Spalten maximiert die Spur

$$\text{spur}\left(\underbrace{\tilde{\boldsymbol{Z}}^\top \cdot \boldsymbol{D}^{-1/2}}_{\boldsymbol{Z}^\top} \cdot \boldsymbol{A} \cdot \underbrace{\boldsymbol{D}^{-1/2} \cdot \tilde{\boldsymbol{Z}}}_{\boldsymbol{Z}}\right)$$

unter der Bedingung $\tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}} = \boldsymbol{E}$.

Spektrale Gruppierung statt K-means



2 Ringwolken — 2-means-Gruppierung

K-means Algorithmus

modelliert ausschließlich **konvexe** Ballungsgebiete und findet nur **lokale** Verzerrungsminima.

1. Bestimme $\tilde{\boldsymbol{A}}$
2. Berechne $\tilde{\boldsymbol{Z}}$ (EWP)
3. Ermittle $\boldsymbol{Z} = \boldsymbol{D}^{-1/2}\tilde{\boldsymbol{Z}}$
4. Errate (?!) $\boldsymbol{C}$ aus $\boldsymbol{Z}$

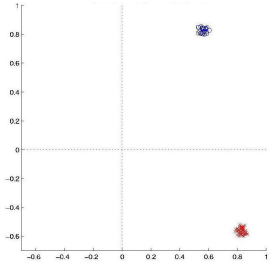
Spektrale Gruppierung

Die metrische Struktur der Datenobjekte wird in einen gewichteten Graphen transformiert; anschließend wird der Graph in Polynomialzeit durch Berechnung eines Semi-Schnittes in K Teilgraphen partitioniert.

Gruppieren im Spektralraum

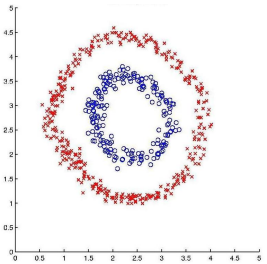
des Ähnlichkeitsgraphen

$\mathbb{R}^{T\times N}$ —Affinität→ $\mathbb{R}^{T\times T}$ —Eigenraum→ $\mathbb{R}^{T\times K}$ —Gruppen→ $\mathbb{R}^{K\times K}$



Zeilen der $\mathbb{R}^{T\times 2}$-Matrix

Gruppen-ID



Spektrale Gruppen ($K = 2$)

Eigenraummatrix $\boldsymbol{U} \in \mathbb{R}^{T\times K}$

Die K-dimensionalen Zeilenvektoren weisen eine hochdiskriminante Gruppenstruktur (innerer Ring — äußerer Ring) auf.

Ng-Jordan-Weiss Algorithmus

⟨Algorithmus⟩

1. AFFINITÄTSMATRIX
   $\boldsymbol{A} \in \mathbb{R}^{T\times T}$ mit $A_{st} \stackrel{\text{def}}{=} \begin{cases} 0 & s = t \\ \exp\{-\|\boldsymbol{x}_s - \boldsymbol{x}_t\|^2/2\sigma^2\} & s \neq t \end{cases}$
2. LAPLACEMATRIX
   $\boldsymbol{L} \in \mathbb{R}^{T\times T}$ via Normierung $L_{st} \stackrel{\text{def}}{=} \frac{A_{st}}{\sqrt{\rho_s\cdot\rho_t}}$, $\rho_s \stackrel{\text{def}}{=} \sum_{r=1}^{T} A_{rs}$
3. K FÜHRENDE EIGENVEKTOREN
   $$\boldsymbol{L} = \sum_{t=1}^{T} d_t^2 \cdot \boldsymbol{u}_t\boldsymbol{u}_t^\top, \qquad \boldsymbol{U} \stackrel{\text{def}}{=} (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_K) \in \mathbb{R}^{T\times K}$$
4. ZEILENWEISE NORMIERUNG
   $$\tilde{\boldsymbol{U}} \stackrel{\text{def}}{=} \boldsymbol{C}^{-1/2}\cdot\boldsymbol{U}, \qquad C_{st} = \begin{cases} \sum_{\kappa=1}^{K} U_{t\kappa}^2 & s = t \\ 0 & s \neq t \end{cases}$$
5. K-MEANS GRUPPIERUNG
   der Matrixzeilen $\tilde{\pi} : \{\tilde{\boldsymbol{v}}_1,\ldots,\tilde{\boldsymbol{v}}_T\} \to \{1,2,\ldots,K\}$
6. PARTITIONIERUNG
   der Originaldaten $\pi : \begin{cases} \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\} & \to \{1,2,\ldots,k\} \\ \boldsymbol{x}_t & \mapsto \tilde{\pi}(\tilde{\boldsymbol{v}}_t) \end{cases}$

⟨Algorithmus⟩

Warum funktioniert der NJW-Algorithmus ?

Beobachtung (geodätische Gruppenbildung)

Die Zeilen der Matrix $\boldsymbol{U}$ bilden eine K-dimensionale Repräsentation der Daten, in der Objekte mit kurzem Verbindungsweg — geodätisch, nicht Luftlinie — nahe beieinander liegen. warum?

Idealtypisches Szenarium ($K = 3$)

Objekte unterschiedlicher Gruppe besitzen den euklidischen Abstand ∞.

- Affinitätsmatrix und Laplacematrix sind von **Blockdiagonalform** (geeignete Nummerierung der $\boldsymbol{x}_t$ vorausgesetzt)

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_3 \end{pmatrix}, \qquad \boldsymbol{L} = \begin{pmatrix} \boldsymbol{L}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{L}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{L}_3 \end{pmatrix}$$

- Die Eigenvektoren von $\boldsymbol{L}$ sind die *mit Nullen aufgefüllten* Eigenvektoren der Blockmatrizen $\boldsymbol{L}_\kappa$.
- Dank der Doppelnormierung von $\boldsymbol{L}$ besitzt jeder Block *genau einen* maximalen **Eigenwert Eins**.

Warum funktioniert der NJW-Algorithmus ?

Idealtypisches Szenarium ($K = 3$)

Die K Haupteigenvektoren von $\boldsymbol{L}$ rekrutieren sich aus den K **Blockgewinnern**.

