Materialien zum Modul *Fortgeschrittene Verfahren sozialwissenschaftlicher Datenanalyse*

Teil II: Methoden der Datenrepräsentation und Klassifikation

G. Rohwer

Version 2 **Oktober 2009**

Fakultät für Sozialwissenschaft
Ruhr-Universität Bochum, GB 1

goetz.rohwer@ruhr-uni-bochum.de

Vorbemerkung

Dieser Text soll als Grundlage für eine Lehrveranstaltung dienen, die sich mit Methoden der Datenrepräsentation und Klassifikation beschäftigt. Der Datenbegriff wird weit gefasst; es werden nicht nur statistische Daten mit einer Objekt-Variablen-Struktur betrachtet, sondern auch relationale und Netzwerkdaten.

Obwohl sich der Text schwerpunktmäßig mit Methoden beschäftigt, haben wir uns bemüht, sie ausführlich durch Beispiele zu illustrieren, die (in den meisten Fällen) praktisch nachvollzogen werden können. Soweit dafür Computerprogramme erforderlich sind, wird auf die verwendeten Skripte (zumeist für das Programm TDA) hingewiesen.[1] Anhang B enthält Hinweise zu den verwendeten Statistikprogrammen.

Die meisten Kapitel erfordern keine besonderen mathematischen Kenntnisse. Eine Ausnahme bilden die Abschnitte, in denen Methoden besprochen werden, die auf linearer Algebra beruhen (wie beispielsweise die Korrespondenzanalyse). Für ihr Verständnis sind Grundkenntnisse der Matrizenrechnung erforderlich. Die jeweiligen mathematischen Überlegungen haben wir jedoch in einen Anhang ausgegliedert, so dass zum Verständnis der Grundzüge der Methoden elementare Kenntnisse der Matrizenrechnung ausreichen.[2]

Hinweise zum Text

- Wie im Inhaltsverzeichnis angegeben wird, gliedert sich der Text in Kapitel, die meisten von ihnen auch in Abschnitte. Eine weitere Untergliederung in Paragraphen wird zu Beginn jedes Kapitels angegeben.
- Einfache Anführungszeichen werden zur Kennzeichnung sprachlicher Ausdrücke verwendet, doppelte Anführungszeichen werden verwendet, um Zitate kenntlich zu machen oder um anzudeuten, dass ein Ausdruck unklar ist und/oder metaphorisch verwendet wird. Innerhalb von Zitaten wird versucht, die im Original verwendeten Anführungszeichen zu reproduzieren. Wenn nicht anders angegeben, folgen Hervorhebungen in Zitaten stets dem Original; eigene Zusätze, Änderungen und Auslassungen werden durch eckige Klammern kenntlich gemacht.
- Wir unterscheiden die Zeichen '=' und ':='. Ein Gleichheitszeichen mit vorangestelltem Doppelpunkt wird verwendet, um anzudeuten, dass eine definitorische Gleichsetzung vorgenommen wird, d.h. der Ausdruck auf der linken Seite wird durch den Ausdruck auf der rechten Seite definiert. Dagegen dient ein einfaches Gleichheitszeichen zur Formulierung

[1] Einige der verwendeten TDA-Prozeduren wurden für diesen Text entwickelt oder modifiziert und stehen erst ab Version 6.4p zur Verfügung.

[2] Etwa im Umfang des Anhang *Rechnen mit Matrizen* bei Rohwer und Pötter (2002a).

einer Gleichheitsbehauptung und setzt deshalb voraus, dass beide Seiten schon definiert sind.

- Als Dezimalpunkt wird ein Punkt und nicht, wie im Deutschen üblich, ein Komma verwendet.
- Bei den Notationen aus der Mengenlehre und zum Funktionsbegriff folgen wir Rohwer und Pötter (2001, S. 21ff.); bei den Notationen zur Matrizenrechnung folgen wir Rohwer und Pötter (2002a).

Inhalt

Kapitel 1

Einleitung

1.1 Variablen, Abstände, Ähnlichkeiten

1. Statistische und relationale Variablen.
2. Abstände und Ähnlichkeiten.
3. Abstandsfunktionen für Merkmalsräume.
4. Abstandsfunktionen für Objektmengen.
5. Abstands- bzw. Ähnlichkeitsmatrizen.
6. Metrische und nichtmetrische Abstände.

