# Multivariate Analysemethoden


Thema: Multidimensionale Skalierung (MDS)

***Günter Meinhardt***
***Johannes Gutenberg Universität Mainz***

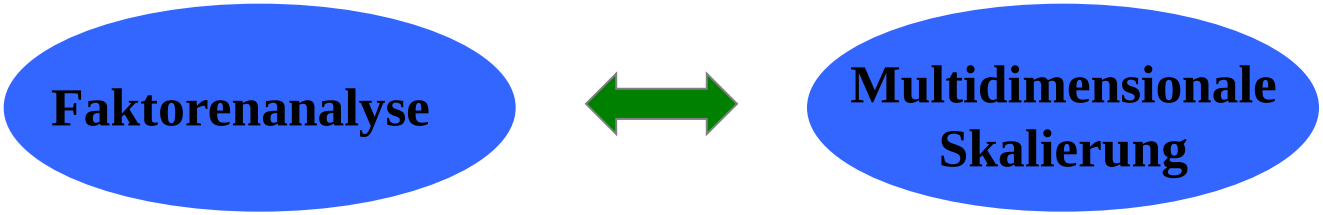
**Thema**

# Multidimensionale Skalierung

Problem:

Positionierung von Messobjekten in einem latenten Raum (hier: Wahrnehmungsraum)

Möglichkeiten:



**Faktorenanalyse**

**Vorgehen**

Man lässt Personen Eigenschaftsausprägungen von Objekten einschätzen (Item-Schätzskalen). Man faktorisiert die Skalen und betrachtet die Koordinaten der Objekte auf den neuen (unabhängigen) Dimensionen (= latenter Wahrnehmungsraum).

**MDS**

**Vorgehen**

Man lässt nur die Ähnlichkeit der Objekte beurteilen (ohne den direkten Bezug auf konkrete Eigenschaften) und probiert die Anordnung („Konfiguration“) der Objekte in einem latenten Raum derart, dass die Ähnlichkeitsurteile möglichst gut reproduziert werden.

**Latente Variable**

# Faktorenanalyse

Man möchte Objekte (Personen) in einem Raum latenter Dimensionen (Fähigkeiten, Traits) anordnen. Gegeben ist ein Set von Beobachtungen (Messvariablen)

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$$

Problem: Finde latente Variablen

$$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$$

$r \leq p$, so dass jede Variable $\mathbf{x}_k$ eine Linearkombination der $\mathbf{w}_l$ ist:

$$\mathbf{x}_k = b_{k1}\mathbf{w}_1 + b_{k2}\mathbf{w}_2 + \ldots + b_{kr}\mathbf{w}_r$$

Beispiel:
Das Abschneiden im Abitur mit Deutsch, Mathe, Physik, Latein und Geographie wird erklärt aus latenten Variablen Memory, Induction, Perceptual Speed, Space, Verbal Comprehension.

**Latente Variable**

# Multidimensionale Skalierung

Man möchte Objekte (Personen) in einem Raum latenter Dimensionen anordnen. Gegeben ist ein Set von Beobachtungen über die (sensorischen) Distanzen der Objekte: (Distanzmatrix)

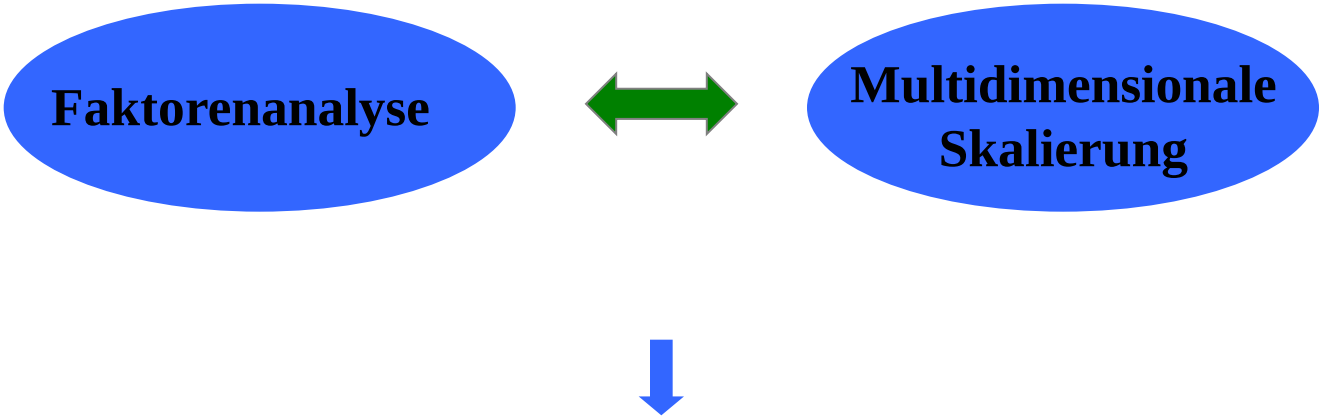
$$\mathbf{D} = \begin{array}{c|cccccc} & o_1 & o_2 & \cdots & o_j & \cdots & o_n \\ \hline o_1 & 0 & & & & & \\ o_2 & d_{21} & 0 & & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ o_j & d_{j1} & d_{j2} & \cdots & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ o_n & d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & 0 \end{array}$$

