Faktorenanalyse 1

Ziele

Reduzierung von Variablen/Merkmalen auf wenige, relevante Faktoren

Entdeckung von untereinander unabhängigen Beschreibungs- und Erklärungsvariablen

Beispielanwendung

Consumer-Einschätzungen von Streichfetten (Butter, Margarine) 4 verschiedener Anbieter (Rama, Sanella, Becel, Kerrygold)

Erhobene Eigenschaften

Streichfähigkeit

Anteil ungesättigter Fettsäuren

Kalorien

Vitaminzusätze

Haltbarkeit

Preis

Skalenniveau?

Warum?

n = 30 befragte Hausfrauen

Schritt 1 **Deskription/Merkmalsverteilung**

Mittelwerte		Rama	Sanella		Becel	Kerrygold
Streichfähigkeit	x1	2	1		6	5
ungesättigte Fettsäuren	x2	1	1	vs.	5	5
Kalorien	x3	1	2		6	6
Vitaminzusätze	x4	1	1		4	5
Haltbarkeit	x5	3	6		6	4
Preis	x6	3	7		5	4

Korrelation der Merkmale

Vermutung:

Streichfähigkeit x1
ungesättigte Fettsäuren x2
Kalorien x3
Vitaminzusätze x4
→ GESUNDHEIT

Haltbarkeit x5
Preis x6
→ WIRTSCHAFTLICHKEIT

FAKTOREN

Schritt 2

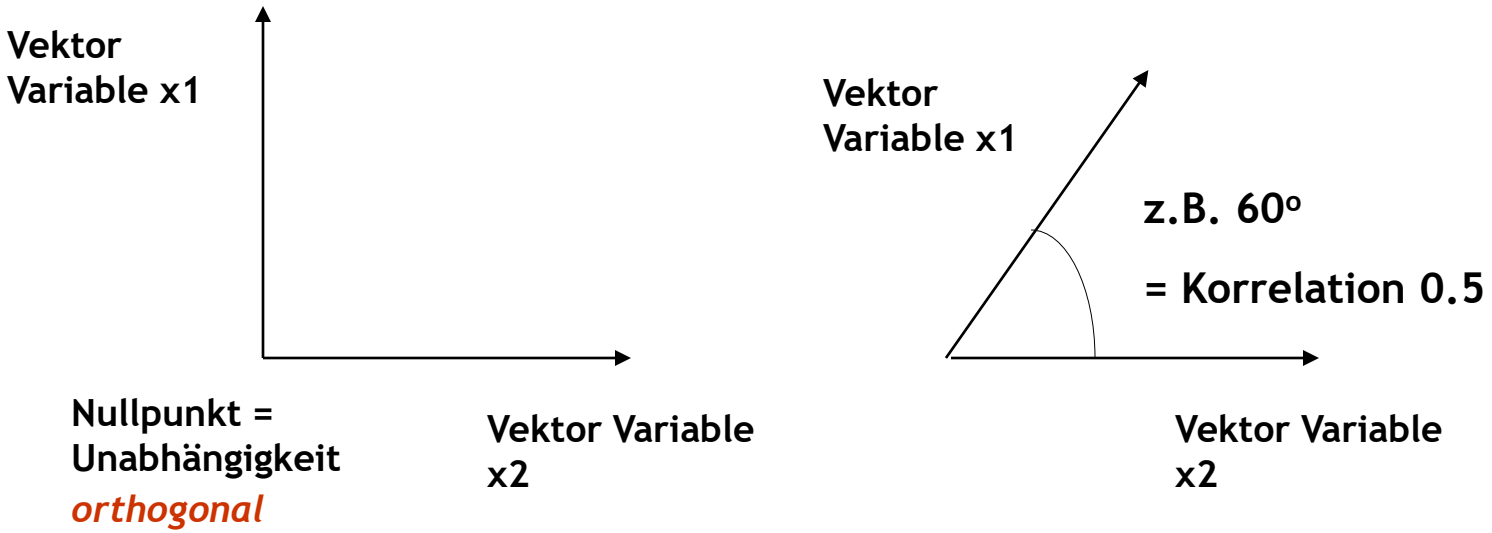
Korrelationsmatrix (r)

		x1	x2	x3	x4	x5	x6
Streichfähigkeit	x1	*1.0000*	0.9701	0.9317	0.9169	0.1400	-0.286
ung. Fettsäuren	x2	0.9701	*1.0000*	0.9898	0.9802	0.1924	-0.169
Kalorien	x3	0.9317	0.9898	*1.0000*	0.9683	0.3168	-0.185
Vitaminzusätze	x4	0.9169	0.9802	0.9683	*1.0000*	0.0808	-0.213
Haltbarkeit	x5	0.1400	0.1924	0.3168	0.0808	*1.0000*	0.8783
Preis	x6	-0.286	-0.169	-0.185	-0.213	0.8783	*1.0000*

Mathematischer Hintergrund

Korrelationskoeffizienten lassen sich auch als Winkel zwischen Vektoren darstellen.

Einfaches Beispiel: 2 Variablen/Vektoren



Die Faktorenanalyse trachtet danach, das sich in den Winkeln bzw. Korrelationskoeffizienten ausdrückende Verhältnis der Variablen zueinander in einem möglichst gering dimensionierten Raum zu reproduzieren. **Zahl der dafür benötigten Achsen = Zahl der Faktoren.**

Faktoren und Faktorladungen

Annahme

Jeder Messwert einer Ausgangsvariable x_i (**eines Objekts k**, hier: Marken) lässt sich als lineare Kombination mehrerer ***hypothetischer* Faktoren** beschreiben.

$$x_{ik} = a_{i1} * p_{1k} + a_{i2} * p_{2k} + \ldots a_{iq} * p_{qk}$$

Faktor 1

Faktorladung

Wie viel hat ein Faktor mit der Ausgangsvariable zu tun?