1.2 Klassifikation und Typologien

1. Klassifikationen für Objektmengen.
2. Scharfe und unscharfe Klassifikationen.
3. Darstellung durch statistische Variablen.
4. Typologien für Merkmalsräume.
5. Verfahren zur Bildung von Klassifikationen.
6. Ansätze zur Konstruktion von Typologien.
7. Verwendungen des Typenbegriffs.
8. Zum Verständnis „fliessender Grenzen“.

1.3 Ansätze zur Datenrepräsentation

1. Aufgaben der Datenrepräsentation.
2. Darstellung von Häufigkeitsverteilungen.
3. Datenrepräsentation via Klassifikation.
4. Repräsentation von Ähnlichkeiten.
5. Darstellung relationaler Daten.

In diesem einleitenden Kapitel wird der formale Begriffsrahmen erläutert und werden einige der Fragestellungen angegeben, die in den folgenden Kapiteln genauer besprochen werden.

1.1 Variablen, Abstände, Ähnlichkeiten

1. Statistische und relationale Variablen. Als formalen Rahmen verwenden wir statistische und relationale Variablen. Eine *statistische Variable* hat die Form

$$X : \Omega \longrightarrow \mathcal{X} \tag{1.1}$$

Ω repräsentiert eine endliche Menge von Objekten (irgendeiner Art, es kann sich auch um abstrakte Objekte oder um Situationen handeln) und wird als *Referenzmenge* der Variablen bezeichnet. Die statistische Variable X ordnet jedem Element $\omega \in \Omega$ genau ein Element $X(\omega)$ des Merkmalsraums $\mathcal{X}$ zu (der als eine Charakterisierung von ω interpretiert werden kann). Es wird angenommen, dass Merkmalswerte durch Zahlen repräsentiert werden, also $\mathcal{X}$ als eine Teilmenge der reellen Zahlen aufgefasst werden kann.

Um Beziehungen zu erfassen, werden relationale Variablen verwendet. Eine (unimodale) *relationale Variable* hat die Form

$$R : \Omega \times \Omega \longrightarrow \mathcal{R} \tag{1.2}$$

Jeweils zwei Elementen der Referenzmenge Ω, etwa ω' und ω'', wird ein Wert $R(\omega', \omega'')$ in einem Merkmalsraum $\mathcal{R}$ zugeordnet, der als eine Information über das Vorhanden- oder Nichtvorhandensein einer Beziehung zwischen ω' und ω'' (und ggf. über die Art der Beziehung) interpretiert werden kann. Für den Merkmalsraum $\mathcal{R}$ wird wiederum angenommen, dass es eine numerische Repräsentation gibt. Im einfachsten Fall ist $\mathcal{R} = \{0, 1\}$, und die Werte geben an, ob eine Beziehung einer bestimmten Art besteht (1) oder nicht besteht (0).

2. Abstände und Ähnlichkeiten. Relationale Variablen liefern einen allgemeinen formalen Rahmen, um relationale Strukturen beliebiger Art zu repräsentieren. In diesem Text verwenden wir sie in erster Linie, um Ähnlichkeiten bzw. Unterschiede zwischen Objekten zu erfassen. Als Hilfsmittel dient der Begriff einer *Abstandsfunktion*. Eine Abstandsfunktion für eine beliebige Menge M ist eine Funktion

$$d : M \times M \longrightarrow \mathbf{R}$$

die für alle $m, m' \in M$ folgende drei Bedingungen erfüllt:

$$\text{a)}\ \ d(m, m') \geq 0 \tag{1.3}$$

$$\text{b)}\ \ d(m, m') = d(m', m) \tag{1.4}$$

$$\text{c)}\ \ d(m, m) = 0 \tag{1.5}$$

Das ist natürlich nur ein formaler Rahmen. In inhaltlicher Hinsicht wird angenommen, dass $d(m, m')$ als eine Größe interpretiert werden kann, die

Aufschluss über den Abstand zwischen m und m' oder über das Ausmaß ihrer Unterschiedlichkeit gibt: Je größer $d(m, m')$, desto größer der Abstand zwischen oder die Unterschiedlichkeit oder Unähnlichkeit von m und m'. Und umgekehrt: Je kleiner $d(m, m')$, desto näher oder ähnlicher sind sich m und m'.