Problem: Finde latente Variablen

$$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$$

$r \leq n$, so dass die Distanzen zwischen den Objekten auf den Koordinaten reproduziert werden.

Beispiel:
Man lässt Filmschauspieler paarweise nach Ähnlichkeit/Unähnlichkeit bewerten. Die MDS soll den latenten Wahrnehmungsraum liefern, auf dem die Schauspieler angeordnet werden können, so dass die Ähnlichkeitsurteile reproduziert werden.




**Demo - Beispiel mit Excel und Statistica**

MDS

**Vorteile**

- relevante Eigenschaften dürfen unbekannt sein (keine Verzerrung durch Vorauswahl)
- kann bereits bei Rangdaten eingesetzt werden (Ergebnisse sind mit metrischer MDS quasi identisch)

**Nachteile**

- Aggregation über Personen ist problematisch (Bezug auf verschiedene latente Dimensionen beim Urteil)
- Großer Interpretationsfreiraum beim Untersucher bei der inhaltlichen Benennung der Dimensionen (vage)
- MDS- Lösung ist nicht algorithmisch (Keine Garantie die beste Lösung gefunden zu haben)
- MDS Lösung ist von weiteren Parametern abhängig (Distanzmodell, Anzahl der Dimensionen)

## Städte-Beispiel

**MDS**

Distanzen von Städten in km

| | Basel | Berlin | Frank-furt | Ham-burg | Hanno-ver | Kassel | Köln | Mün-chen | Nürn-berg | Stutt. |
|---|---|---|---|---|---|---|---|---|---|---|
| Basel | | | | | | | | | | |
| Berlin | 847 | | | | | | | | | |
| Frankfurt | 337 | 555 | | | | | | | | |
| Hamburg | 820 | 294 | 495 | | | | | | | |
| Hannover | 677 | 282 | 352 | 154 | | | | | | |
| Kassel | 517 | 378 | 193 | 307 | 164 | | | | | |
| Köln | 496 | 569 | 189 | 422 | 287 | 243 | | | | |
| München | 438 | 584 | 400 | 782 | 639 | 482 | 578 | | | |
| Nürnberg | 437 | 437 | 228 | 609 | 466 | 309 | 405 | 167 | | |
| Stuttgart | 268 | 634 | 217 | 668 | 526 | 366 | 376 | 220 | 207 | |