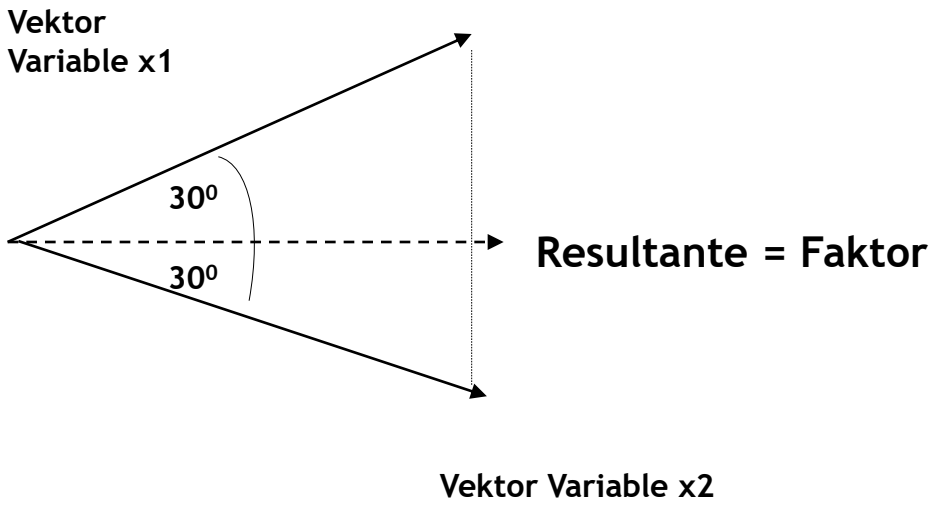
= Maßgröße für deren Zusammenhang

= Korrelationskoeffizient zwischen Variable und Faktor

Behauptung eines linearen Zusammenhangs

Bestimmung

Idealtypisches, vereinfachtes Beispiel:
2 Variablen (= Vektoren), Korrelation 0.5 (= Winkel von 60°)

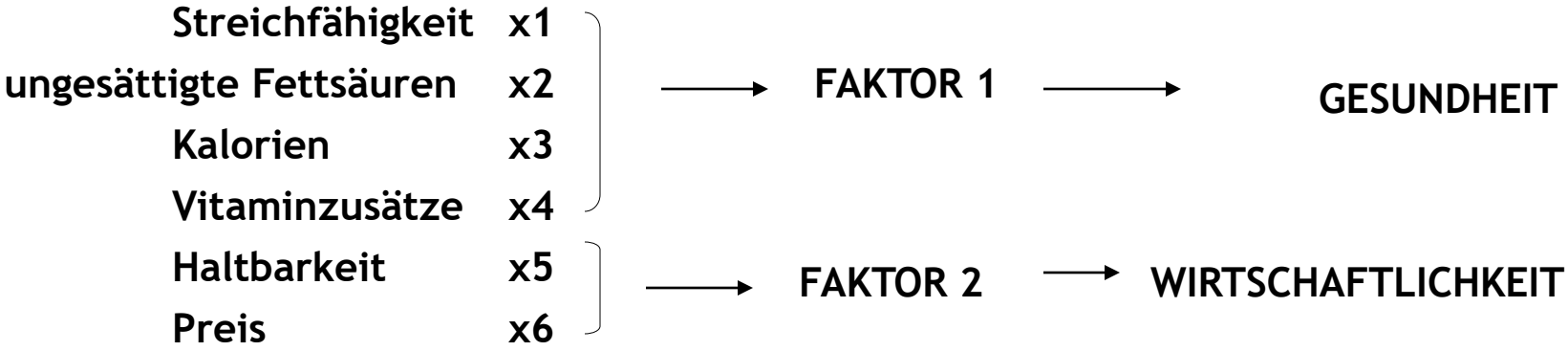


Faktorladung = Korrelationskoeffizient zwischen Variable(n) und Faktor

Faktorladung hier = cosinus 30° = 0.8660

Verfahren: HAUPTACHSENANALYSE

Faktorladungen



	Faktor 1	Faktor 2
x1	0.9673	-0.0902
x2	0.9998	0.0006
x3	0.9869	0.1470
x4	0.9757	-0.0794
x5	0.1882	0.9341
x6	-0.1660	0.9410

Ladungsmatrix oder Faktormuster

Schritt 3

Wie viel der Gesamtvarianz werden durch die Faktoren erklärt?

Gesamtvarianz = durch Faktoren erklärt + nicht durch Faktoren erklärt

KOMMUNALITÄT

theoretisches Maximum = 1

Und 1 bedeutet?

„Eigenvarianz“ des Messwerts / Variable **+** **Messfehler etc.**

(1) Der jeweils höchste Korrelationskoeffizient einer Variablen mit einer anderen (-> Korrelationsmatrix) -> Schätzung

(2) Hauptkomponentenanalyse – diese unterstellt jedoch den Grenzfall, dass die gesamte Varianz auf die Faktoren zurückzuführen ist.

(3) Iteratives Verfahren

Schritt 4

Bestimmung der Faktoren. Wie viele Faktoren, und aufgrund welchen Kriteriums?