3. Abstandsfunktionen für Merkmalsräume. Abstandsfunktionen können für beliebige Mengen definiert werden. Wir verwenden sie in diesem Text hauptsächlich für Merkmalsräume und Objektmengen. Abstandsfunktionen für Merkmalsräume erlauben es, über Abstände zwischen Merkmalswerten zu sprechen. Zum Beispiel könnte man bei einem Merkmalsraum zur Erfassung von Einkommen eine Abstandsfunktion durch absolute Einkommensdifferenzen definieren.

Im allgemeinen gibt es unterschiedliche Möglichkeiten, um Abstandsfunktionen zu definieren. Man betrachte zur Illustration einen Merkmalsraum $\mathcal{X} = \{1, 2, 3, 4, 5\}$, durch den fünf Schulabschlüsse unterschieden werden: 1 = ohne Hauptschulabschluss, 2 = Hauptschulabschluss, 3 = Realschulabschluss, 4 = Fachhochschulreife, 5 = Abitur. Eine Abstandsfunktion könnte beispielsweise durch $d(i, j) := |\, i - j \,|$ definiert werden. Offenbar gibt es viele andere Möglichkeiten.

4. Abstandsfunktionen für Objektmengen. Mit Abstandsfunktionen für Objektmengen wird versucht, das Reden über Unterschiede zwischen bzw. Ähnlichkeiten von Objekten zu präzisieren. Es gibt hauptsächlich drei Ansätze.

a) Hat man eine statistische Variable, etwa $X : \Omega \longrightarrow \mathcal{X}$, und gibt es bereits eine Abstandsfunktion für den Merkmalsraum $\mathcal{X}$, etwa d, dann können durch $d(X(\omega), X(\omega'))$ Abstände zwischen den Objekten definiert werden. Wir sprechen dann von der durch d und X *induzierten Abstandsfunktion*.

b) In einigen Fällen können Werte einer Abstandsfunktion direkt (ohne Umweg über statistische Variablen) ermittelt werden. Beispielsweise kann man an Daten über (soziale) Netzwerke denken.

c) Schließlich können Werte von Abstandsfunktionen auch durch subjektive Ratingverfahren erzeugt werden. Ein Verfahren besteht darin, Menschen zu bitten, vorgegebene Objekte paarweise zu vergleichen und eine Meinung über den Grad ihrer (irgendwie wahrgenommenen) Ähnlichkeit zu äußern.[1]

5. Abstands- bzw. Ähnlichkeitsmatrizen. Es ist oft praktisch, die Werte einer Abstandsfunktion für eine Menge M mit n Elementen durch eine

[1] Man vgl. hierzu etwa Green, Carmone und Smith (1989: 60ff.); Bortz und Döring (1995: 157f.).

quadratische (n,n)-Matrix $\mathbf{D} = (d_{ij})$ darzustellen, wobei d_{ij} den Abstand zwischen den Elementen i und j angibt.[2] Wir sprechen dann von einer *Abstands-* oder *Ähnlichkeitsmatrix*. So lässt sich beispielsweise die in § 3 definierte Abstandsfunktion für Schulabschlüsse durch eine Abstandsmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix} \tag{1.6}$$

erfassen. Man sieht auch sogleich, wie man beliebige andere Abstandsmatrizen bilden kann. Erforderlich ist nur, dass es sich um symmetrische Matrizen mit nichtnegativen Elementen handelt und dass alle Elemente in der Hauptdiagonale gleich Null sind.

Es sei angemerkt, dass Abstandsmatrizen bei praktischen Anwendungen unvollständig sein können, d.h. es kann vorkommen, dass für einige Elemente der Matrix Werte fehlen. Wir verwenden die Konvention, fehlende Werte durch negative Zahlen zu kennzeichnen.