## Städte-Beispiel

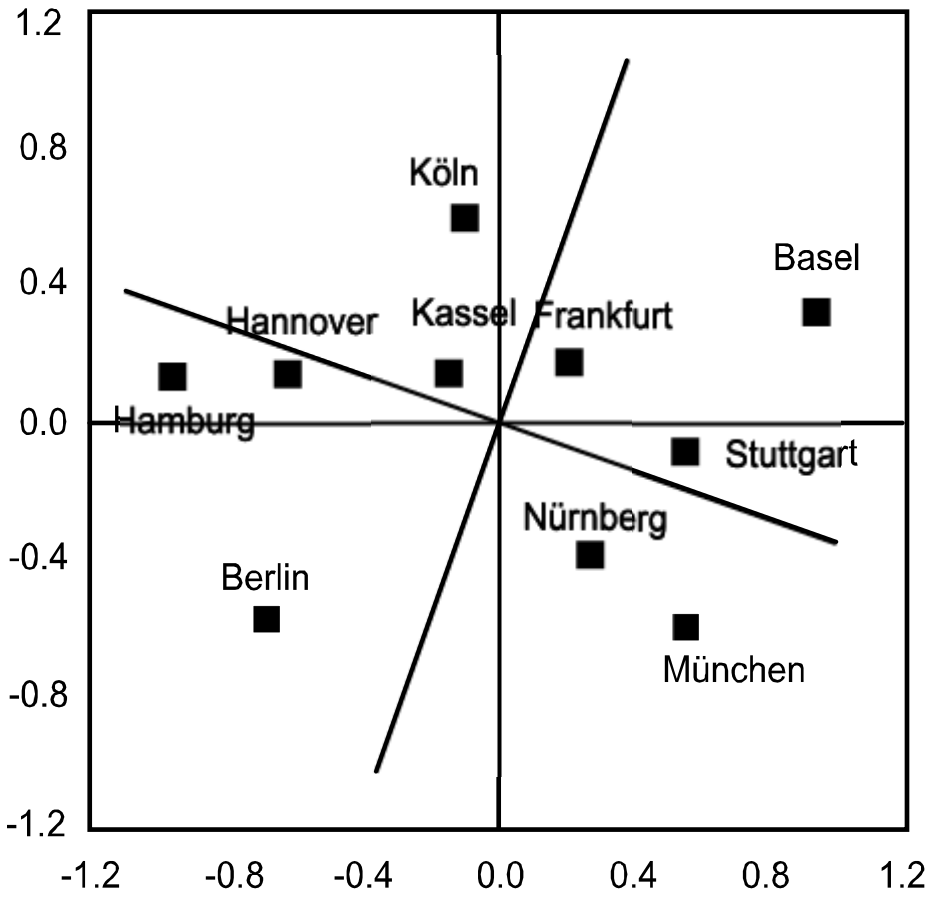
**MDS**

Rangreihe der Distanzen von Städten

| | Basel | Berlin | Frank-furt | Ham-burg | Hanno-ver | Kassel | Köln | Mün-chen | Nürn-berg | Stutt. |
|---|---|---|---|---|---|---|---|---|---|---|
| Basel | | | | | | | | | | |
| Berlin | 45 | | | | | | | | | |
| Frankfurt | 17 | 34 | | | | | | | | |
| Hamburg | 44 | 14 | 30 | | | | | | | |
| Hannover | 42 | 12 | 18 | 1 | | | | | | |
| Kassel | 32 | 21 | 5 | 15 | 2 | | | | | |
| Köln | 31 | 35 | 4 | 24 | 13 | 10 | | | | |
| München | 27 | 37 | 22 | 43 | 40 | 29 | 36 | | | |
| Nürnberg | 25 | 25 | 9 | 38 | 28 | 16 | 23 | 3 | | |
| Stuttgart | 11 | 39 | 7 | 41 | 33 | 19 | 20 | 8 | 6 | |