2 gebräuchliche Kriterien

(die zu unterschiedlichen Lösungen führen können)

Kaiser

Zahl der Faktoren = Zahl der Faktoren mit **Eigenwert** > 1

< 1 bedeutet: Faktor erklärt weniger Varianz als die Variablen selbst.

Eigenwert = Erklärungsanteil eines Faktors in bezug auf die Varianz aller Variablen

*Also nicht zu verwechseln mit: **Kommunalität** = Erklärungsanteil aller Faktoren in bezug auf eine Variable / eines Messwerts.*

Berechnung: Summe der quadrierten Ladungsquadrate (-> Ladungsmatrix)

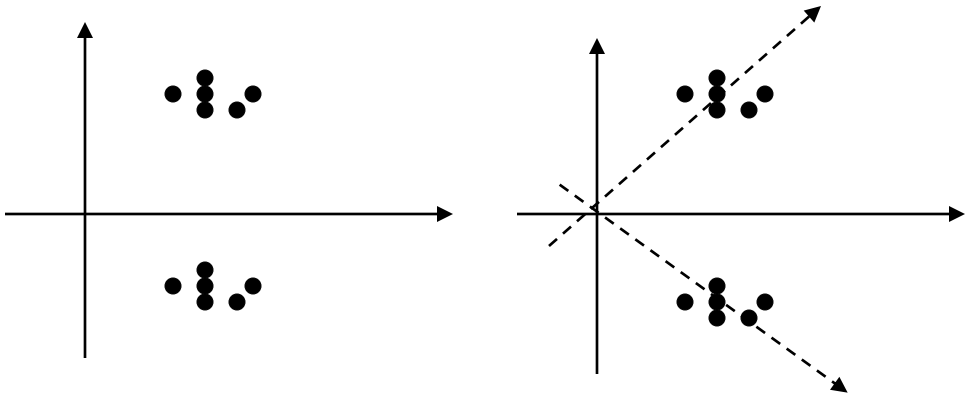
Nfactors

Zahl der Faktoren < Hälfte der Anzahl der Variablen

Schritt 5 Interpretation der Ladungsmatrix / des Faktormusters

	Faktor 1	Faktor 2
x1	0.9673	-0.0902
x2	0.9998	0.0006
x3	0.9869	0.1470
x4	0.9757	-0.0794
x5	0.1882	0.9341
x6	-0.1660	0.9410

Aussagekraft einer Hauptachsenanalyse ändert sich durch Drehung des Koordinatensystems im Ursprung nicht.



Nicht immer so eindeutig; häufig laden Variablen / Messwerte deutlich auf mehr als einem Faktor.

Rotation

- **VARIMAX-ROTATION**
- **schiefwinklige Rotation**

bedeutet? **Aufgabe der Unabhängigkeitsprämisse**

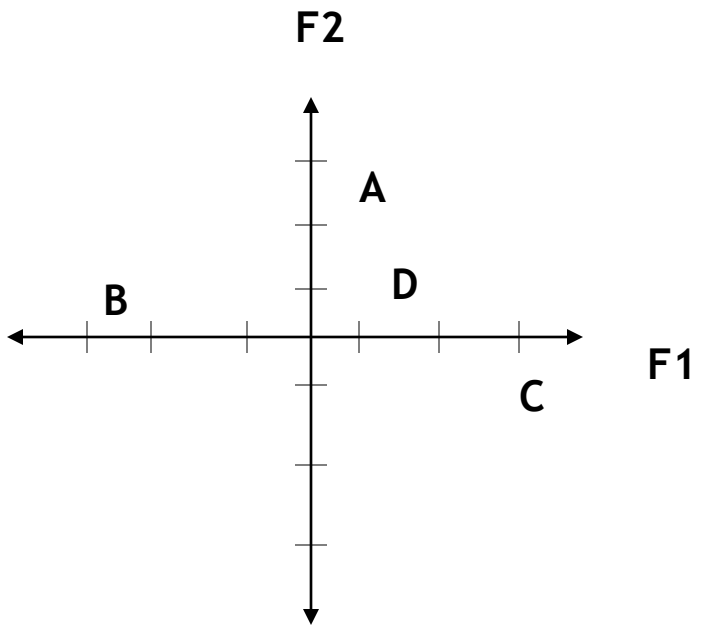
Schritt 6

Wie nah oder fern sind sich die Objekte (hier: Streichfettmarken) in bezug auf die Faktoren?

Positionierungsanalyse

	Faktor 1	Faktor 2
(A) Rama	0.6	1.3
(B) Sanella	-1.2	0.8
(C) Becel	1.6	-0.7
(D) Kerrygold	0.9	1.1

Schätzung i.d.Regel durch multiple Regression



Clusteranalyse

Ziel

Gruppierung von Untersuchungsobjekten zu natürlichen („empirischen“) Gruppen → **Cluster**

Entdeckung einer Struktur → **Exploration**

Zentrale Entscheidungen:

Welches Proximitätsmaß?
Welcher Fusionsalgorithmus?