6. Metrische und nichtmetrische Abstände. Abstandsfunktionen für Merkmalsräume sind oft metrisch, d.h. sie genügen zusätzlich zu den Bedingungen (1.1) – (1.3) auch der sogenannten *Dreiecksungleichung*:

$$\text{Für alle } m, m', m'' \in M: \; d(m,m') + d(m',m'') \geq d(m,m'') \tag{1.7}$$

und der folgenden Eindeutigkeitsbedingung:

$$\text{Wenn } d(m,m') = 0, \text{ dann ist } m = m'. \tag{1.8}$$

Man spricht dann von einer *metrischen Abstandsfunktion* oder kurz von einer *Metrik*. Zum Beispiel ist die in § 3 definierte Abstandsfunktion für Schulabschlüsse ($d(i,j) = |\, i - j \,|$) metrisch. Eine Menge von Objekten, für die eine Metrik definiert ist, wird auch ein *metrischer Raum* genannt.

Dagegen sind Abstandsfunktionen für Objektmenge oft nicht metrisch. Werden sie durch eine Metrik für einen Merkmalsraum induziert, erfüllen sie zwar die Dreiecksungleichung, oft jedoch nicht die Eindeutigkeitsbedingung (1.8). Als Beispiel kann man sich vorstellen, dass Abstände zwischen Schulabgängern durch Abstände zwischen ihren Schulabschlüssen definiert werden. In einigen Fällen, insbesondere wenn Abstandsfunktionen durch Ratingverfahren ermittelt werden, wird auch die Dreiecksungleichung (1.7) verletzt.[3]

[2] Wir folgen in diesem Text der Konvention, für Matrizen und Vektoren fettgedruckte Buchstaben zu verwenden. Vgl. den Anhang *Rechnen mit Matrizen* bei Rohwer und Pötter (2002a).

[3] Wenn bei einer Abstandsmatrix $\mathbf{D} = (d_{ij})$ die Dreiecksungleichung verletzt wird, kann man jedoch eine neue Abstandsmatrix mit Koeffizienten $d_{ij} + c$ (für $i \neq j$) bilden, wobei

$$c := \max\{0, \max_{i,j,k}\{d_{ij} - d_{ik} - d_{kj}\}\}$$

ist. Dann ist für die modifizierte Abstandsmatrix die Dreiecksungleichung erfüllt. Zur praktischen Durchführung kann die TDA-Prozedur `dmet` verwendet werden.

1.2 Klassifikationen und Typologien

1. Klassifikationen für Objektmengen. Es sei Ω eine Objektmenge. Unter einer *Klassifikation für die Objektmenge* Ω verstehen wir eine Einteilung von Ω in zwei oder mehr paarweise disjunkte Teilmengen:

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_m \tag{1.9}$$

Eine Klassifikation für Ω kann also auch als eine Partition von Ω bezeichnet werden. Hier sind einige einfache Beispiele:

a) Klassifikation einer Menge von Personen nach ihrem Geschlecht.

b) Klassifikation einer Menge von Studierenden nach ihrem Studiengang.

c) Klassifikation einer Menge von Haushalten durch eine Zuordnung zu Einkommensklassen.

2. Scharfe und unscharfe Klassifikationen. Klassifikationen in der in § 1 gegebenen Definition können auch als *scharfe Klassifikationen* bezeichnet werden, da jedes Element der Objektmenge genau einer Teilmenge zugeordnet wird. Lässt man dagegen zu, dass sich die Teilmengen, in die man eine Objektmenge Ω einteilt, überschneiden können, gelangt man zum Begriff einer *unscharfen Klassifikation*.[4] Wiederum können einige einfache Beispiele der Verdeutlichung dienen:

a) Einteilung einer Menge von Personen in drei Gruppen: Personen, die ein Fahrrad besitzen; Personen, die ein Auto besitzen; Personen, die weder ein Fahrrad noch ein Auto besitzen.

b) Ungenau erfasste Daten, die nur durch Teilmengen eines Merkmalsraums charakterisiert werden können.[5]

3. Darstellung durch statistische Variablen. Klassifikationen können durch statistische Variablen dargestellt werden. Ist eine scharfe Klassifikation der Form (1.9) gegeben, kann man sie durch eine statistische Variable

$$K : \Omega \longrightarrow \{1, \ldots, m\} \tag{1.10}$$

[4] Es sei angemerkt, dass der Ausdruck 'unscharfe Klassifikation' auch noch in einer anderen Bedeutung, nämlich als Bezeichnung für Fuzzy-Klassifikationen, verwendet wird; man vgl. etwa Höppner, Klawonn und Kruse (1997).