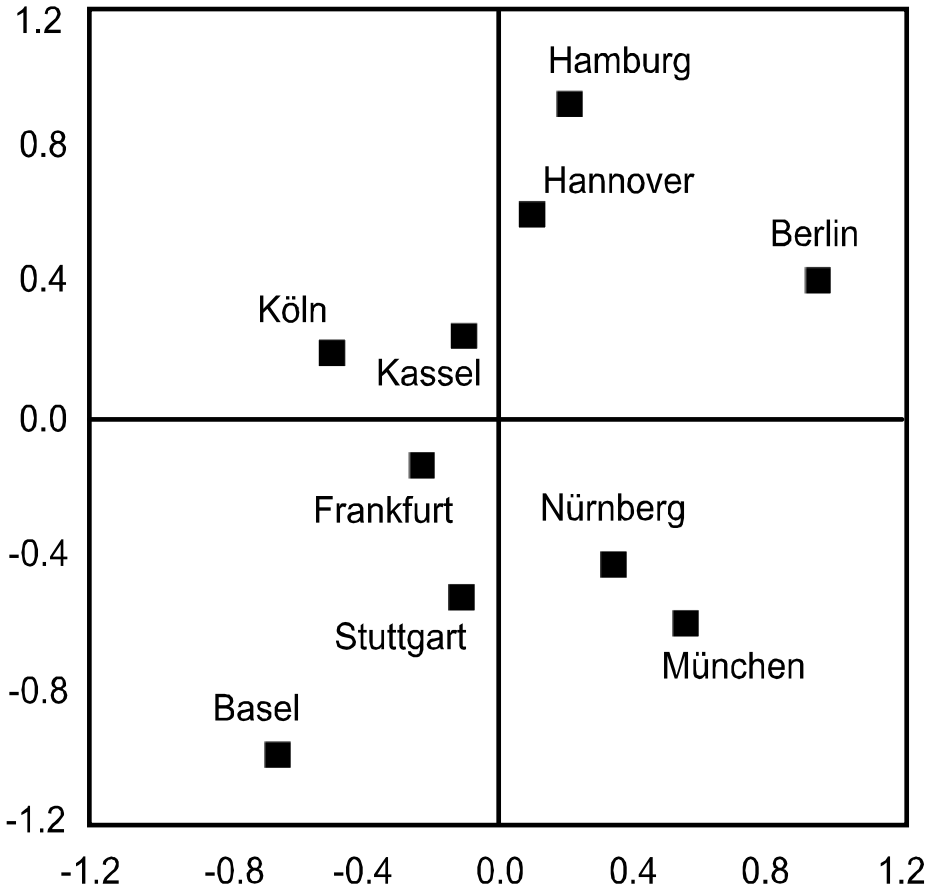
## Städte-Beispiel

MDS - Konfiguration



## Städte-Beispiel

MDS – Konfiguration nach Rotation und Spiegelung



# MDS

**Anwendung**

Die MDS ist ein exploratives Verfahren und nicht zur strengen Hypothesenprüfung geeignet

**Probleme**

- Auffinden der Konfiguration (relative Lage der Objekte zueinander im Wahrnehmungsraum, wenn nur die Distanzen bekannt sind
- Bestimmung der Dimensionalität
- Bestimmung der Metrik

**Kommentar**

- Die Konfiguration ist unabhängig von Rotation und Spiegelung
- Es finden fast nur nichtmetrische MDS Prozeduren Verwendung (Kruskal)

## Ablauf

**MDS**

1. Messung von Ähnlichkeiten
2. Wahl des Distanzmodells
3. Ermittlung der Konfiguration
4. Zahl und Interpretation der Dimensionen
5. Aggregation von Personen

**Methoden**

- Rangreihungsmethode
- Ankerpunktmethode
- Ratingverfahren

**Rangreihe**

Es werden „n über 2“ Paare geordnet von „unähnlichstes Paar“ zu „ähnlichstes Paar“ (bei grossem n kaum möglich)

**Ankerpunkt**

Jedes Objekt ist einmal Vergleichsobjekt (Anker) für alle anderen Objekte. Es werden soviele Rangreihen wie Objekte erstellt. Man erhält eine asymmetrische quadratrische Distanzmatrix, die in eine symmetrische überführt werden kann.

**Rating**

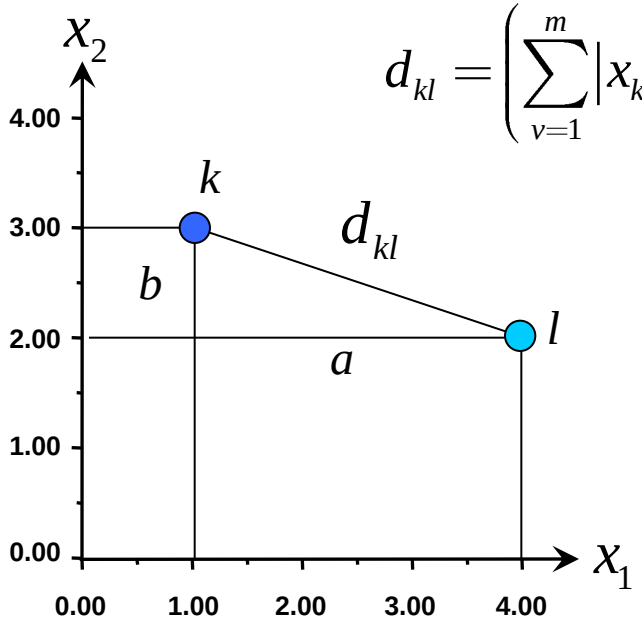
Man bildet ale möglichen Paare und lässt diese, randomisiert dargeboten, auf einer Ratingskala nach Ähnlichkeit bewerten.

