Multivariate Verfahren

	Lineare Regression	Hauptkomponentenanalyse	Korrespondenzanalyse	Clusteranalyse	Diskriminanzanalyse
Zweck:					
Vorhersage	x				
Dimensionsreduktion		x	x		
Klassifizierung				x	x
Eigenschaften:					
nicht-linear					
verteilungsfrei			x	x	
nominal skalierte Var.			x		x

Clusteranalyse

Ziel: Aufteilung des Datensatzes in Gruppen (= Cluster), so dass

- die Unterschiede ***zwischen*** den einzelnen Gruppen möglichst ***groß*** sind
- die Unterschiede ***innerhalb*** der einzelnen Gruppen möglichst ***klein*** sind

Hauptkomponentenanalyse vs. Clusteranalyse

Hauptkomponentenanalyse (HKA):
Zusammenfassung von ***Variablen***, die dem gleichen "Prozess" zugeordnet werden können, zu Hauptkomponenten (Faktoren)

Clusteranalyse (CA):
Zusammenfassung von ***Messstellen***, die ähnliche Messwerte aufweisen, zu Clustern

	Messstelle 1	Messstelle 2	Messstelle 3	
Variable 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	} **HKA**
Variable 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	
Variable 3	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	
		CA		

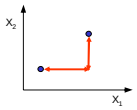
Ähnlichkeits-/Distanzmaße (I)

$x_{1,m}$, $x_{2,m}$, ... $x_{n,m}$: Werte für die Variablen x_1, x_2, ..., x_n der Messstelle m

in n Dimensionen

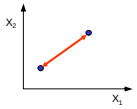
City-Block-Metrik (Manhattan-, Taxifahrer-Distanz):

$$d_{m=1,m=2} = \sum_{i=1}^{n} \left| x_{i,1} - x_{i,2} \right|$$



Euklidische Distanz:

$$d_{m=1,m=2} = \sqrt{\sum_{I=1}^{n} (x_{i,1} - x_{i,2})^2}$$



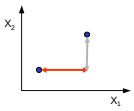
allgemein:
Minkowski-Distanz:
(L_p-Norm)

$$d_{m=1,m=2} = \sqrt[p]{\sum_{I=1}^{n} \left| x_{i,1} - x_{i,2} \right|^p}$$

Ähnlichkeits-/Distanzmaße (II)

$x_{1,m}$, $x_{2,m}$, ... $x_{n,m}$: Werte für die Variablen x_1, x_2, ..., x_n der Messstelle m

Tschebyscheff: $d_{m=1,m=2} = \max\left(\left|x_{i,1} - x_{i,2}\right|\right)$



Pearson-Korrelation: $d_{m=1,m=2} = corr(m=1, m=2)$

Mahalanobis-Distanz: $d_{m=1,m=2} = \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{n} c_{j,k} \cdot (x_{j,1} - x_{j,2}) \cdot (x_{k,1} - x_{k,2})}$

$\rightarrow$ hohe Werte für geringe Kovarianz entspricht der Euklidischen Distanz für

- z-transformierte und
- voneinander unabhängige (HKA!)

Variablen

$c_{j,k}$: Element der Inversen der Kovarianz-Matrix

Cluster-Algorithmen

Vorgehensweise:

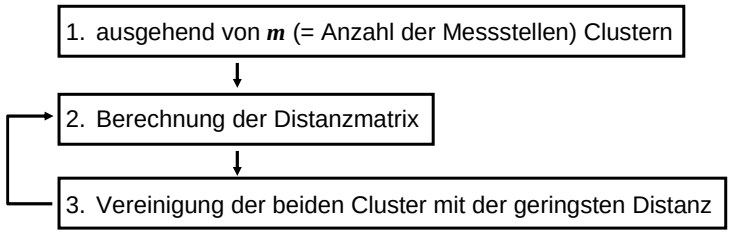
1. Sukzessive Aufteilung in Cluster: ***Hierarchische Clusteranalyse***

   2.1 Sukzessive Zusammenfassung von Clustern: ***agglomerativ***

   2.2 Sukzessive Aufspaltung in Cluster: ***divisiv***
   (faktisch unbedeutend)

2. Anzahl der Gruppen wird vorgegeben: ***Partitionierende Clusteranalyse***, ***Clusterzentrenanalyse***

Agglomeratives Verfahren



Vereinigung von Clustern (I)

Ziel: Aufteilung des Datensatzes in Gruppen (= Cluster), so dass

1. die Unterschiede ***zwischen*** den einzelnen Gruppen möglichst ***groß*** sind
2. die Unterschiede ***innerhalb*** der einzelnen Gruppen möglichst ***klein*** sind

Methodik: Quantifizierung der ***Unterschiede*** mittels:

1. ***Distanzen zwischen*** den Gruppen (Clustern)

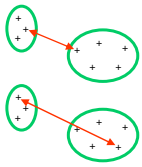
2. ***Streuung innerhalb*** der Gruppen (Cluster)