[5] Dazu ausführlich Rohwer und Pötter (2001: Teil V).

darstellen, wobei $K(\omega) = k$ gdw. $\omega \in \Omega_k$.[6] Handelt es sich um eine unscharfe Klassifikation, muss stattdessen eine Abbildung in die Potenzmenge, also

$$K^* : \Omega \longrightarrow \mathcal{P}(\{1, \ldots, m\}) \tag{1.11}$$

verwendet werden, wobei $K^*(\omega)$ die Indizes derjenigen Teilmengen von Ω angibt, denen ω zugerechnet werden kann.

4. Typologien für Merkmalsräume. Klassifikationen (in der hier verwendeten Definition) beziehen sich auf Mengen von Objekten. Die Objekte können beliebiger Art sein; insofern ist der Begriff sehr allgemein anwendbar. Eine besondere Terminologie bietet sich an, wenn es sich um einen Merkmalsraum handelt. Wir sprechen dann von einer (scharfen oder unscharfen) *Typologie* (für einen Merkmalsraum).[7] Der Merkmalsraum kann ein- oder mehrdimensional sein.[8]

Folgt man dieser Definition, betreffen Typologien Fragen der Begriffsbildung.[9] Weder setzen sie voraus noch implizieren sie Klassifikationen von (nicht-begrifflichen) Objekten. Ein Zusammenhang zwischen Typologien und Klassifikationen kann jedoch durch statistische Variablen hergestellt werden. Um das zu verdeutlichen, beziehen wir uns auf eine statistische Variable

$$X : \Omega \longrightarrow \mathcal{X} \tag{1.12}$$

Ist außerdem eine Typlogie $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_m$ gegeben, kann man mit ihrer Hilfe sofort eine Klassifikation von Ω bilden, indem man Teilmengen $X^{-1}(\mathcal{X}_j)$ verwendet. Wir nennen sie die durch die Typologie *induzierte Klassifikation.* Offenbar führt eine scharfe Typologie auch zu einer scharfen Klassifikation.

Ist umgekehrt eine Klassifikation einer Objektmenge Ω gegeben, etwa $\Omega = \Omega_1 \cup \cdots \cup \Omega_m$, kann man durch sie auch eine Typologie für den Merkmalsraum $\mathcal{X}$ definieren, indem man die Teilmengen $X(\Omega_j)$ verwendet. In diesem Fall erhält man jedoch durch eine scharfe Klassifikation nicht unbedingt eine scharfe Typologie und die Typologie wird von den jeweils verwendeten Objekten abhängig.

[6] 'gdw.' wird als Abkürzung für 'genau dann wenn' verwendet.

[7] Die Idee, Typologien und Typenbegriffe nicht auf Objekte, sondern auf Merkmalsräume zu beziehen, findet man bei zahlreichen Autoren, die sich mit methodischen Fragen der Typenbildung beschäftigt haben; man vgl. etwa Hempel und Oppenheim (1936); Lazarsfeld (1937); Ziegler (1973).

[8] In der Literatur findet man gelegentlich die Auffassung, dass von „eigentlichen Typenbegriffen" erst gesprochen werden sollte, wenn zwei oder mehr Merkmalsarten kombiniert werden; man vgl. etwa Hempel und Oppenheim (1936: 65f.), Lazarsfeld (1937: 120), Stinchcombe (1968: 43).

[9] Dies betont auch Bailey (1994: 4f., 66), um die Konstruktion von Typologien von der Klassifikation von Objekten zu unterscheiden.