**Probleme:**

Ties (Rangbindungen), Reliabilität der Ränge

**Euklidische Metrik**
***p* = 2**

$$d_{kl} = \left( \sum_{v=1}^{m} \left| x_{kv} - x_{lv} \right|^{p} \right)^{\frac{1}{p}}$$




$$a = x_{k1} - x_{l1} = 1 - 4 = -3$$

$$b = x_{k2} - x_{l2} = 3 - 2 = 1$$

$$d_{kl} = \sqrt{3^2 + 1^2} = \sqrt{10} = 3.16$$

**Objektdistanz**

## Minkowski-Metriken

$$d_{kl} = \left( \sum_{v=1}^{m} \left| x_{kv} - x_{lv} \right|^{p} \right)^{\frac{1}{p}}$$

**$p = 2$** Euklidische Metrik: Abstand der Objekte ist die Länge der Verbindungslinie.

**$p = 1$** City-Block Metrik: Abstand der Objekte ist die Summe der einzelnen Koordinatendistanzen

**$p = \infty$** Supremum Metrik: Abstand der Objekte ist die größte der auftretenden Koordinatendistanzen

**Wahlkriterium** Metrik muss nach inhaltlichen Gesichtspunkten gewählt Sein, Abstände werden in diesem Sinne interpretiert.

## Konfiguration ermitteln

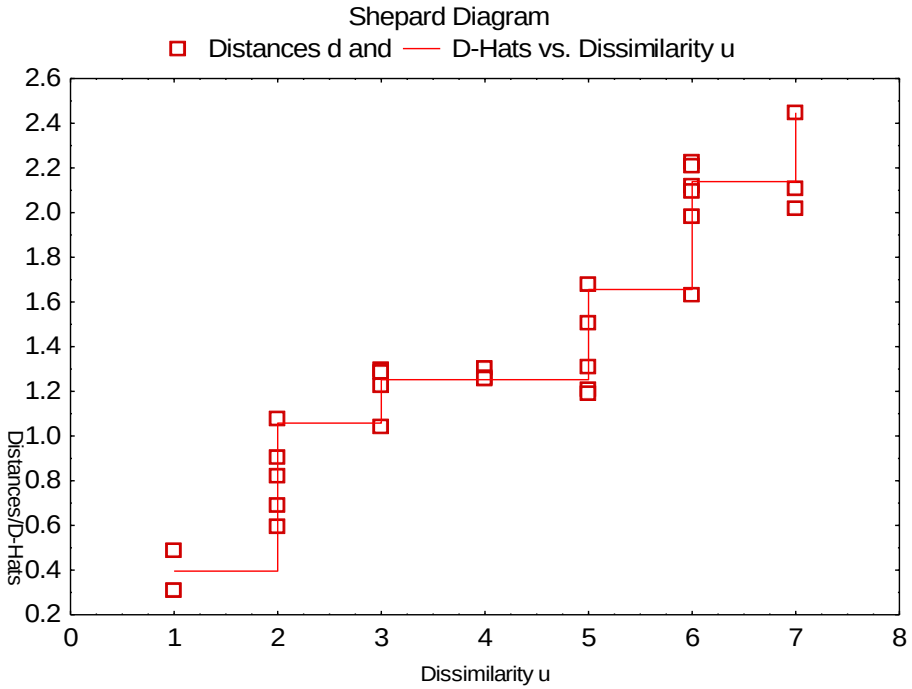
*p* = 2

***Shephard Diagramm***

Ausgehend von den Unähnlichkeiten $u$ ist ein möglichst niedrig dimensionierter Raum zu finden, in dem die Distanzen $d$ möglichst der Monotoniebeziehung

$$\text{Wenn } u_{kl} > u_{ij} \text{ dann } d_{kl} > d_{ij}$$

genügen.