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{u}_1' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{u}_2' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{u}_3' \end{pmatrix} \in \mathbb{R}^{T \times 3}, \qquad \tilde{\boldsymbol{U}} = \begin{pmatrix} \mathbf{100} \\ \mathbf{010} \\ \mathbf{001} \end{pmatrix} \in \mathbb{R}^{T \times 3}$$

- Die Matrix $\boldsymbol{U}$ besitzt die K Eigenvektoren als Spalten.
- Alle Zeilen von $\boldsymbol{U}$ enthalten **genau einen** Eintrag ungleich Null.
- In der *zeilensummennormierten* Matrix $\tilde{\boldsymbol{U}}$ wird der Eintrag zur Eins.
- Das Gruppieren der Zeilen ($\approx$ Einheitsvektoren) ergibt zwangsläufig genau die richtigen Cluster!

Hierarchisch K-Means EM Relational Konzeptuell Spektral Clustergüte Σ

Warum funktioniert der NJW-Algorithmus ?

Denkfehler

Die K Haupteigenvektoren besitzen den gemeinsamen Eigenwert Eins.

- Sie sind also keineswegs eindeutig bestimmt und voller Nullen, sondern spannen lediglich einen eindeutig bestimmten K-dimensionalen Unterraum auf.

Rettung der Argumentation

Statt $\boldsymbol{U}$ erhalten wir $\boldsymbol{U}' = \boldsymbol{UR}$ mit einer Rotationsmatrix $\boldsymbol{R} \in \mathbb{R}^{K \times K}$.

Bezeichne $\boldsymbol{v}_t^\top$ die t-te Zeile von $\boldsymbol{U}$.

- $\boldsymbol{v}_t^\top \boldsymbol{R}$ ist die t-te Zeile von $\boldsymbol{U}'$ und für die Quadratnorm gilt:

$$(\boldsymbol{v}_t^\top \boldsymbol{R}) \cdot (\boldsymbol{R}^\top \boldsymbol{v}_t) = \boldsymbol{v}_t^\top \cdot (\boldsymbol{R}\boldsymbol{R}^\top) \cdot \boldsymbol{v}_t = \boldsymbol{v}_t^\top \cdot \boldsymbol{v}_t = \|\boldsymbol{v}_t\|^2$$

Die Zeilennormierung macht $\boldsymbol{UR}$ zu $\tilde{\boldsymbol{U}}\boldsymbol{R}$.

Zu clustern sind nicht mehr die **Einheitsvektoren** des $\mathbb{R}^K$, aber immerhin noch die Vektoren einer Orthonormalbasis $\boldsymbol{R}$ des Raumes.

Hierarchisch K-Means EM Relational Konzeptuell Spektral Clustergüte Σ

Details zum NJW-Algorithmus

Lineare Störungstheorie

Analyse des NJW **ohne** Intercluster-Distanzen $= \infty$

Stewart & Sun: *Matrix Perturbation Theory*, 1990

- **Eigengap** $\lambda_K - \lambda_{K+1}$ als untere Schranke der Gruppierungsstabilität

Abklingparameter $\sigma^2 > 0$

Minimale Endverzerrung nach dem K-means Clustering

- Skalarer Optimierungslauf für σ^2

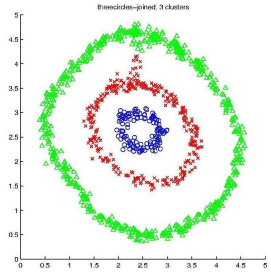
Startpartition für K-means

Die (idealen) Gruppenzentren liegen auf der Einheitssphäre.
Sie stehen paarweise senkrecht aufeinander.

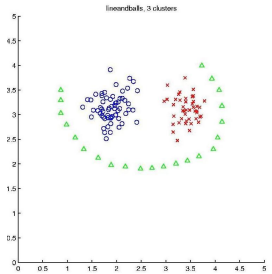
- Sukzessive Auswahl derjenigen $\boldsymbol{v}_t$ als Saatpunkte, die zu allen bereits selektierten Kandidaten maximal orthogonal sind.

Hierarchisch K-Means EM Relational Konzeptuell Spektral Clustergüte Σ

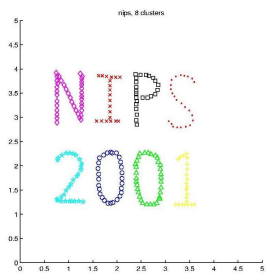
Beispiele — NJW & Wettbewerber



Ng-Jordan-Weiss, $K = 3$

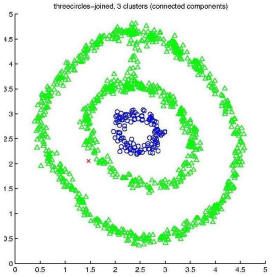


Ng-Jordan-Weiss, $K = 3$

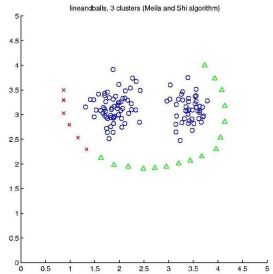


Ng-Jordan-Weiss, $K = 3$

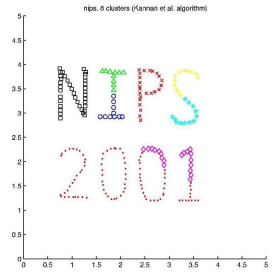
Ng-Jordan-Weiss



„connected components“



„random walk“



„geodesic K-means“

Die Konkurrenz

Sonstige spektral orientierte Verfahren

Gruppierung nach Zusammenhangskomponenten

Bilde den ungerichteten Graphen mit den Kanten

$$(s,t) \in \mathcal{E} \quad \Leftrightarrow \quad \|\boldsymbol{x}_s - \boldsymbol{x}_t\| \leq \theta \ .$$

Wähle die Schranke θ so, daß #(ZSH-Komponenten) $= K$ ist.

Gruppierung nach dem Irrfahrtprinzip (Meila & Shi)

Unterm Strich derselbe Ablauf wie beim NJW-Algorithmus, aber:
· Nur die *Zeilen* der Affinitätsmatrix $\boldsymbol{A}$ werden normiert.
· Die Zeilen der Eigenvektormatrix $\boldsymbol{U}$ werden *nicht* normiert.