Voraussetzungen

Bereinigung fehlender Werte

Bei stark unterschiedlichen Wertebereichen der Variablen: z-Transformation

Anpassung der Skalenniveaus, ggf. Transformation auf das jeweils niedrigste Niveau (*kein Muss*)

Ziel

Bestimmung der Distanz/Ähnlichkeiten zwischen den Objekten

→ (Varianz) innerhalb der Gruppen möglichst homogen,
(Varianz) zwischen den Gruppen möglichst heterogen

Verfahren

Intervallskalierte Variablen

Quadrierte Euklidische Distanz

Einfache Euklidische Distanz

City-Block-Distanz

Grundlage Minkowski-Metrik

Anzahl Clustervariablen

$$d_{k,i} = \left(\sum_{j=1}^{J} \left| x_{kj} - x_{ij} \right|^{r}\right)^{1/r}$$

Distanz zwischen den Objekten k und i

Minkowski-Konstante

Variablenwerte der Objekte k und i

Verfahren

Binäre Variablen

→ Merkmal vorhanden / nicht vorhanden

Beispiel:

	Konfiguration				
Auto	**Airbag**	**ESP**	**Metallic**	**Navi**	**ABS**
Mercedes	1	1	0	1	0
BMW	1	1	1	0	0
Fall	A	A	B	C	D

$$S_{ij} = \frac{\alpha \times a + \delta \times d}{\alpha \times a + \lambda(b+c) + \delta \times d}$$

griechische Buchstaben:

mögliche Kombinationen, Anzahl

↓

Gewichtung

Objekte i, j

Häufigkeit, mit der ein bestimmter Fall vorkommt

Die Gewichte werden je nach Proximitätsmaß gewählt.
Sie entscheiden darüber, ob und inwiefern die Fälle A, B, C und D berücksichtigt werden.

Beispiel:

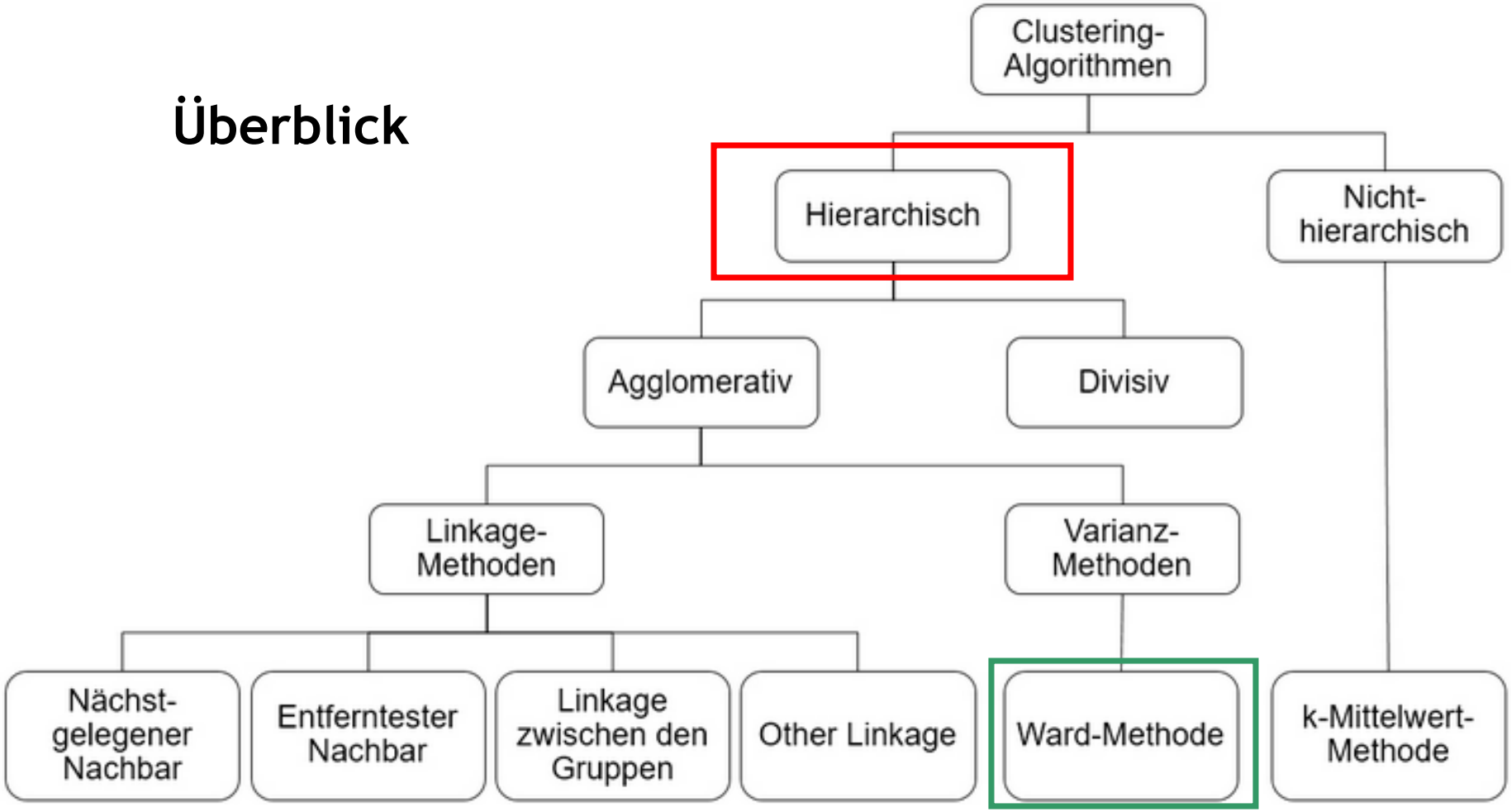
	Konfiguration				
Auto	**Airbag**	**ESP**	**Metallic**	**Navi**	**ABS**
Mercedes	1	1	0	1	0
BMW	1	1	1	0	0
Fall	A	A	B	C	D

