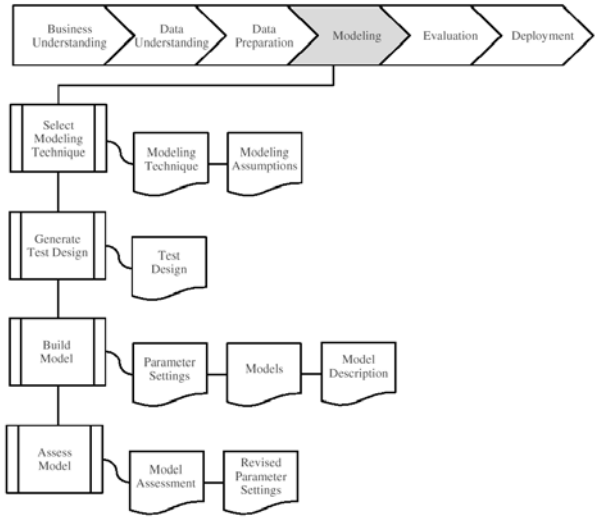
VII Unüberwachte Data-Mining-Verfahren

- Clusteranalyse
- Assoziationsregeln
- Generalisierte Assoziationsregeln mit Taxonomien
- Formale Begriffsanalyse
- Self Organizing Maps

© Institut AIFB, 2002.
Alle Rechte vorbehalten. Nachdruck oder photomechanische Wiedergabe nur mit Genehmigung des Verfassers.
Zuwiderhandlungen unterliegen den strafrechtlichen Bedingungen des Urheberrechtgesetzes.



VII.1 **Clusteranalyse**

VII.1.1 **Einleitung** (Bacher 1994)

- Zusammenfassung von Objekten in homogene Gruppen (Cluster, Klassen)
- Ziel dabei ist eine möglichst **große**
  - **Homogenität** innerhalb der Cluster
  - **Heterogenität** zwischen den Clustern

- geg. Menge von Objekten **kann** sich für Clusterbildung eignen, **muss** aber **nicht:**

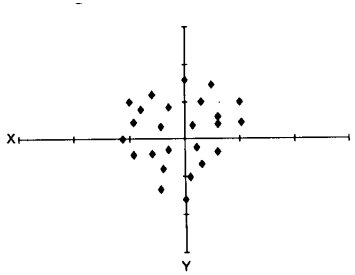


Abb. 1.1a: Eine Clusterstruktur ist nicht erkennbar. Die Klassifikationsobjekte bilden eine große Punktewolke.

- geg. Menge von Objekten **kann** sich für Clusterbildung eignen, **muss** aber **nicht:**

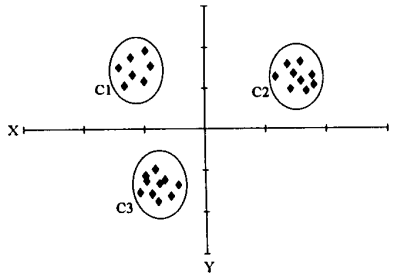


Abb. 1.1b: Es sind drei Cluster erkennbar. Beide Grundvorstellungen sind erfüllt. Die Cluster sind homogen und voneinander gut getrennt.

- geg. Menge von Objekten **kann** sich für Clusterbildung eignen, **muss** aber **nicht:**

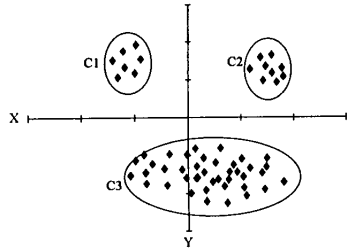


Abb. 1.1c: Es lassen sich drei Cluster erkennen. Cluster C3 ist aber sehr langgestreckt und erfüllt daher in einem geringeren Ausmaß die Vorstellung der Homogenität innerhalb der Cluster. Bei der Clusterbildung wurde der Heterogenität zwischen den Clustern ein größeres Gewicht beigemessen und das Cluster daher nicht getrennt.