Geodätischer K-means (Kannan & Vempala & Vetta)

Wiederum derselbe Ablauf wie beim NJW-Algorithmus, aber:
· Nur die *Zeilen* der Affinitätsmatrix $\boldsymbol{A}$ werden normiert.
· Die *Urbilder* der Clusterzentroide werden als *Repräsentanten* genutzt.

Dualisierte Berechnungen für K-means

Lemma

Sei $\omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\} \subset \boldsymbol{\Omega}$ und $\phi : \boldsymbol{\Omega} \to \mathbb{H}$ eine Expansion mit dem zugehörigen Kernoperator $K(\cdot,\cdot)$.

1. *Das Zentroidelement der termexpandierten Daten $\phi(\omega)$ bezüglich des quadratischen euklidischen Abstandes $\|\cdot\|_{\mathbb{H}}^2$ ist der Mittelwertvektor*
$$\mu \ = \ \frac{1}{T} \cdot \sum_{t=1}^{T} \phi(\boldsymbol{x}_t) \ .$$
2. *Der Abstand zwischen μ und einem expandierten Objekt $\phi(\boldsymbol{y})$, $\boldsymbol{y} \in \boldsymbol{\Omega}$, läßt sich mit $\mathrm{O}(T^2)$ Kernoperatorauswertungen berechnen.*
3. *Die Berechnung der Abstände von μ zu allen $\phi(\boldsymbol{x}_t)$, $t = 1, \ldots, T$, erfordert i.a. den Aufwand $\mathrm{O}(T^2 N)$, wenn $\boldsymbol{\Omega} = \mathbb{R}^N$ ist.*

K-means mit Termexpansion

Der Kernel Trick

Gruppieren in einem (impliziten) Expansionsraum $\phi(\Omega)$:

$$\phi : \boldsymbol{\Omega} \ \to \ \mathbb{H} \ , \qquad \langle \phi\boldsymbol{x}, \phi\boldsymbol{y} \rangle_{\mathbb{H}} = K(\boldsymbol{x}, \boldsymbol{y})$$

Der **Kernoperator** $K(\cdot,\cdot)$ simuliert das „*Rechnen*“ im RKHS $\mathbb{H}$.

Optimale K-Gruppierung in $\phi(\boldsymbol{\Omega})$

Ein Kodebuch $\{\mu_1, \ldots, \mu_K\}$ mit minimaler **Verzerrung**

$$\varepsilon(\{\omega_\kappa\}_\kappa) \ = \ \sum_{\kappa=1}^{K} \sum_{\boldsymbol{x} \in \omega_\kappa} \|\phi(\boldsymbol{x}) - \mu_\kappa\|^2$$

Berechnung von Gruppenzentr(oid)en μ_κ

$\mu_\kappa \in \mathbb{H}$ mittelt (endlich viele) expandierte Objekte.

Berechnung von Prototypdistanzen $\|f - g\|_{\mathbb{H}}^2$

f ist ein expandiertes Objekt.
g ist ein Gruppenzentrum.

Beweis.

1. $\mathbb{H}$ als Hilbertraum ist insbesondere auch ein Vektorraum.
2. Kernbasierte Distanzberechnung (speziell $\boldsymbol{y} = \boldsymbol{x}_r \in \omega$):
$$\begin{aligned}
\|\phi\boldsymbol{y} - \mu\|_{\mathbb{H}}^2 &= \left\| \phi\boldsymbol{y} - {}^1\!/_T \sum_{\boldsymbol{x}_t} \phi\boldsymbol{x}_t \right\|_{\mathbb{H}}^2 \\
&= \langle \phi\boldsymbol{y}, \phi\boldsymbol{y} \rangle - 2 \cdot \frac{1}{T} \cdot \sum_{\boldsymbol{x}_t} \langle \phi\boldsymbol{y}, \phi\boldsymbol{x}_t \rangle + \frac{1}{T^2} \cdot \sum_{\boldsymbol{x}_s, \boldsymbol{x}_t} \langle \phi\boldsymbol{x}_s, \phi\boldsymbol{x}_t \rangle \\
&= K(\boldsymbol{y}, \boldsymbol{y}) - 2 \cdot \frac{1}{T} \cdot \sum_{\boldsymbol{x}_t} K(\boldsymbol{y}, \boldsymbol{x}_t) + \frac{1}{T^2} \cdot \sum_{\boldsymbol{x}_s, \boldsymbol{x}_t} K(\boldsymbol{x}_s, \boldsymbol{x}_t) \\
&= \frac{1}{T^2} \cdot \left\{ T^2 \cdot G_{rr} - 2T \cdot \sum_{\boldsymbol{x}_t} G_{rt} + \sum_{\boldsymbol{x}_s, \boldsymbol{x}_t} G_{st} \right\} \\
&= G_{rr} - 2\bar{\boldsymbol{g}}_r + \bar{\boldsymbol{G}}
\end{aligned}$$
3. Jede Kernoperatorauswertung kostet $\mathrm{O}(N)$, die Berechnung der Gramschen Matrix kostet $\mathrm{O}(T^2 N)$, und die Mittelungen über $\boldsymbol{G}$ und ihre Zeilen $\boldsymbol{g}_r$ bleiben bei $\mathrm{O}(T^2)$.

□

Kernel K-means Algorithmus

Kostenpunkt: $\mathrm{O}(T^2N + T^2I))$

⟨Algorithmus⟩

1. STARTWERTE
   Startgruppierung $\{\omega_\kappa^{(0)}\}$ und Kernmatrix $\boldsymbol{G}$, $G_{st} = K(\boldsymbol{x}_s, \boldsymbol{x}_t)$.
2. NEUE OBJEKTZUGEHÖRIGKEIT
$$\kappa^*(\boldsymbol{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\lambda=1..K} \left\| \phi\boldsymbol{x} - \mu_\lambda^{(i)} \right\|$$
3. IMPLIZITE ZENTROIDBERECHNUNG
$$\omega_\lambda^{(i+1)} \leftarrow \{\boldsymbol{x}_t \mid \kappa^*(\boldsymbol{x}_t) = \lambda\}$$
   (keine *explizite* Berechnung der Mittelwertvektoren $\mu_\lambda^{(i+1)} \in \mathbb{H}$)
4. TERMINIERUNG
   Wenn $\varepsilon^{(i)}(\{\omega_\kappa\}) \leq \theta$ dann Ende sonst $i \leftarrow i+1$ und $\rightsquigarrow$ 2.