2 einfache Varianten

Simple Matching $S_{ij} = \frac{a+d}{a+b+c+d}$

Dice $S_{ij} = \frac{2a}{2a+b+c}$ → **Gewichtung von A, D nicht berücksichtigt**

Überblick


Clustering-Algorithmen
Hierarchisch
Nicht-hierarchisch
Agglomerativ
Divisiv
Linkage-Methoden
Varianz-Methoden
Nächst-gelegener Nachbar
Entferntester Nachbar
Linkage zwischen den Gruppen
Other Linkage
Ward-Methode
k-Mittelwert-Methode

Fusionsalgorithmen 2

Verfahren

SINGLE LINKAGE nächstgelegener Nachbar	*Minimum* aller möglichen Distanzen zwischen den Datenpunkten in Cluster 1 und denen in Cluster 2
COMPLETE LINKAGE entferntester Nachbar	*Maximum* aller möglichen Distanzen zwischen den Datenpunkten in Cluster 1 und denen in Cluster 2
AVERAGE LINKAGE Linkage zwischen Gruppen	*Mittelwert* aller möglichen Distanzen zwischen den Datenpunkten in Cluster 1 und denen in Cluster 2
OTHER LINKAGE Verschiedenes	z.B. Distanz zwischen dem *Median* von Cluster 1 und Cluster 2

Ein sehr stabiles, aber rechenintensives Verfahren:

WARD

Pro Cluster Berechnung der Summe der quadrierten Distanzen der Einzelfälle vom jeweiligen Zentroid. Fusion jeweils der 2 Cluster, deren Zusammenfügung die geringste Erhöhung der Gesamtsumme der quadrierten Distanzen zur Folge hat

Gesucht: Berufscluster nach Einkommen / Markenbewusstsein

Beruf	Einkommen	Marke
A	6861	21765
Ing	5150	28245
Che	5474	25179
M	7389	19048
Pr	5152	24608
CEO	12810	27611
Anw	7203	21536
K	4162	24823
Arch	6779	22499
F	3204	7465
PH	5335	17471
L	4311	14735
B	3949	17921
F	2132	8822
Serv	3018	12201

metrische Variablen

Analyse mit SPSS

```
CLUSTER Einkommen Marke
/METHOD WARD
/MEASURE= SEUCLID
/ID=Beruf
/PRINT SCHEDULE CLUSTER(2,5)
/PRINT DISTANCE
/PLOT DENDROGRAM VICICLE
/SAVE CLUSTER(2,5).
```

Distanzmatrix

Schritt 1

Fall	Quadriertes euklidisches Distanzmaß														
	1:A	2:Ing	3:Che	4:M	5:Pr	6:F	7:PH	8:L	9:B	10:CEO	11:Anw	12:K	13:F	14:Arch	15:Serv
1:A	,000	1,444	,570	,218	,636	6,897	,793	2,164	1,641	6,196	,019	1,330	7,385	,014	4,422
2:Ing	1,444	,000	,240	2,774	,315	10,844	2,766	4,447	2,754	8,935	1,711	,427	10,357	1,189	6,813
3:Che	,570	,240	,000	1,452	,024	8,246	1,416	2,800	1,606	8,327	,770	,265	8,061	,430	4,923
4:M	,218	2,774	1,452	,000	1,496	5,855	,701	1,883	1,830	6,214	,152	2,377	6,690	,340	4,021
5:Pr	,636	,315	,024	1,496	,000	7,566	1,216	2,426	1,283	9,135	,864	,150	7,313	,508	4,353
6:F	6,897	10,844	8,246	5,855	7,566	,000	3,072	1,443	2,684	23,688	7,141	7,305	,219	7,319	,539
7:PH	,793	2,766	1,416	,701	1,216	3,072	,000	,338	,297	10,944	,924	1,495	3,339	,918	1,477
8:L	2,164	4,447	2,800	1,883	2,426	1,443	,338	,000	,261	14,930	2,372	2,423	1,554	2,360	,407
9:B	1,641	2,754	1,606	1,830	1,283	2,684	,297	,261	,000	14,176	1,921	1,140	2,471	1,717	,910
10:CEO	6,196	8,935	8,327	6,214	9,135	23,688	10,944	14,930	14,176	,000	5,660	11,561	25,739	6,154	20,232
11:Anw	,019	1,711	,770	,152	,864	7,141	,924	2,372	1,921	5,660	,000	1,664	7,756	,049	4,736
12:K	1,330	,427	,265	2,377	,150	7,305	1,495	2,423	1,140	11,561	1,664	,000	6,715	1,170	3,988
13:F	7,385	10,357	8,061	6,690	7,313	,219	3,339	1,554	2,471	25,739	7,756	6,715	,000	7,733	,391
14:Arch	,014	1,189	,430	,340	,508	7,319	,918	2,360	1,717	6,154	,049	1,170	7,733	,000	4,673
15:Serv	4,422	6,813	4,923	4,021	4,353	,539	1,477	,407	,910	20,232	4,736	3,988	,391	4,673	,000

Schritt 2

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	14	,007	0	0	3
2	3	5	,019	0	0	6
3	1	11	,039	1	0	8
4	6	13	,148	0	0	10
5	8	9	,279	0	0	7
6	3	12	,414	2	0	9
7	7	8	,582	0	5	12
8	1	4	,752	3	0	11
9	2	3	,961	0	6	11
10	6	15	1,234	4	0	12
11	1	2	3,506	8	9	13
12	6	7	6,058	10	7	14
13	1	10	12,654	11	0	14
14	1	6	28,000	13	12	0