# Konfiguration ermitteln

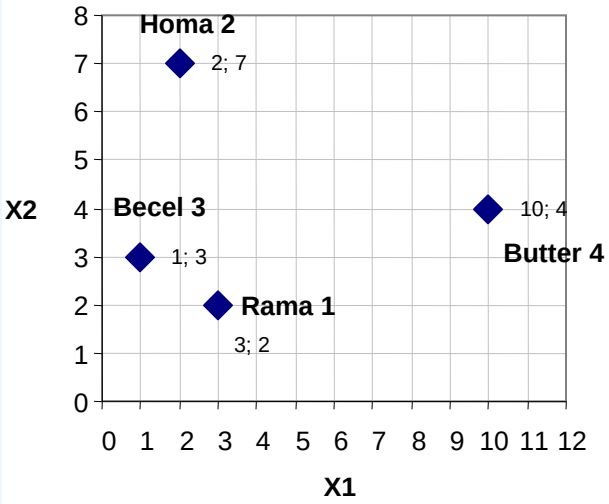
*Start-Konfiguration*

**Unähnlichkeiten u**

| | | Rama | Homa | Becel |
|---|---|---|---|---|
| 1 | Rama | | | |
| 2 | Homa | 3 | | |
| 3 | Becel | 2 | 1 | |
| 4 | Butter | 5 | 4 | 6 |

**Koordinaten x1, x2**

| | | x1 | x2 |
|---|---|---|---|
| 1 | Rama | 3 | 2 |
| 2 | Homa | 2 | 7 |
| 3 | Becel | 1 | 3 |
| 4 | Butter | 10 | 4 |

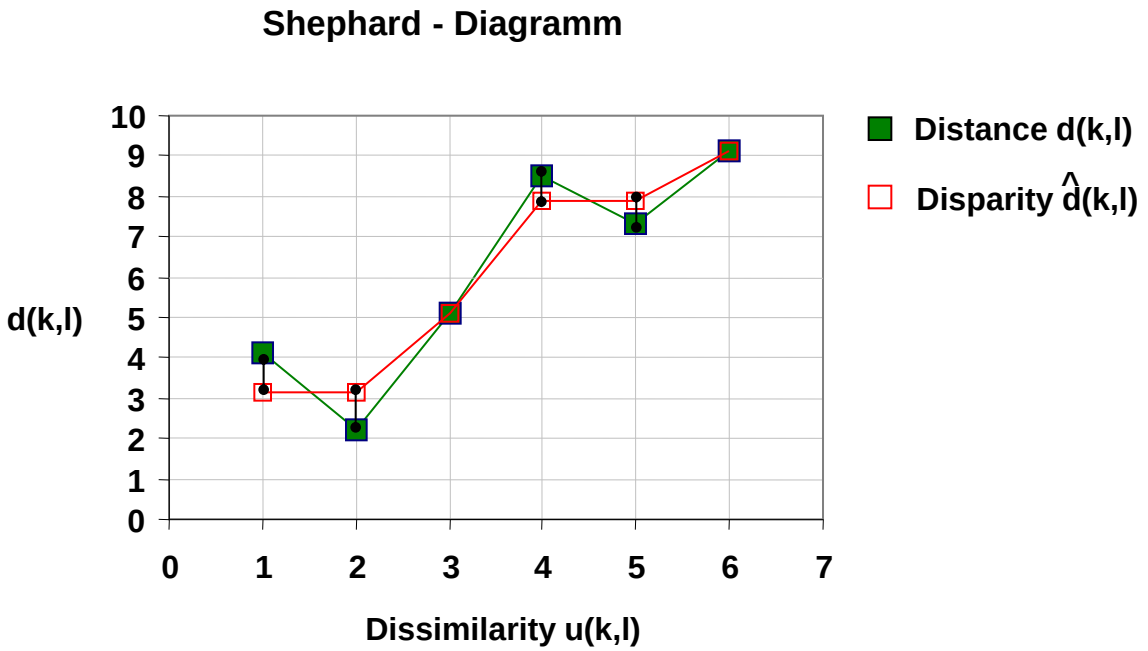



| Objekte(k,l) | $\sum \lvert x_{kv} - x_{lv} \rvert^2$ | d(kl) | Rg[d(k,l)] | u(k,l) |
|---|---|---|---|---|
| 1;2 | 1+25=26 | 5.1 | 3 | 3 |
| 1;3 | 4+1=5 | 2.2 | 1 | 2 |
| 1;4 | 49+4=53 | 7.3 | 4 | 5 |
| 2;3 | 1+16=17 | 4.1 | 2 | 1 |
| 2;4 | 64+9=73 | 8.5 | 5 | 4 |
| 3;4 | 81+1=82 | 9.1 | 6 | 6 |