⟨zumEnde⟩

Gewichteter Kernel K-means

Gewichtetes Verzerrungskriterium

Objektabhängige Gewichtung der Zentrumsabstände:

$$\varepsilon(\{\omega_\kappa\}, \boldsymbol{w}) = \sum_{\kappa=1}^{K} \sum_{\boldsymbol{x}\in\omega_\kappa} w(\boldsymbol{x}) \cdot \|\phi(\boldsymbol{x}) - \mu_\kappa\|^2$$

- Gruppenzentroide $\hat{=}$ **gewichtete** Mittelwertvektoren.

Optimale Gruppenstruktur

Minimale Verzerrung $\Leftrightarrow$ maximale Spur (NC):

- Kernmatrix $\boldsymbol{W}^{1/2}\boldsymbol{G}\boldsymbol{W}^{1/2}$ korrespondiert mit normierter Affinität $\tilde{\boldsymbol{A}}$.
- Mantelmatrix $\boldsymbol{U}$ korrespondiert mit $\tilde{\boldsymbol{Z}}$.

  Jede Zeile von $\boldsymbol{U}$ ist ein skalierter K-Einheitsvektor.

- $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{E}_{(K)}$ ($\boldsymbol{U}$ besitzt orthonormale Spalten)

WKKM kann (im Relaxationssinne) auch durch NJW gelöst werden!

WKKM *vs.* normierter Schnitt

Minimale Verzerrung $\Leftrightarrow$ maximale NC-Matrixspur

Lemma

Die Gruppenverzerrung des WKKM ist gleichwertig zum Ausdruck

$$\varepsilon(\{\omega_\kappa\}, \boldsymbol{w}) = \operatorname{spur}\left(\boldsymbol{W}^{1/2}\boldsymbol{G}\boldsymbol{W}^{1/2}\right) - \operatorname{spur}\left(\boldsymbol{U}^\top \cdot \boldsymbol{W}^{1/2}\boldsymbol{G}\boldsymbol{W}^{1/2} \cdot \boldsymbol{U}\right)$$

mit

$$\begin{aligned}
\boldsymbol{\Phi} &= (\phi\boldsymbol{x}_1, \ldots, \phi\boldsymbol{x}_T) \\
\boldsymbol{G} &= \boldsymbol{\Phi}^\top\boldsymbol{\Phi} \\
\boldsymbol{W} &= \operatorname{diag}(\{w(\boldsymbol{x}_t)\}_t) \\
\boldsymbol{U} &= \operatorname{diag}(\{s_\lambda^{-1/2} \cdot \boldsymbol{W}_\lambda^{1/2} \cdot \boldsymbol{1}_\lambda\}) , \quad s_\lambda = \boldsymbol{1}_\lambda^\top \boldsymbol{W}_\lambda \boldsymbol{1}_\lambda .
\end{aligned}$$

Die ersten K Eigenvektoren der Matrix $\boldsymbol{W}^{1/2}\boldsymbol{G}\boldsymbol{W}^{1/2}$ liefern eine verzerrungsminimale Lösung unter der ***Relaxationsbedingung*** *$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{E}$ für die blockstrukturierte $(T \times K)$-Matrix $\boldsymbol{U}$.*

(Beweis durch exzessives Nachrechnen)

Spektrale Gruppierung

Vorzügliche und nachteilige Eigenschaften

NJW-Algorithmus

· einstufiges Verfahren
· metrische Distanzen
· K-means über $\mathbb{R}^K$
· Eigenvektoren $\mathrm{O}(T^2K)$
· Gruppen raten $\mathrm{O}(TK^2I)$
· Relaxationslösung !?!

WKKM-Algorithmus

· iteratives Verfahren
· Mercer-reskalierbar
· K-means über $\mathbb{H}$/dual
· Gram-Matrix $\mathrm{O}(T^2N)$
· dualer K-means $\mathrm{O}(T^2I)$
· Startpartition ??

Hybride spektrale Gruppierung

⟨Algorithmus⟩

1. Berechne ggf. die Affinitäten $\boldsymbol{A}$, Zeilensummen $\boldsymbol{D}$ und $\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$.
2. Startgruppierung $\{\omega_\kappa^{(0)}\}$ via **NJW-Algorithmus** auf $\tilde{\boldsymbol{A}}$.
3. Berechne die (virtuelle) Gram-Matrix $\boldsymbol{G} = \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{D}^{-1}$ und setze $\boldsymbol{W} = \boldsymbol{D}$.
4. Führe den i-ten **Weighted-Kernel-K-Means**-Doppelschritt durch.
5. Wenn $\varepsilon^{(i)}(\{\omega_\kappa\}) \leq \theta$ dann Ende sonst $i \leftarrow i+1$ und $\rightsquigarrow$ 4.

⟨zumEnde⟩

Gütemaße für Gruppen & Partitionen

Fragen über Fragen

- Mit welcher Vorgabeordnung $K \in \mathbb{N}$ **starte** ich K-means ?
- Welche Zerlegungsebene **wähle** ich als Resultat aus? (`agnes`, `diana`, `pam` & Co.)
- Zerfällt ω_κ in **noch kleinere** Gruppen ?
- Trifft gelernte Partition $\{\hat{\omega}_\kappa\}_\kappa$ die **wahre** Gruppenstruktur ?

Problem

*Weder **Gruppenverzerrungen** $\varepsilon(\omega_\kappa)$*
*noch **Gesamtverzerrung** $\varepsilon(\{\omega_\kappa\}_\kappa)$*
beantworten auch nur eine dieser Fragen !