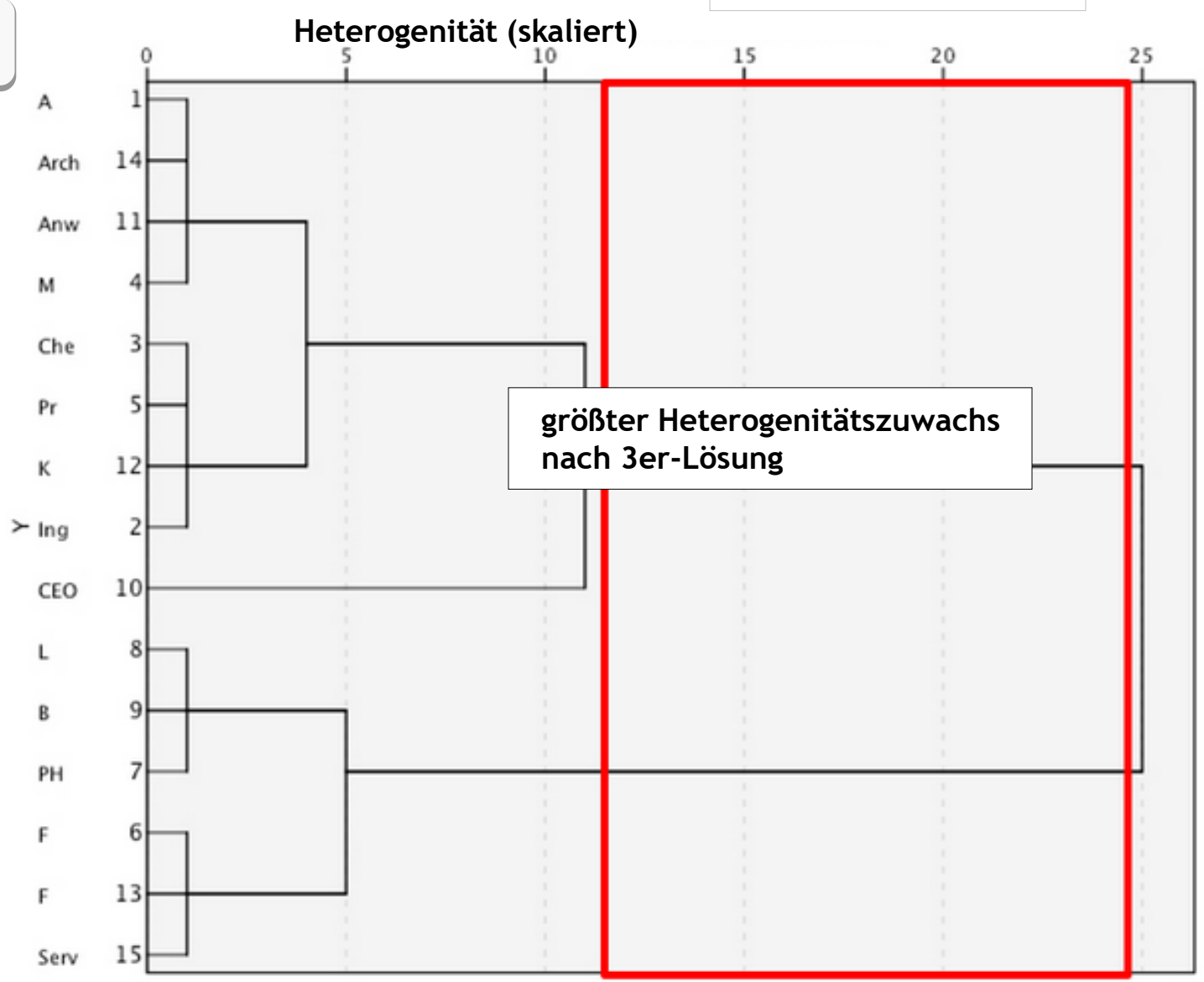
Start: jeder Fall = eignes Cluster
Ende: alle Fälle = ein Cluster
n = 15

Heterogenitätsmaß
steigt bei jedem Schritt

Was ist die optimale Lösung?