# Konfiguration ermitteln

*Start-Konfiguration*

*Gütemaß "Stress"*

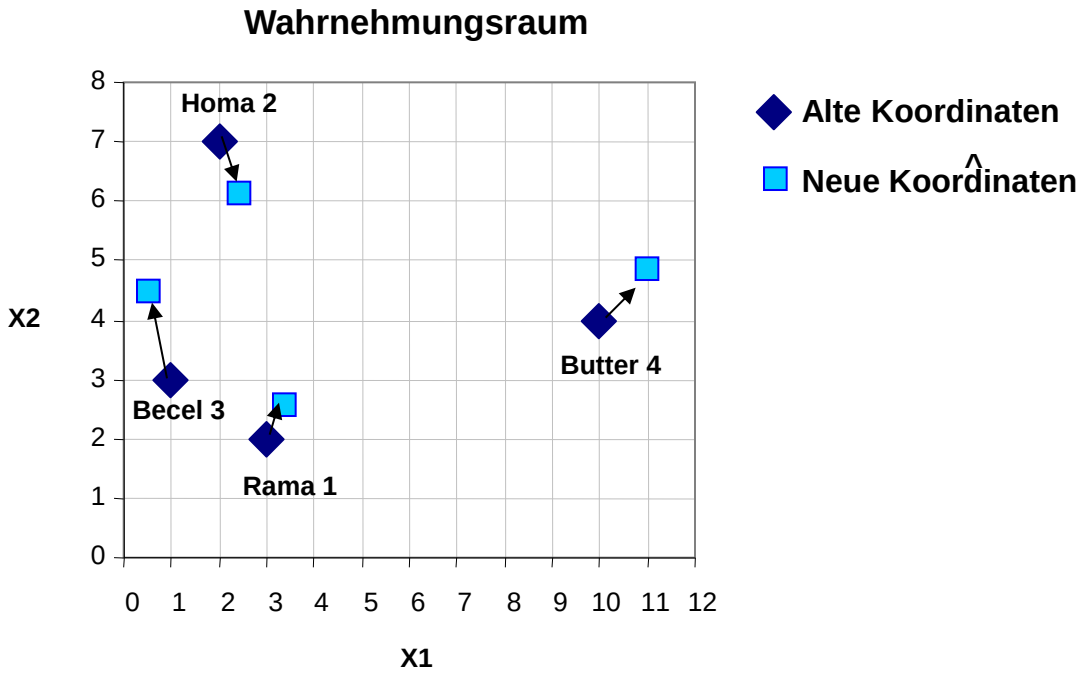



Abweichung von Distanz $d$ und Disparität $\hat{d}$

$$\text{Stress} = \sqrt{\frac{\sum_{k,l} d_{kl} - \hat{d}_{kl}^{\,2}}{\text{Faktor}}}$$