- geg. Menge von Objekten **kann** sich für Clusterbildung eignen, **muss** aber **nicht:**

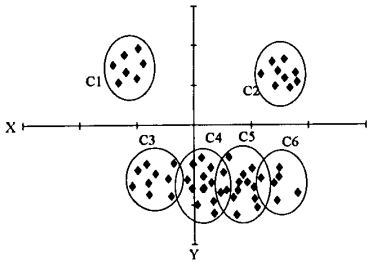


Abb. 1.1d: Im Unterschied zur Abbildung 1.1c wurde der Homogenität innerhalb der Cluster bei der Clusterbildung ein größeres Gewicht beigemessen. Das langgestreckte Cluster wurde daher in 4 überlappende Teilcluster zerlegt.

- Clusteranalyseverfahren unterscheiden sich u.a.
  - in den **Zuordnungsprinzipien**
    - exakte Zuordnung
    - probabilistische Zuordnung
    - possibilistische Zuordnung
  - in den benutzten **Informationen**
    - partielle Verfahren
      paarweiser Vergleich
    - globale Verfahren
      Distanz aller Objekte wird für Clusterbildung genutzt.
  - in der **Vorgehensweise**
    - hierarchisch
    - partitionierend
    - heuristisch
    - objective function based
    - begrifflich

- **Zuordnungsprinzipien**
  - **exakte Zuordnung**
    - Objekte werden mit Wahrscheinlichkeit 1 einem Cluster (**nicht-überlappende** Zuordnung) oder mehreren Clustern (**überlappende** Zuordnung) zugeordnet.
  - **probabilistische Zuordnung**
    - Objekte werden mit einer zwischen 0 und 1 liegenden Wahrscheinlichkeit einem oder mehreren Clustern zugeordnet
    - Verallgemeinerung der deterministischen Verfahren
  - **possibilistische Zuordnung**
    - Objekte werden über eine Zugehörigkeitsfunktion, die Werte zwischen 0 und 1 annehmen kann, jedem Cluster zu einem bestimmten Zugehörigkeitsgrad zugeordnet.

Vorgehensweise

- legt fest, nach welcher Vorgehensweise ein Cluster erzeugt wird.
- **Partitionierende Verfahren**
  - zufällig gewählte Anfangspartition (Menge nicht-überlappender Cluster) der zu clusternden Objekte wird schrittweise verbessert durch **Neuzuordnung** der Objekte in den Clustern
  - im folgenden betrachtet:
    - K-Means Verfahren
    - EM-Algorithmus
- **heuristische Vorgehensweise**
  - Dimensionalität der zu clusternden Objekte wird reduziert, um eine auf zwei bis drei Dimensionen reduzierte graphische Darstellung zu erreichen
- **objective function based**
  - kein prozedurales Vorgehen wie bei hierarchischen Verfahren
  - Basis bildet die Objektfunktion, die jedem Cluster einen Qualitätswert zuordnet

- **hierarchische Verfahren**
  hierarchische Verfahren werden unterschieden in
  - **agglomerative** Verfahren
    Cluster werden **bottom-up** erzeugt, ausgehend von **einelementigen** Clustern, den zu clusternden Objekten
  - **divisive Verfahren**
    Cluster werden **top-down** erzeugt, ausgehend von **einem** Cluster, das alle zu clusternden Objekte enthält

    divisive Verfahren waren in der Vergangenheit eher weniger bedeutend, gewinnen aber gerade für das Clustering von Dokumenten an Bedeutung

Hierarchisch agglomerativer Algorithmus

- bei n geg. Objekten werden (n-1) überlappungsfreie Clusterlösungen berechnet
- Algorithmus kann mit verschiedenen **Ähnlichkeitsmaßen** bzw. **Unähnlichkeitsmaßen** arbeiten, u.a.
  - **Complete Linkage**
  - **Single Linkage**

- **Complete Linkage**
  - Unähnlichkeit zwischen zwei Clustern wird durch das **Maximum** der paarweisen Unähnlichkeiten der Clusterelemente bestimmt:
  - für c_1, c_2 Cluster, d Abstandsmaß:

$$D(c_1, c_2) = \max_{x \in c_1, y \in c_2} d(x, y)$$

  - **hohe** Anforderungen an die Homogenität der zu bildenen Cluster

- **Single Linkage**
  - Unähnlichkeit zwischen zwei Clustern wird durch das **Minimum** der paarweisen Unähnlichkeiten der Clusterelemente bestimmt:
  - für c_1, c_2 Cluster, d Abstandsmaß:

$$D(c_1, c_2) = \min_{x \in c_1, y \in c_2} d(x, y)$$

  - **geringe** Anforderungen an die Homogenität der zu bildenen Cluster

- **Algorithmus** (*hierarchisch agglomerativ*)

Schritt 1: Jedes Klassifikationsobjekt bildet zu Beginn ein selbständiges Cluster. Setze daher die Clusterzahl K gleich der Klassifikationsobjektzahl n.