Cluster Recovery Index

Externe Qualitätskriterien — Vergleich mit „*Goldstandard*“

Ausgangssituation

Objektmenge $\{o_1, \ldots, o_T\}$ mit **wahrer** und **hypothetischer** Gruppierung:

$$\omega_1^* \uplus \omega_2^* \uplus \ldots \uplus \omega_{K^*}^* \qquad \textit{versus} \qquad \hat{\omega}_1 \uplus \hat{\omega}_2 \uplus \ldots \uplus \hat{\omega}_{\hat{K}}$$

Aufgabenstellung

Berechnen eines Übereinstimmungsmaßes zwischen $\{\omega_\kappa^*\}$ und $\{\hat{\omega}_\kappa\}$.

- Vergleich einer Ist-Lösung mit der Soll-Lösung
- Vergleich zweier Lösungen zweier Methoden

Problem

*Gruppierungen sind nur eindeutig bis auf **Indexpermutation**.*

Der Rand-Index

Überschneidungsfreie, scharfe Gruppen (W. M. Rand, 1971)

Kreuzadjazenzstatistiken

Die Objekte $o_1, \ldots, o_T$ bilden $M = \binom{T}{2}$ ungeordnete Paaren $\{o_s, o_t\}$.

	gleiche $\{\hat{\omega}_\kappa\}$-Gruppe	verschiedene $\{\hat{\omega}_\kappa\}$-Gruppen	
gleiche $\{\omega_\kappa^*\}$-Gruppe	M_{11}	M_{10}	$M_{1\cdot}$
verschiedene $\{\omega_\kappa^*\}$-Gruppen	M_{01}	M_{00}	$M_{0\cdot}$
	$M_{\cdot 1}$	$M_{\cdot 0}$	M

Definition

Unter dem **Rand-Index** zweier scharfer Objektpartitionen verstehen wir den relativen Anteil

$$C_{\text{rand}} \stackrel{\text{def}}{=} \frac{M_{11} + M_{00}}{M}$$

der kohärent gruppierten Punktepaare $\{o_s, o_t\}$.

Der bereinigte Rand-Index

„adjusted Rand index“ (Hubert & Arabie, 1985)

Problem

⊕ *Maximum $C_{rand} = 1$ wird für äquivalente Partitionen angenommen.*
⊖ *Hohe C_{rand}-Werte entstehen auf Grund zufälliger Korrespondenzen.*

Definition

Unter dem **bereinigten Rand-Index** zweier scharfer Objektpartitionen verstehen wir den Quotienten

$$C_{\text{ari}} \stackrel{\text{def}}{=} \frac{\text{observed} - \text{expected}}{\text{maximum} - \text{expected}} = \frac{C_{\text{rand}} - \{M_{1.}M_{.1} + M_{0.}M_{.0}\} \,/\, M^2}{1 - \{M_{1.}M_{.1} + M_{0.}M_{.0}\} \,/\, M^2}$$

aus **beobachtetem** und **größtmöglichem** Übertreffen der allein zufallsbedingten Gruppenkohärenz.

Bemerkung

Ein *störbereinigtes* und *permutationsinvariantes* Vergleichsmaß ist auch die **Transinformation** $\mathcal{H}(\mathbb{K}_1) + \mathcal{H}(\mathbb{K}_2) - \mathcal{H}(\mathbb{K}_1, \mathbb{K}_2)$ zwischen dem wahren und dem hypothetischen Gruppenindex der Objekte.

Die Heterogenität einer Punktmenge

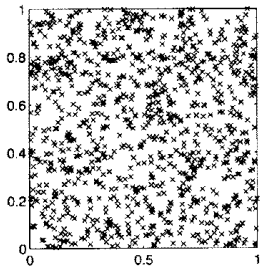
Zerfällt $\omega \subset \mathbf{\Omega}$ in noch kleinere Gruppen ?

Definition

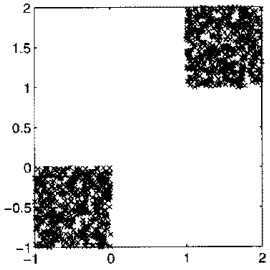
Für die Punktmenge ω mit der konvexen Hülle $\bar{\omega} \supset \omega$ ist der **Hopkins-Index** durch die relative Punktdichte

$$C_{\text{hop}} \stackrel{\text{def}}{=} \frac{\mathcal{E}_{\bar{\omega}}[d^*(\mathbb{X}, \omega)]}{\mathcal{E}_{\bar{\omega}}[d^*(\mathbb{X}, \omega)] + \mathcal{E}_{\omega}[d^*(\mathbb{X}, \omega^{(\mathbb{X})})]}$$

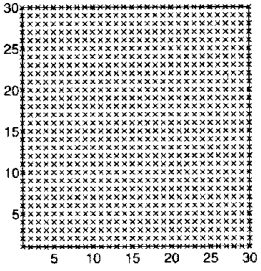
im Gesamtbereich der ω-Hülle definiert.