Vereinigung von Clustern (II)

→ Distanz zwischen ***ausgewählten*** Objekten der zu vergleichenden Cluster:

Single-Linkage (Nearest Neighbour): ***kleinste*** Distanz zwischen Objekten der beiden Cluster (→ *Bestimmung von „Ausreißern“*)

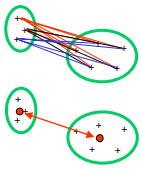
Complete-Linkage (Furthest Neighbour): ***größte*** Distanz zwischen Objekten der beiden Cluster



→ gemittelte Distanz ***aller*** Objekte:

Average Linkage ("mittlerer Nachbar"): Mittelwert aller Distanzen

Zentroid-Distanz: Distanz zwischen Cluster-Zentroiden (= Cluster-Mittelwerte)



Vereinigung von Clustern (III)

→ Heterogenitätsmaß:

Ward

- Vereinigung der Cluster, die zur geringstmöglichen Erhöhung der ***Fehlerquadratsumme*** (***Varianzkriterium***, V_g) im neuen Cluster g führt:

$$V_g = \sum_{k=1}^{K_g} \sum_{i=1}^{n} (x_{k,i} - \bar{x}_{i,g})^2 \quad \text{mit Gruppen-Mittelwert } \bar{x}_{i,g} = \frac{1}{K_g} \cdot \sum_{k=1}^{K_g} x_{k,i}$$

K_g: Anzahl der Objekte im Cluster g; n: Anzahl der Variablen

Eigenschaften der hierarchischen Linkage-Verfahren

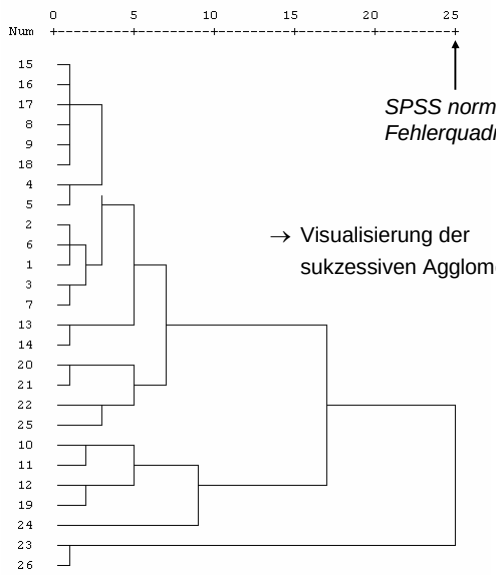
Verfahren	Eigenschaft	neigt zu
Single-Linkage	kontrahierend	Kettenbildung
Complete-Linkage	dilatierend	Bildung kleiner Gruppen
Average Linkage	konservativ	
Zentroid	konservativ	
Ward	konservativ	Bildung etwa gleich großer Gruppen

kontrahierend: Bildung weniger großer und vieler kleiner Gruppen (→ "Ausreisser")

dilatierend: Bildung in etwa gleich großer Gruppen

konservativ: weder kontrahierend noch dilatierend

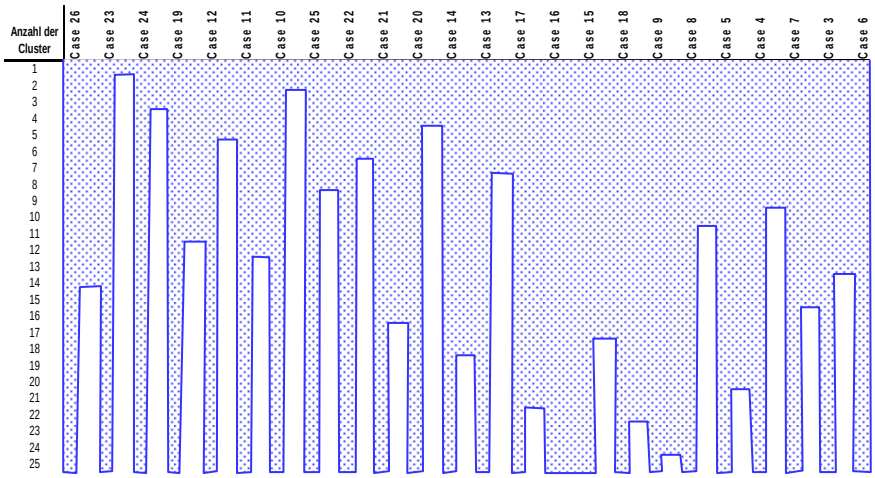
Dendrogramm



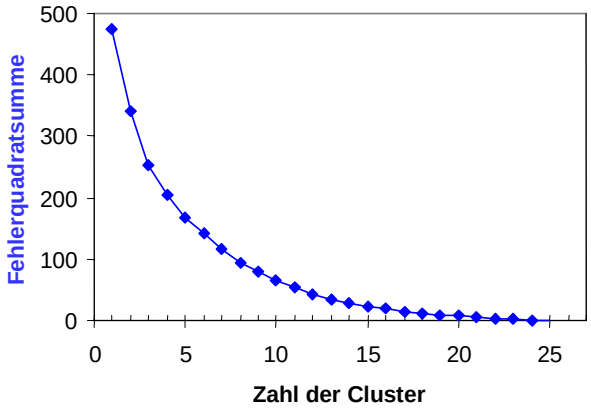
SPSS normiert die maximale Fehlerquadratsumme auf 25