Schritt 2: Suche das Clusterpaar ({p},{q}) mit der größten Ähnlichkeit bzw. der geringsten Unähnlichkeit, verschmelze das Clusterpaar zu einem neuen Cluster {p,q} und reduziere die Clusterzahl K um 1 (K=K-1).

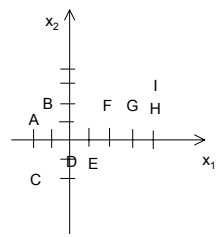
Schritt 3: Prüfe, ob K gleich 1 ist. Ist das der Fall, *beende* den Algorithmus, da alle Klassifikationsobjekte einem einzigen Cluster angehören. Bei nein fahre mit *Schritt 4* fort.

Schritt 4: Berechne die Ähnlichkeiten bzw. Unähnlichkeiten des neu gebildeten Clusters {p,q} zu den verbleibenden Clustern k.

Schritt 5: Gehe zu *Schritt 2*.

Beispiel (Bacher 1994): geg. Datenmatrix mit 9 Objekten und 2 Variablen

	Datenmatrix		Matrix der quadrierten euklidischen Distanzen								
	X1	X2	A	B	C	D	E	F	G	H	I
A	-2	1	0								
B	-1	2	2	0							
C	-1	-2	10	16	0						
D	0	-1	8	10	**2**	0					
E	1	-1	13	13	**5**	1	0				
F	2	2	17	**9**	25	13	10	0			
G	3	2	26	**16**	32	18	13	1	0		
H	4	2	37	**25**	41	25	18	4	1	0	
I	4	3	40	**26**	50	32	25	5	2	1	0



VII.1.2 K-Means Verfahren

- K-Means ist ein **partitionierendes, globales** Verfahren mit **exakter** Zuordnung, das **Clusterzentren** zur Clusterbildung verwendet

- **Grundidee:**
  - Annahme: Objekte *g* durch numerische Variablen *j* charakterisiert, d.h. jedes Objekt ist ein Punkt im $\mathbf{R}^m$
  - berechne die Clusterzentren für K Cluster derart, dass **Streuungsquadratsumme** in den Clustern ein **Minimum** ist.
  - sei *K* = Anzahl der zu bildenden Cluster (*k* = 1, ... , *K*)
    m = Anzahl der Variablen (*j* = 1, ... , *m*)
    x_{gj} = Wert der Variablen *j* für Objekt *g*
    $\bar{x}_{kj}$ = Clusterzentrum für Variable *j* im Cluster *k*

damit: $SQ_{in}(K) = \sum_k \sum_{g \in k} \sum_j (x_{gj} - \bar{x}_{kj})^2 \rightarrow \min \quad (*)$

- da für die quadrierte euklidische Distanz zwischen Objekt *g* und Clusterzentrum *k* gilt, dass

$$d^2_{g,k} = \sum_j (x_{gj} - \bar{x}_{kj})^2 \quad ,$$

kann Minimierungsaufgabe (*) spezifiziert werden als

$$SQ_{in}(K) = \sum_k \sum_{g \in k} d^2_{g,k} \rightarrow \min$$

- da die **Gesamtstreuungsquadratsumme** SQ_{ges} für eine geg. Objektmenge konstant ist, ergibt sich mit

$SQ_{in}(K)$: Streuungsquadratsumme in den Clustern
$SQ_{zw}(K)$: Streuungsquadratsumme zwischen den Clustern

$$SQ_{zw}(K) = SQ_{ges} - SQ_{in}(K)$$

- Minimierung von $SQ_{in}(K)$ ist gleichbedeutend zur **Maximierung** von $SQ_{zw}(K)$

$$SQ_{zw}(K) = SQ_{ges} - SQ_{in}(K)$$

$$\boldsymbol{SQ_{ges} = SQ_{zw}(K) + SQ_{in}(K)}$$

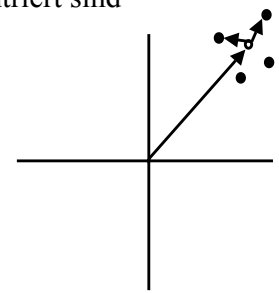
Ohne Einschränkung der Allgemeinheit:

Annahme, dass Daten um (0,...0) zentriert sind

$$\boldsymbol{SQ_{ges} = \sum_g \sum_j x^2_{gj}}$$

$$\boldsymbol{SQ_{in}(K) = \sum_k \sum_{g \in k} \sum_j (x_{gj} - \bar{x}_{kj})^2}$$

$$\boldsymbol{SQ_{zw}(K) = \sum_k \sum_{g \in k} \sum_j \bar{x}^2_{kj} = \sum_k |k| \bar{x}^2_k}$$



- K-Means Alogrithmus:
  (1) Lege **Clusteranzahl** *K* fest
  (2) Wahl von **Startwerten** für die Clusterzentren,
  z.B. zufällig gewählte Werte
  (3) **Zuordnung der Objekte** zu den Clusterzentren:
  - jedes Objekt *g* wird jenem Clusterzentrum *k* zugeordnet, zu dem die quadrierte euklidische Distanz minimal ist.

$$g \in k \Leftrightarrow k = \arg\min_{k'=1,\ldots,K} (d^2_{g,k'})$$

- damit: $SQ_{in}(K) = \sum_k \sum_{g \in k} d^2_{g,k}$ wird minimiert,
  da in jedem Schritt
  $SQ_{in}(K) = \sum_g \min_{k'=1,\ldots,K} d^2_{g,k'}$
  berechnet wird.

(4) **Neuberechnung der Clusterzentren:**

- nach der Zuordnung aller Objekte zu den *K* Clustern werden die Clusterzentren neu berechnet:
- sei n_{kj} = Zahl der Objekte des Clusters *k* mit gültigem Angaben in der Variablen *j*

damit: $\bar{x}_{kj} = \sum_{g \in k} \frac{x_{gj}}{n_{kj}}$

- $\bar{x}_{kj}$ ist **Mittelwert** für Variable *j* über alle Objekte *g* in Cluster *k*

(5) **Iteration**:

- sofern sich im Schritt (3) die Zuordnung der Objekte geändert hat, wird bei Schritt (3) fortgefahren; andernfalls endet der Algorithmus

Beispiel

(Bacher 1994)

- geg. Datamatrix mit 9 Objekten und 2 Variablen:

	Datenmatrix		Matrix der quadrierten euklidischen Distanzen								
	X1	X2	A	B	C	D	E	F	G	H	I
A	-2	1	0								
B	-1	2	2	0							
C	-1	-2	10	16	0						
D	0	-1	8	10	**2**	0					
E	1	-1	13	13	**5**	1	0				
F	2	2	17	**9**	25	13	10	0			
G	3	2	26	**16**	32	18	13	1	0		
H	4	2	37	**25**	41	25	18	4	1	0	
I	4	3	40	**26**	50	32	25	5	2	1	0

- Bildung von 3 Clustern (*K* = 3)

• Anwendung des K-Means Algorithmus auf geg. Objekte

			1. Iteration				2. Iteration				3. Iteration			
			Clusterzentren (Startwerte)				Clusterzentren (Startwerte)				Clusterzentren (Startwerte)			
			C1	C2	C3		C1	C2	C3		C1	C2	C3	
		X1	-2,00	-1,00	-1,00		-2,00	2,40	0,00		-1,50	3,25	0,00	
		X2	1,00	2,00	-2,00		1,00	2,20	-1,33		1,50	2,25	-1,33	
	X1	X2				Zuord.				Zuord.				Zuord.
A	-2	1	**0,00**	2,00	10,00	C1*	**0,00**	20,80	9,43	C1	**0,50**	29,13	9,43	C1
B	-1	2	2,00	**0,00**	16,00	C2*	**2,00**	11,60	12,09	C1*	**0,50**	18,13	12,09	C1
C	-1	-2	10,00	16,00	**0,00**	C3*	10,00	29,20	**1,45**	C3	12,50	36,13	**1,45**	C3
D	0	-1	8,00	10,00	**2,00**	C3*	8,00	16,00	**0,11**	C3	8,50	21,13	**0,11**	C3
E	1	-1	13,00	13,00	**5,00**	C3*	13,00	12,20	**1,11**	C3	12,50	15,63	**1,11**	C3
F	2	2	17,00	**9,00**	25,00	C2*	17,00	**0,20**	15,09	C2	12,50	**1,63**	15,09	C2
G	3	2	26,00	**16,00**	32,00	C2*	26,00	**0,40**	20,09	C2	20,50	**0,13**	20,09	C2
H	4	2	37,00	**25,00**	41,00	C2*	37,00	**2,60**	27,09	C2	30,50	**0,63**	27,09	C2
I	4	3	40,00	**26,00**	50,00	C2*	40,00	**3,20**	34,75	C2	32,50	**1,13**	34,75	C2
			neue Clusterzentren				neue Clusterzentren				neue Clusterzentren			
			C1	C2	C3		C1	C2	C3		C1	C2	C3	
			-2,00	2,40	0,00		-1,50	3,25	0,00		-1,50	3,25	0,00	
			1,00	2,20	-1,33		1,50	2,25	-1,33		1,50	2,25	-1,33	
			Zahl der Vertauschungen = 9				Zahl der Vertauschungen = 1				Zahl der Vertauschungen = 0			