Dendrogramm

Schritt 3

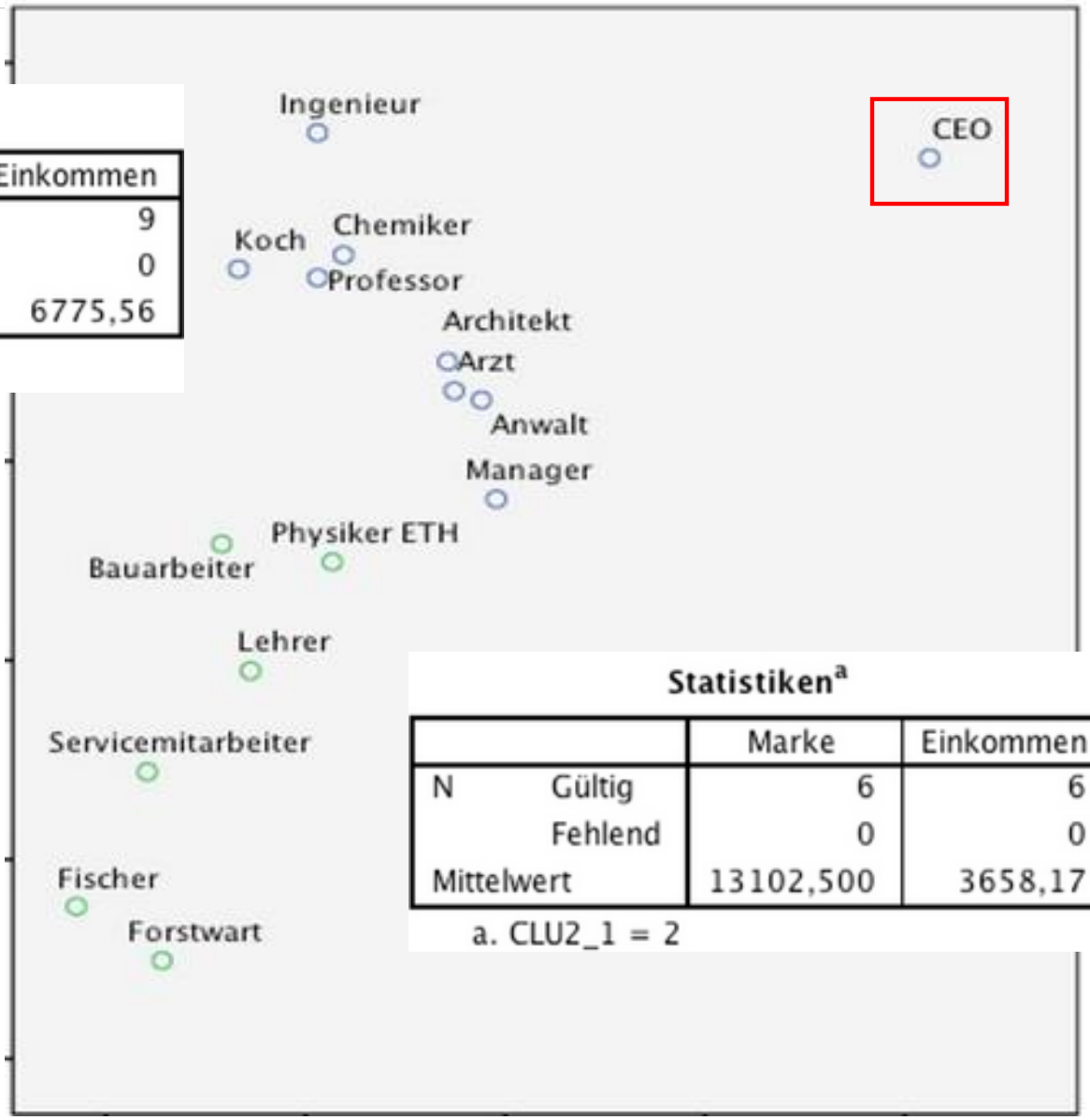


Clusterzugehörigkeit

Statistiken[a]

		Marke	Einkommen
N	Gültig	9	9
	Fehlend	0	0
Mittelwert		23923,778	6775,56

a. CLU2_1 = 1



Statistiken[a]

		Marke	Einkommen
N	Gültig	6	6
	Fehlend	0	0
Mittelwert		13102,500	3658,17

a. CLU2_1 = 2

Diskriminanzanalyse 1

Fragestellungen

Besteht zwischen zwei oder mehreren vorgegebenen Gruppen von Objekten oder Personen ein signifikanter Unterschied hinsichtlich der Gesamtstruktur mehrerer Merkmale?

Welche Kombination von Merkmalen ermöglichst die bestmögliche Trennung der vorgegebenen Gruppen?

Welche relative Bedeutung kommt einzelnen Merkmalen bei der Unterscheidung der Gruppen zu?

Welche der bereits unterschiedenen Gruppen sind neu zu untersuchende Objekte oder Personen aufgrund ihrer Merkmalsausprägungen zuzuordnen?

Charakter

spezielle Form der Regression: unabhängige (Gruppen-)Variable nominalskaliert

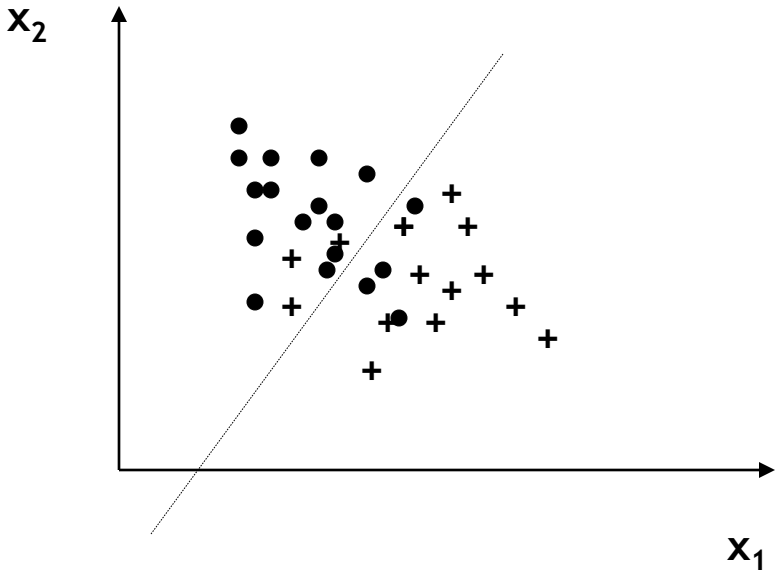
spezielle Form der Faktorenanalyse / Verfahren der Datenreduktion

Diskriminanzanalyse 2

Einfachster Fall

Zwei Gruppen, zwei Variablen, Befragung

(A)	Becel	•	Streichfähigkeit	(x1)
(B)	Kerrygold	+	Haltbarkeit	(x2)



Ziel

Berechnung einer Geraden, die bestmöglichst trennt

Diskriminanz- bzw. Trennfunktion

Diskriminanzanalyse 3

Definition

$$y_{ik} = b_1 * x_{1ik} + b_2 * x_{2ik} + \ldots$$

Diskriminanzwert y_{ik} für (das Eigenschaftsurteil der) Person i bzgl. Marke k.