→ Visualisierung der sukzessiven Agglomeration

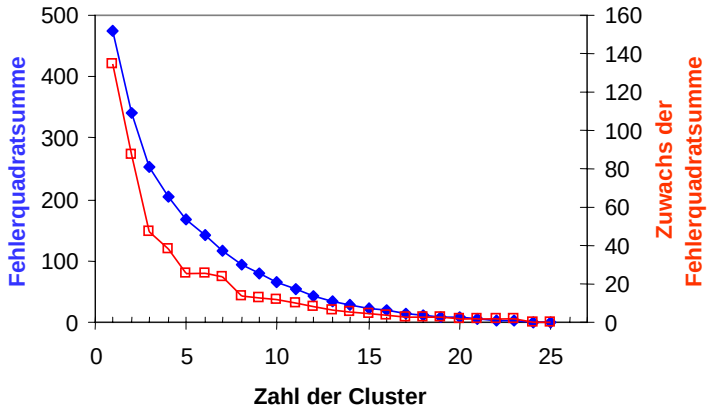
Eiszapfen-Diagramm



Screeplot: Bestimmung der "optimalen" Clusterzahl



Bestimmung der "optimalen" Clusterzahl



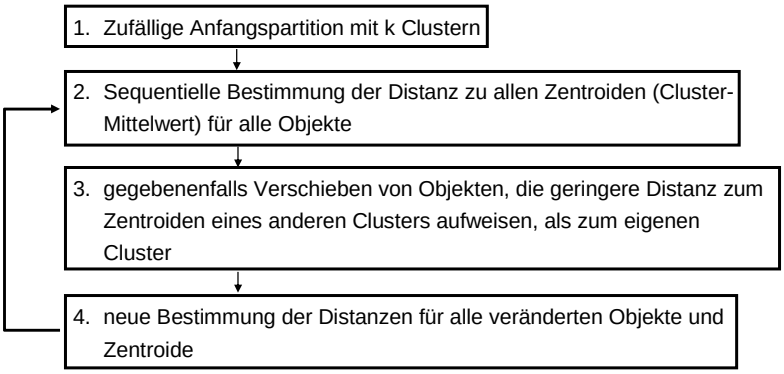
Partitionierendes Verfahren

1. ausgehend von Anfangspartition
2. Berechnung der Gruppenzentroide
3. Berechnung des Gütemaßes der Gruppeneinteilung
4. Verlagerung des Objektes, das zur maximal möglichen Verbesserung der Gruppeneinteilung führt

Partitionierendes Linkage-Verfahren:

K-Means

- Optimierung der Aufteilung in ***k*** Cluster
- Cluster werden jeweils durch ihre Zentroide (Mittelwerte = ***means***) repräsentiert



Empfohlene Vorgehensweise

1. Kompensation für unterschiedliche Wertebereiche und Streuungen der verschiedenen Variablen mittels ***z-Transformation***
2. Berücksichtigung der Korrelationen zwischen einzelnen Variablen durch Verwendung der ***Hauptkomponenten*** (alternativ: Verwendung der ***Mahalanobis-Distanz***)
3. Bestimmung der optimalen Clusterzahl mittels ***hierarchischer Clusteranalyse*** (Ward-Verfahren)
4. Optimierung der Clustereinteilung mittels ***partitionierender Clusteranalyse*** (K-Means-Verfahren)
5. Überprüfung der Clustereinteilung mittels ***Kreuzvalidierung*** bzw. ***Diskriminanzanalyse***

Aufgabe

1. Führen Sie eine **hierarchische** Clusteranalyse mit den **Hauptkomponentenwerten** durch (*Single Linkage, Complete Linkage, Ward*), bestimmen Sie die optimale Zahl der Cluster, und lassen Sie Sich für diese Clusterzahl die Clusterzugehörigkeiten ausgeben.

2. Führen Sie eine **partitionierende** Clusteranalyse mit der unter 1. bestimmten optimalen Zahl der Cluster durch. Interpretieren Sie die Cluster anhand der Clusterzentroid-Werte.

3. Wiederholen Sie die Analyse mit den z-transformierten Daten.

4. Vergleichen und bewerten Sie die mit den verschiedenen Verfahren vorgenommenen Klassifizierungen.

5. Vergleichen Sie die Ergebnisse mit denen der Hauptkomponenten- und der Korrespondenzanalyse.