(Bacher 1994)

• Erläuterungen:

- die Objekte A, B, C werden als Startwerte für Clusterzentren der Cluster C1, C2, C3 gewählt
- die restlichen Objekte werden jenem Cluster zugeordnet, zu dem es die kleinste quadrierte euklidische Distanz besitzt (fettgedruckte Werte)
- das neue Clusterzentrum für C2 ergibt sich in der 1. Iteration (C2 besteht aus den Objekten B, F, G, H, I):

$$\bar{x}_{21} = (-1+2+3+4+4)/5 = 2.40$$

$$\bar{x}_{22} = (2+2+2+2+3)/5 = 2.20$$

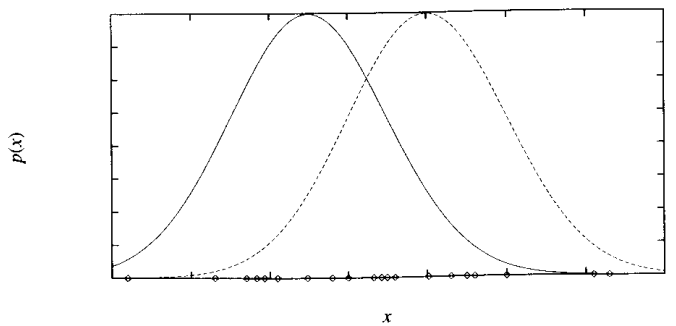
- in der 2. Iteration wird das Objekt B einem neuen Cluster zugeordnet: C1
- in der 3. Iteration tritt keine Veränderung der Zuordnung mehr auf, Algorithmus stoppt

- **Bemerkung:**
  - in jedem Iterationsschritt wird die Streuungsquadratsumme in den Clustern $SQ_{in}(K)$ kleiner oder bleibt gleich
  - Algorithmus findet für $SQ_{in}(K)$ ein **lokales Minimum.** D.h. das Ergebnis ist von den gewählten Startwerten abhängig!! D. h. diese sind geeignet auszuwählen und das Ergebnis ist ggf.hinterher kritisch zu hinterfragen.
  - diese Variante des K-Means Algorithmus wurde 1965 von Forgy entwickelt und wird deshalb auch als Forgy Methode bezeichnet
  - zu dieser Basis-Variante des Algorithmus existieren verschiedene Modifikationen
  - in K-Means können auch andere Distanzmaße verwendet werden (damit ist auch Behandlung nicht-numerischer Variablen möglich, wenn für diese die Durchschnittsbildung Bedeutung trägt. (Vorsicht z.B. bei Schulnoten!))
  - Der Aufwand pro Iteration ist linear in |G|, d.h. der Algorithmus hat geringe Komplexität, da nicht alle |G|² Distanzen berücksichtigt werden müssen.