Es sei betont, dass in diesem Text mit dem Wort 'Typologie' stets eine (scharfe oder unscharfe) Klassifikation eines Merkmalsraums gemeint ist. Liegt beispielsweise nur eine Ordnungsrelation für einen Merkmalsraum vor, wird diese nicht als eine Typologie bezeichnet.[10]

5. Verfahren zur Bildung von Klassifikationen. Es gibt sehr viele Verfahren zur Erzeugung von Klassifikationen.[11] Zwei Vorgehensweisen können grundsätzlich unterschieden werden:

a) Man kann von einer Typologie ausgehen und die jeweils gegebenen Objekte ihr entsprechend in (ggf. sich überlappende) Klassen einteilen.

b) Man kann Objekte klassifizieren, ohne eine Typologie vorauszusetzen. Die hierfür verwendeten Verfahren beruhen meistens darauf, dass zunächst eine Abstandsfunktion für die Objekte angenommen oder konstruiert wird; dann wird die Idee verfolgt, ähnliche Objekte zu Klassen zusammenzufassen.[12]

Wir besprechen Klassifikationsverfahren in den Kapiteln 7 – 8.

6. Verwendungen des Typenbegriffs. Das Wort 'Typ' wird in unterschiedlichen Bedeutungen verwendet. Orientiert man sich an der oben gegebenen Definition von Typologien, kann man *eine* Bedeutung fixieren: *Typen* sind Elemente von Typologien. Bei dieser Definition sind Typen Begriffe (Konstellationen begrifflicher Merkmale) und müssen von Objekten (womit hier stets nicht-sprachliche Objekte gemeint sind) unterschieden werden. Objekte können durch Typen *charakterisiert* werden; und andererseits können Typen durch Objekte *illustriert* (oder *exemplifiziert*) werden.[13]

Von diesem Typbegriff, bei dem man im engeren Sinne auch von *klassifizierenden Typen* sprechen könnte, unterscheiden wir *Idealtypen*, worunter wir in diesem Text explizit festgelegte Werte in einem Merkmalsraum verstehen, die dem Zweck dienen, gewissermaßen als Ankerpunkte für eine relationale Ordnung des Merkmalsraums zu dienen.

7. Zum Verständnis „fliessender Grenzen". Autoren, die sich mit Klassifikationen und Typologien beschäftigen, sprechen oft von „fliessenden

[10] Eine Ausdehnung des Typologiebegriffs, die auch Ordnungsrelationen einschließt, wurde von Hempel und Oppenheim (1936) vorgeschlagen.

[11] Aus der umfangreichen Literatur seien hier genannt: Anderberg (1973), Sneath und Sokal (1973); Bock (1974), Späth (1975), Lorr (1983), Jain und Dubes (1988), Kaufman und Rousseeuw (1990), Everitt (1993), Bacher (1994), Mirkin (1996).

[12] Manchmal wird diese Idee bereits verwendet, um „klassifizieren" zu erläutern; beispielsweise von Gordon (1987: 119): „Classification can be described as the activity of dividing a set of objects into a smaller number of classes in such a way that objects in the same class are similar to one another and dissimilar to objects in other classes."

[13] Anders als beispielsweise Sodeur (1974: 9) unterscheiden wir also auch Typen (als begriffliche Konstruktionen) und Cluster (als irgendwie abgegrenzte) Mengen von Objekten.

Grenzen“.[14] Wie kann man das verstehen?

1.3 Ansätze zur Datenrepräsentation

1. Aufgaben der Datenrepräsentation. In einer allgemeinen Formulierung kann man sagen, dass Verfahren der Datenrepräsentation die Aufgabe haben, eine Menge gegebener Daten übersichtlich und informativ darzustellen. In einer etwas engeren Bedeutung geht es um Verfahren, mit denen man Daten „anschaulich“ machen, also bildliche Darstellungen erzeugen kann. Wir sprechen dann von *graphischer* oder *bildlicher* Datenrepräsentation.

Als Ausgangspunkt kommen sowohl statistische als auch relationale Daten in Betracht. Es gibt keine strenge Unterscheidung zwischen diesen beiden Arten von Daten; denn oft können statistische Daten durch (konstruierte) Relationen ergänzt werden, und andererseits gibt es bei relationalen Daten oft zusätzliche statistische Informationen (über ihre Knoten). Wir werden uns deshalb in diesem Text nicht nur mit statistischen Daten im engeren Sinne beschäftigen, sondern auch mit relationalen Daten.