$C_{\text{hop}} = 0.4229$



$C_{\text{hop}} = 0.9988$



$C_{\text{hop}} = 0.1664$

Bestimmung des Hopkins-Index

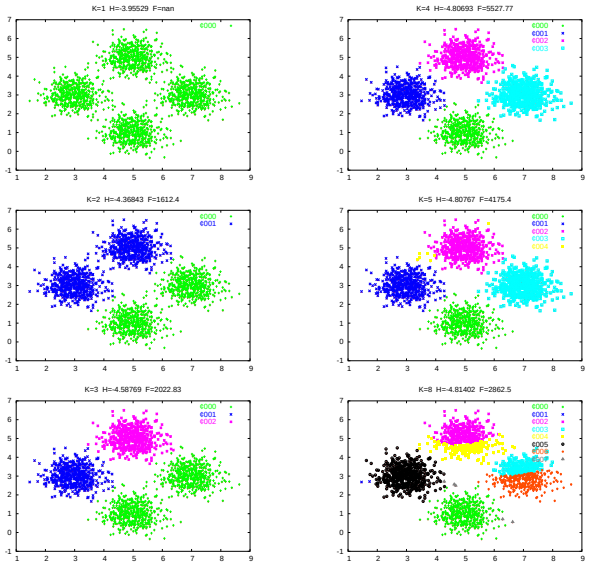
⟨Algorithmus⟩

1. RESAMPLING ω und $\bar{\omega}$
   Ziehe $S \in \mathbb{N}$ Vektoren $z_1, \ldots, z_S$ aus ω.
   Ziehe $S \in \mathbb{N}$ Vektoren $y_1, \ldots, y_S$ aus der konvexen Hülle
   $$\bar{\omega} \stackrel{\text{def}}{=} \left\{ \sum_{t=1}^{T} a_t x_t \mid \sum_{t=1}^{T} a_t = 1,\ a_t \geq 0 \right\}$$
2. PUNKTDICHTE IN $\bar{\omega}$
   Berechne die kumulative Punkt-Mengen-Distanz
   $$\mathcal{E}_{\bar{\omega}}[d^*(\mathbb{X}, \omega)] \approx D_{\bar{\omega}} = \sum_{s=1}^{S} \min_{x \in \omega} d(y_s, x)$$
3. PUNKTDICHTE IN ω
   Berechne die *„leave-one-out“* kumulative Punkt-Mengen-Distanz
   $$\mathcal{E}_{\omega}[d^*(\mathbb{X}, \omega^{(\mathbb{X})})] \approx D_{\omega} = \sum_{s=1}^{S} \min_{x \in \omega \setminus \{z_s\}} d(z_s, x)$$
4. AUSGABE
   Berechne den Quotienten $C_{\text{hop}} = D_{\bar{\omega}} \,/\, (D_{\bar{\omega}} + D_{\omega})$.

⟨Algorithmus⟩

Die beste Anzahl von Gruppen

Gretchenfrage $K \in \{1, 2, 3, 4, 5, 8\}$ bei der GMM-Identifikation

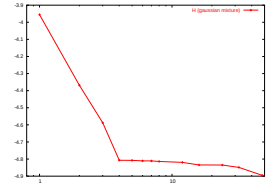


#1 — Raute^4

Vier sphärische Gaußdichten á 500 Punkte mit Varianz $\sigma = 0.5$

Verzerrungsmaß

Kreuzentropien für 15 Ordnungen $K \leq 96$



Ellenbogenheuristik

Konkavknick bei $K = 4$ nebst Sättigungsauslauf

Ordnung und Güte einer Gruppierung

Je mehr Gruppen — desto paßgenauer das Datenmodell

Lemma

Für einen Datensatz $\omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\} \subset \mathbb{R}^N$ *und die Ordnung* $K \in \mathbb{N}$ *bezeichne*

$$\varepsilon_{VQ}^{(K)}(\omega) \stackrel{def}{=} \underset{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_K}{argmin} \sum_{t=1}^{T} \min_{\kappa} \|\boldsymbol{x}_t - \boldsymbol{z}_\kappa\|^2$$

die minimale ***quadratisch-euklidische Verzerrung*** *der* K*-Partition und*

$$\ell_{GMM}^{(K)}(\omega) \stackrel{def}{=} \underset{\theta \in \mathcal{M}_K}{argmax} \sum_{t=1}^{T} \log \sum_{\kappa=1}^{K} \left\{ \pi_\kappa^{(\theta)} \cdot \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_\kappa^{(\theta)}, \boldsymbol{S}_\kappa^{(\theta)}) \right\}$$

die maximale Güte eines K*-****Mischverteilungsmodells*** *für den Datensatz. Dann gelten die Antitonie und die Monotonie*

$$K \le K' \quad \Rightarrow \quad \begin{cases} \varepsilon_{VQ}^{(K)}(\omega) & \ge & \varepsilon_{VQ}^{(K')}(\omega) \\ \ell_{GMM}^{(K)}(\omega) & \le & \ell_{GMM}^{(K')}(\omega) \end{cases} .$$

Der Pseudo-F-Wert

belohnt gute Gruppentrennung & bestraft große Gruppenanzahl

Problem

Die allermeisten Gruppierungskriterien ***verbessern sich systematisch*** *mit wachsender Gruppenzahl* K*, sind also zur Auswahl der Gruppenzahl völlig ungeeignet.*

Bemerkung

Die Monotonie ist in den meisten Kurven verletzt, denn K-means und GMM-Identifikation sind **lokale** Optimierungsverfahren.

Definition (Calinski-Harabasz)

Die Vergleichsgröße

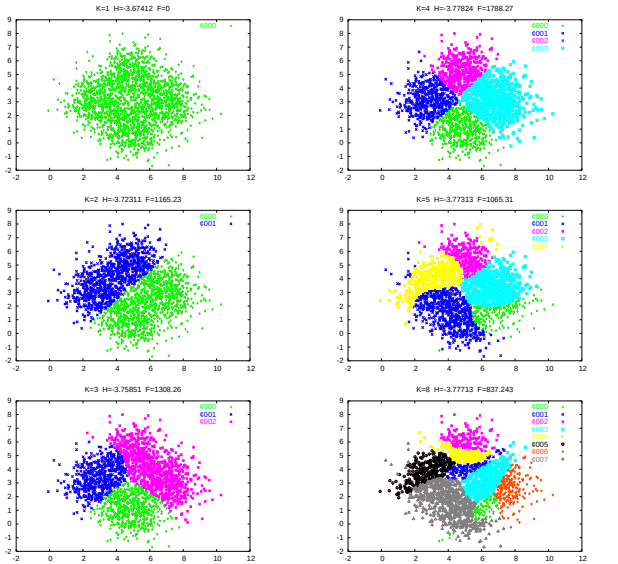
$$C_{\text{pseudo}}(\{\omega_1, \ldots, \omega_K\}) \stackrel{\text{def}}{=} \frac{\text{spur}\,(\boldsymbol{S}_B) \,/\, (K-1)}{\text{spur}\,(\boldsymbol{S}_W) \,/\, (T-K)}$$

heißt **Pseudo-F-Wert** der Gruppierung $\{\omega_1, \ldots, \omega_K\}$.

Es bezeichnen $\boldsymbol{S}_B$ die *Zwischengruppenstreuung* und $\boldsymbol{S}_W$ die *Innergruppenstreuung* der Datenpartition.