VII.1.3 EM Algorithmus

- EM-Algorithmus (**Expected-Maximum-Likelihood-Estimator**) ist eine Verallgemeinerung des K-Means-Verfahrens:
  - Zugehörigkeit eines Objektes *g* zu einem Cluster *k* ist mit einer bestimmten Wahrscheinlichkeit gegeben
    ⇒ **probabilistisches** Verfahren
  - im Schritt (3) des K-Means-Algorithmus wird für jedes Objekt *g* die Zuordnungswahrscheinlichkeit zum Cluster *k* berechnet
  - im Schritt (4) werden die Klassenzentren $\bar{x}_{kj}$ als Maximum-Likelihood-Schätzung berechnet

- Annahmen für EM-Algorithmus:
  - den zu clusternden Objekten liegen *K* unbekannte (= nicht direkt beobachtbare) Klassen (Cluster) zugrunde (auch **latente** Klassen genannt)
  - diese Klassen erklären die beobachteten Variablen
  - jede Klasse *k* besitzt in jeder Variablen *j* eine **Normalverteilung** mit Mittelwert μ_{kj} (Klassenzentrum) und Varianz σ_{kj}^2 (d.h. μ_{kj} entspricht $\bar{x}_{kj}$ in K-Means)
  - in jeder Klasse *k* sind alle Variablen voneinander **unabhängig**



(Mitchell 1997)

Beispiel: (Mitchell 1997)

- *g* Objekte werden durch 2 latente Klassen erzeugt (*K* = 2)
- x_{gj} beobachteter Wert von Objekt *g* für Variable j
  (1 beobachtete Variable: j = 1
  im folgenden: x_{g1} vereinfacht zu x_g)
- beide Klassen haben Normalverteilung mit identischer Varianz σ^2
- gesucht ist Hypothese $h = (\mu_1, \mu_2)$, d.h. Mittelwerte der beiden Normalverteilungen
- für jeden Mittelwert μ ist Maximum-Likelihood-Schätzung seines Wertes gleichbedeutend mit Minimierung der quadrierten Fehler:

$$\sum_{i=1}^{g} (x_i - \mu)^2 \rightarrow \min \qquad (\text{für } \mu = \mu_1, \mu_2)$$

- *i*-tes Objekt kann beschrieben werden durch (x_i, z_{i1}, z_{i2}) mit
  x_i : beobachteter Wert von Objekt i (i = 1,..., g) (Instanz)
  z_{i1}, z_{i2} : geben an, welche der beiden Normalverteilungen zur Generierung von x_i verwendet wurden ist:
  z_{ie} : sind die nicht beobachteten Variablen
  $z_{ie} \in \{0,1\}$ für $e = 1,2$
  $z_{ie} = 1 : x_i$ ist mit Normalverteilung e erzeugt worden
- Werte von z_{ie} werden durch den EM-Algorithmus sukzessive geschätzt unter Verwendung der Hypothese $h = (\mu_1, \mu_2)$

Schritt (3): berechne Erwartungswert $E[z_{ie}]$ unter Verwendung der Hypothese *h*
Schritt (4): berechne neue Maximum-Likelihood-Schätzung $h' = (\mu_1', \mu_2')$ unter Verwendung der Erwartungswerte $E[z_{ie}]; h'$ wird neue Hypothese h

- $E[z_{ie}]$ ist Wahrscheinlichkeit, daß beobachteter Wert x_i von e-ter Normalverteilung erzeugt worden ist:

$$E[z_{ie}] = \frac{p(x = x_i \mid \mu = \mu_e)}{\sum_{n=1}^{2} p(x = x_i \mid \mu = \mu_n)} = \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_e)^2}}{\sum_{n=1}^{2} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}}$$

damit:
$E[z_{ie}]$ kann aus x_i, μ_1, μ_2 berechnet werden

- μ_e kann neu berechnet werden durch

$$\mu_e = \frac{1}{g} \sum_{i=1}^{g} E[z_{ie}] x_i \qquad (e = 1,2)$$

d.h. μ_e ist gewichteter Mittelwert der x_i

- EM-Algorithmus berechnet gesuchte Mittelwerte der Normalverteilungen **iterativ** - ausgehend von Startwerten für die Mittelwerte (Klassenzentren)
- Algorithmus stoppt, sobald die Verbesserung der Schätzwerte kleiner als ein vorgegebener Schwellwert ist
- EM-Algorithmus findet eine Hypothese h, bei der der Fehler ein **lokales Minimum** erreicht.
- EM-Algorithmus kann i.a. mit *k* Normalverteilungen und *m* beobachteten Variablen definiert werden, siehe (Mitchell 1997), (Bacher 1994)