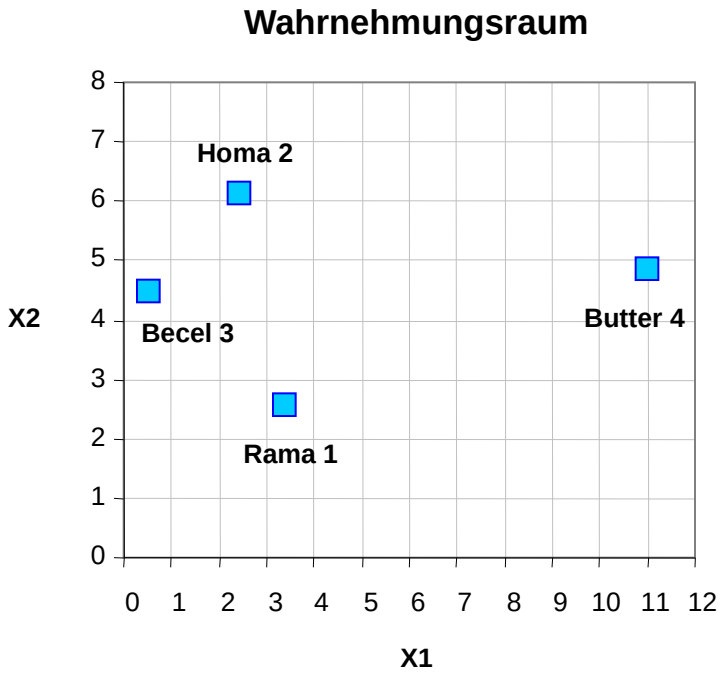
# Konfiguration ermitteln

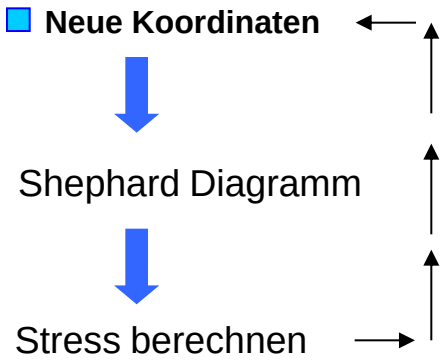
*Iterative Optimierung*



## Konfiguration ermitteln

*Iterative Optimierung*




*Gütemaß "Stress"*

Für jeden Iterationsschritt wird *Stress* bewertet. Iterationen so lange, bis *Stress* sich nicht mehr vermindert.
(Stress ist Führungsfunktion für nichtlineare Optimierung.)

**Anzahl**

- Je mehr Dimensionen, desto geringer wird Stress
- Lösungen mit einer geringeren Anzahl von Dimensionen sind einfacher zu interpretieren
- Stress darf nicht 0 werden (uneindeutige Lösung)

⬇

Trade-Off von Stress und Interpretierbarkeit

**Regeln**

- An Interpretierbarkeit orientieren, ggf. Achsen rotieren
- Stress soll niedrig sein – Anhaltswerte nach Kruskal
- Die Daten sollen einen gewissen Verdichtungsgrad Q haben, Q soll möglichst groß sein (Tabelle)

**Trade-Off**

**Durch Erhöhung der Anzahl der Dimensionen wird trivialerweise eine Repräsentierbarkeit erreicht. Gleichzeitig strebt aber die Datenverdichtung gegen 1. Erhöhung der Anzahl der Objekte n führt zu besserer Verdichtung, aber auch zu schlechterer Urteilspräzision.**

## Verdichtung Q

$$Q = \frac{\binom{n}{2}}{n \cdot m} = \frac{\text{Anzahl der Ähnlichkeiten}}{\text{Anzahl der Koordinaten}}$$

m = Anzahl Dimensionen

## Q - Tabelle

| n | m = 2 | m = 3 |
|---|---|---|
| 7 | 1.5 | 1 |
| 8 | 1.75 | 1.17 |
| 9 | 2 | 1.33 |
| 10 | 2.25 | 1.5 |
| 11 | 2.5 | 1.67 |
| 12 | 2.75 | 1.83 |
| 13 | 3 | 2 |

## Trade-Off

Trade-Off von hohem Q- Wert & niedrigem Stress-Wert

## Stressmaße

$$SM_1 = \sqrt{\frac{\sum_{k,l} d_{kl} - \hat{d}_{kl}^{\,2}}{\sum_{k,l} d_{kl}^2}} \qquad SM_2 = \sqrt{\frac{\sum_{k,l} d_{kl} - \hat{d}_{kl}^{\,2}}{\sum_{k,l} d_{kl} - \bar{d}^{\,2}}}$$

## Stress-Güte

| Güte | SM1 | SM2 |
|---|---|---|
| gering | 0.2 | 0.4 |
| ausreichend | 0.1 | 0.2 |
| gut | 0.05 | 0.1 |
| ausgezeichnet | 0.025 | 0.05 |
| perfekt | 0 | 0 |

## Richtwert

Werte zwischen gut und ausgezeichnet ergeben einen relativ glatten Anstieg im Shephard Diagramm

**Anzahl**

Die MDS als klassisches Verfahren dient der Ermittlung der Konfiguration einer Person. Aggregationen werden durchgeführt:

- Über die Ähnlichkeitsdaten wird aggregiert
- Über die Konfigurationen wird aggregiert
- Über spezielle Rechenverfahren werden MDS Analysen über die Ähnlichkeitsdaten mehrerer Personen (replicated MDS) durchgeführt

**Diskussion**

Vor-und Nachteile der Techniken abwägen