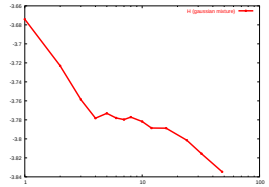
Die beste Anzahl von Gruppen

kann oft durch den Pseudo-F-Wert ermittelt werden

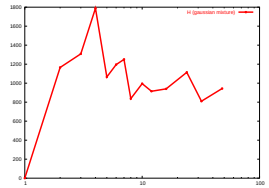


#2 — Raute[4]

Vier sphärische Gaußdichten mit Varianz $\sigma = 1.0$



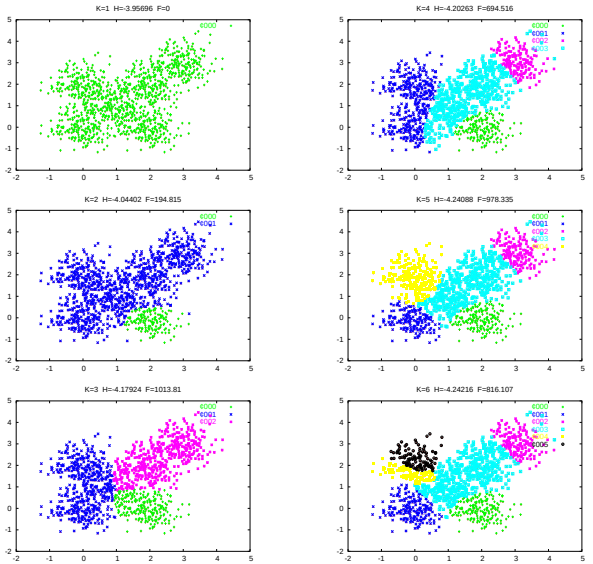
Negatives $\ell_{GMM}^{(K)}$ (EM)



Pseudo-F-Wert

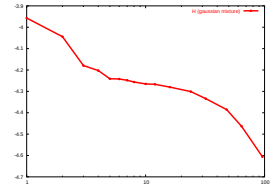
Hier versagt der Pseudo-F-Wert !

Der EM-Algorithmus zur Mischungsidentifikation findet nur suboptimale GMM

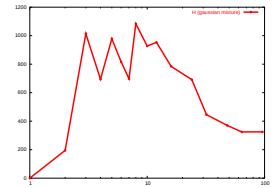


#3 - Kreuz[6]

Sechs Dichten in schrägliegender Kreuzform



Negatives $\ell_{GMM}^{(K)}$ (EM)



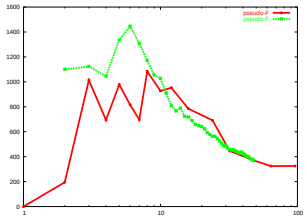
Pseudo-F-Wert

Dreiphasige Gruppierung

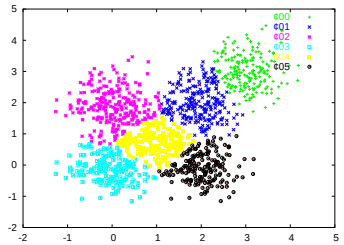
⟨Algorithmus⟩

1. DIVISIVE GRUPPIERUNG
   Sukzessive Zerlegung mit 2-means in $K_{\max} = 2^b$ Gruppen.
2. GMM-IDENTIFIKATION
   Austauschiteration mit Gaußschem Mischverteilungsmodell (EM-Schritte).
3. REAGGLOMERATION
   Bottom-up Gruppierung durch sukzessive ℓ_{GMM}-Maximierung.

⟨Algorithmus⟩



C_{pseudo} für EM & Reagglomeration



Korrekte Partition in $K = 6$ Gruppen

Regularisierte Gütemaße für scharfe Gruppierungen

„cluster validity index“

Dunn's ISODATA

$$C_{\text{iso}} \stackrel{\text{def}}{=} \frac{\min_{\kappa\neq\lambda} \min_{x\in\omega_\kappa} \min_{y\in\omega_\lambda} d(x,y)}{\max_\kappa \max_{x\in\omega_\kappa} \max_{y\in\omega_\kappa} d(x,y)}$$

Rousseeuw's Silhouette

$$C_{\text{sil}} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \frac{b_t - a_t}{\max(a_t, b_t)} \qquad \text{mit } D_{\kappa,t} = \overline{d(\omega_\kappa, x_t)} \text{ und } \begin{cases} a_t = D_{\kappa(t),t} \\ b_t = \min_{\lambda\neq\kappa(t)} D_{\lambda,t} \end{cases}$$

Expected Area Under Curve

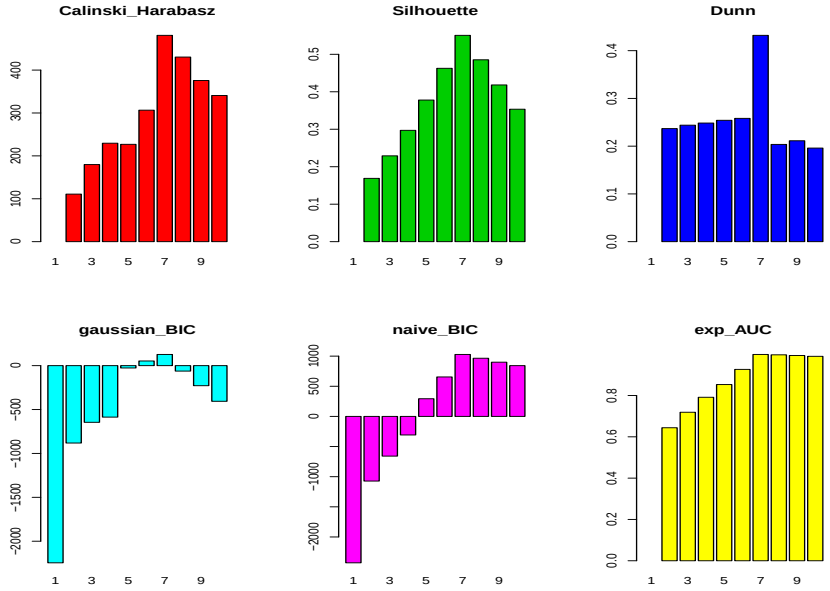
$$C_{\text{auc}} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \text{AUC}([d(x_\bullet, x_t)], [x_\bullet \in \omega_{\kappa(t)}])$$