Diskriminanzkoeffizient der unabhängigen Variable x_1 (hier: Eigenschaft)

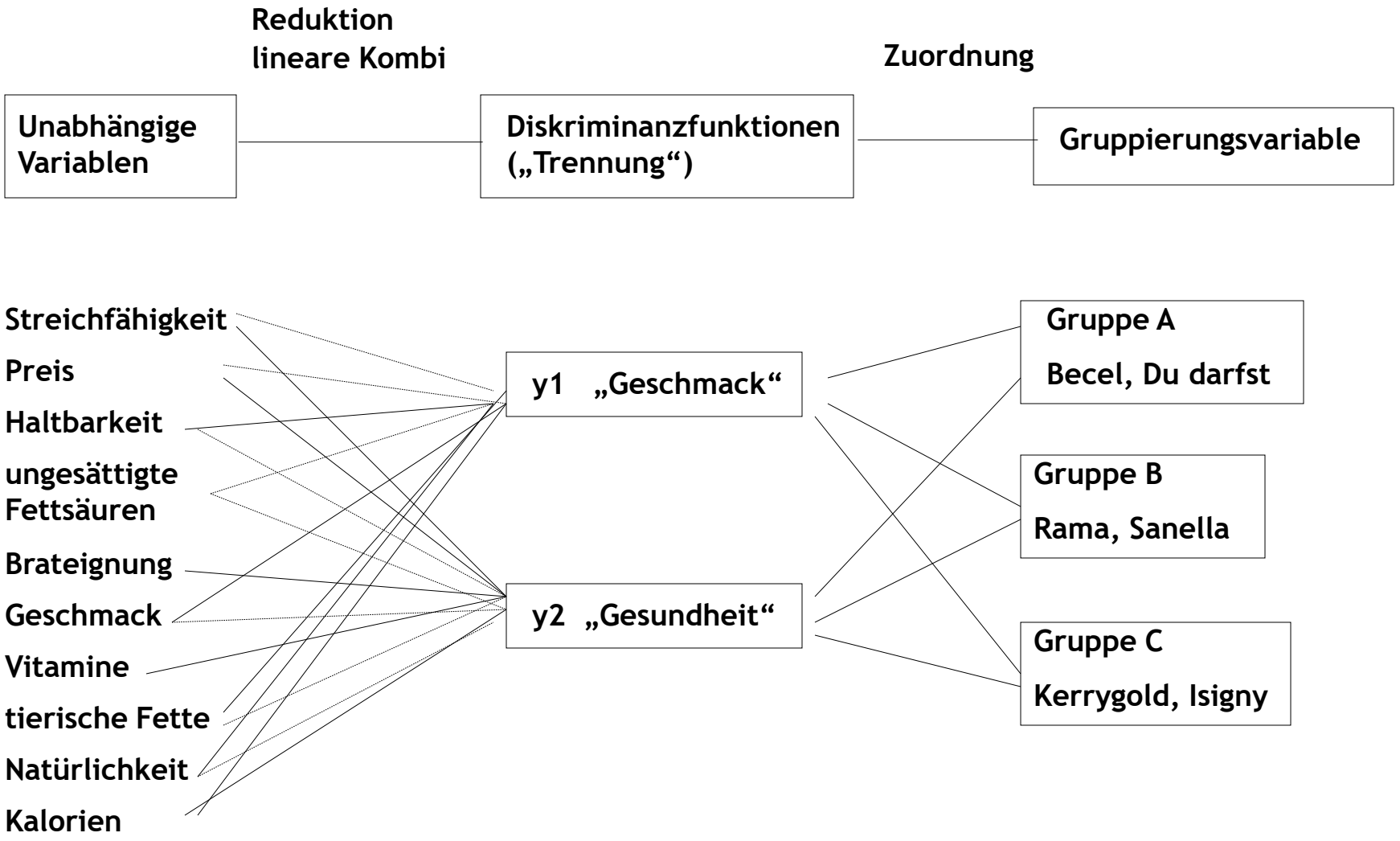
Von Person i bei der Marke k benannte Ausprägung der unabhängigen Variable x_1

Ziel

Festlegung der Diskriminanzkoeffizienten so, dass sich die arithmetischen Mittel der Diskriminanzwerte der Objekte k (hier: Marken) signifikant unterscheiden

→ **bestmögliche Trennung**

Ergebnis


Reduktion
lineare Kombi
Zuordnung
Unabhängige Variablen
Diskriminanzfunktionen („Trennung“)
Gruppierungsvariable
Streichfähigkeit
Preis
Haltbarkeit
ungesättigte Fettsäuren
Brateignung
Geschmack
Vitamine
tierische Fette
Natürlichkeit
Kalorien
y1 „Geschmack“
y2 „Gesundheit“
Gruppe A
Becel, Du darfst
Gruppe B
Rama, Sanella
Gruppe C
Kerrygold, Isigny