Uni-/multivariat gaußsches BIC

$$C_{\text{bic}} \stackrel{\text{def}}{=} -\log \ell^{(K)}_{\text{GMM}} + \log(T)\cdot \text{df}(K,N) \quad \text{mit } \text{df}(K,N) = \begin{cases} K + 2NK & \text{(naiv)} \\ K + (N+3)\frac{NK}{2} & \text{(sonst)} \end{cases}$$

Beispiel: regularisierte Gütemaße im Vergleich

Sieben Cluster im $\mathbb{R}^7$ · Hierarchische Gruppierung für $K = 1, 2, \ldots, 10$



Gütemaße für unscharfe Gruppierungen

Tendenz zur monotonen Verbesserung mit der Gruppenanzahl K

Partitionskoeffizient

$$C_{\text{part}} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\sum_{\kappa=1}^{K} u_\kappa^2(x_t) \quad \stackrel{!}{\longrightarrow} \quad \max$$

Proportionsexponent

$$C_{\text{prop}} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \max_\kappa u_\kappa(x_t) \quad \stackrel{!}{\longrightarrow} \quad \max$$

Klassifikationsentropie

$$C_{\text{entro}} \stackrel{\text{def}}{=} -\frac{1}{T}\sum_{t=1}^{T}\sum_{\kappa=1}^{K} u_\kappa(x_t)\cdot \log_2 u_\kappa(x_t) \quad \stackrel{!}{\longrightarrow} \quad \min$$

Bemerkung

Partitionskoeffizient C_{part}, Proportionsexponent C_{prop} und Klassifikationsentropie C_{entro} nehmen ihre Optimalwerte (1/1/0) für die **scharfen** Gruppierungen $\{u_\kappa(\cdot)\}$ an.

Mischungsidentifikation in R

Ein Zoo konkurrierender Modelle — Kovarianzauslegung $\boldsymbol{S}_\kappa := \sigma_\kappa^2 \cdot \boldsymbol{U}_\kappa \boldsymbol{D}_\kappa \boldsymbol{U}_\kappa^\top$

Sphärische Modelle

EII $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma^2 \cdot \boldsymbol{E})$

VII $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma_\kappa^2 \cdot \boldsymbol{E})$

Global oder κ-variabel?

Volumen, Gesamtstreuung	σ_κ^2
Gestalt, Dynamik	$\boldsymbol{D}_\kappa$
Orientierung, Hauptachsen	$\boldsymbol{U}_\kappa$

Diagonale Modelle

EEI $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma^2 \cdot \boldsymbol{D}\ )$

VEI $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma_\kappa^2 \cdot \boldsymbol{D}\ )$

EVI $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma^2 \cdot \boldsymbol{D}_\kappa)$

VVI $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma_\kappa^2 \cdot \boldsymbol{D}_\kappa)$

Ellipsoidale Modelle

EEE $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma^2 \cdot \boldsymbol{U}\ \boldsymbol{D}\ \boldsymbol{U}^\top)$

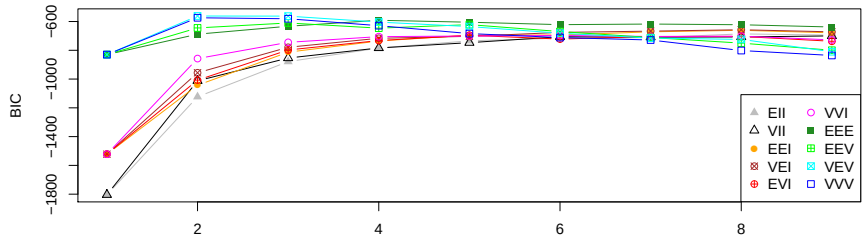
EEV $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma^2 \cdot \boldsymbol{U}_\kappa \boldsymbol{D}\ \boldsymbol{U}_\kappa^\top)$

VEV $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma_\kappa^2 \cdot \boldsymbol{U}_\kappa \boldsymbol{D}\ \boldsymbol{U}_\kappa^\top)$

VVV $\mathcal{N}(\boldsymbol{\mu}_\kappa,\ \sigma_\kappa^2 \cdot \boldsymbol{U}_\kappa \boldsymbol{D}_\kappa \boldsymbol{U}_\kappa^\top)$

Mischungsidentifikation in R

BIC-Entscheidung: optimales GM-Modell & optimale Gruppenzahl K



Agglomeratives Gruppieren

`hc (modelName, data, ...)` *'R'-Paket: `mclust`*

EM-Iteration

`em (modelName, data, parameters, ...)` *Start mit E-Schritt*

`me (modelName, data, z, ...)` *Start mit M-Schritt*

Bayes-Informationsmaß

`Mclust (data, G=1:9, modelNames)` *alle $K \in \{1, ..., 9\}$, alle GMM-Typen*

Zusammenfassung (5)

1. Das Ziel der **Gruppierung** (**Clusteranalyse**) ist die **unüberwachte** Zerlegung eines Datensatzes in **explizit** oder **implizit** charakterisierte Teilmengen von Objekten.
2. Die **hierarchischen** Verfahren arbeiten *bottom-up* (**agglomerativ**) oder *top-down* (**divisiv**); das Resultat ist ein **Gruppendendrogramm**.
3. Die **Austauschverfahren** geben eine **Anzahl** $K \in \mathbb{N}$ vor und erzeugen **scharfe** (*K-means*) oder **unscharfe** (*fuzzy K-means*) K-Partitionen.
4. Die **EM-Gruppierung** modelliert die Daten durch Identifikation einer gaußschen **Mischverteilung**.
5. Neben **sphärischen** Gruppen lassen sich auch **rangdefiziente** Ballungsgebiete ermitteln (**K-Elliptotypes** oder **Probabilistic PCA**).
6. **Relationale** Datensätze werden entweder *agglomerativ* gruppiert oder — wie auch **nominal** skalierte Objekte — mit einem **K-medoids**-Austauschverfahren (*RACE*, *K-Sterne*).
7. Die **spektrale** Methode löst eine **Minimalschnittaufgabe** im Affinitätsgraphen der Datenobjekte und läuft auf eine Art gaußschen **Kernel-K-means**-Algorithmus hinaus.
8. Ermittlung der **Clusteranzahl** durch **Ellenbogenheuristik** oder **Pseudo-F-Wert**.