2. Darstellung von Häufigkeitsverteilungen. Die meisten Methoden zur Darstellung und Analyse statistischer Daten gehen von Häufigkeitsverteilungen aus. Hier können auch Methoden der Datenrepräsentation ansetzen. Einfache Methoden werden bereits in Einführungen in die Statistik vermittelt; zum Beispiel Darstellungen durch Verteilungsfunktionen und durch Streuungsdiagramme. Daran wird in Kapitel 3 angeknüpft. Besprochen werden dort Projektionsmethoden zur Darstellung von statistischen Datenmatrizen und Kontingenztabellen; dies umfasst als einen Spezialfall die sog. Korrespondenzanalyse.

3. Datenrepräsentation via Klassifikation. Die einfachen Verfahren der Datenrepräsentation können bei Daten mit mehr als zwei Dimensionen nicht mehr ohne weiteres verwendet werden. Ein naheliegender Ausweg besteht darin, die Objekte, für die die Daten gegeben sind, zunächst zu klassifizieren. Sobald das geschehen ist, kann man die Daten durch Häufigkeitsverteilungen für die zuvor konstruierten Klassen repräsentieren.[15]

4. Repräsentation von Ähnlichkeiten. Wenn eine Ähnlichkeits- oder Abstandsmatrix gegeben ist oder aus den Daten konstruiert werden kann, kann noch ein anderer Weg zur Datenrepräsentation verfolgt werden. Anstelle einer sofortigen Einteilung der Objekte in Klassen (und einer nachfolgenden Darstellung ihrer Häufigkeiten) kann man die Aufgabe darin sehen, die Ähnlichkeitsstruktur der Objekte darzustellen. Dafür gibt es hauptsächlich zwei Ansätze.

a) Eine Möglichkeit besteht darin, die vorgegebenen Abstände durch sichtbare Abstände in einer bildlichen Darstellung repräsentieren. Diese Idee wird mit Methoden der multidimensionalen Skalierung verfolgt; damit beschäftigen wir uns in Kapitel 4.

b) Ein anderer Ansatz verfolgt die Idee, Ähnlichkeitsstrukturen durch Graphen zu repräsentieren.

[14] Man vgl. beispielsweise die Hinweise bei Kluge (1999: 31); auch bereits Hempel und Oppenheim.

[15] Dies ist nur eine, jedoch unmittelbar naheliegende Möglichkeit, Verfahren der Clusteranalyse für Aufgaben der Datenrepräsentation und -analyse einzusetzen. Weitere Überlegungen über Beziehungen zwischen Statistik und Methoden der Clusteranalyse findet man bei Gower (1988).

Kapitel 2

Abstandskonstruktionen

Viele der in diesem Text besprochenen Methoden beziehen sich auf Abstände zur Erfassung von Ähnlichkeiten oder anderen Arten von Beziehungen. In einigen Fällen können Abstände unmittelbar erhoben werden; beispielsweise in Form von Daten über soziale Netzwerke oder wenn unmittelbar subjektive Ähnlichkeitsurteile erfragt werden.[1] Andererseits können Abstandsfunktionen auch aus anderen Arten von Daten, insbesondere aus statistischen Variablen, konstruiert werden. Einige Möglichkeiten werden in diesen Kapitel besprochen. Die Überlegungen erfolgen anhand von Beispielen, die in späteren Kapiteln zur Illustration von Verfahren der Datenrepräsentation und Klassifikation dienen.

[1] Zur direkten Erfragung von Ähnlichkeitsurteilen vgl. man beispielsweise Green, Carmone und Smith (1989: 60ff.).

2.1 Abstände zwischen Merkmalswerten

1. Objekte und Merkmalswerte. Es ist wichtig, zwischen Objekten und Merkmalswerten zu unterscheiden. Zwar kann der Objektbegriff so weit gefasst werden, dass man auch Merkmalswerte als abstrakte Objekte auffassen kann; dennoch bleibt eine Unterscheidung: Konkrete oder abstrakte Objekte gehören zum jeweils thematisierten Gegenstandsbereich; Merkmalswerte dienen dazu, die konkreten oder abstrakten Objekte zu charakterisieren.[2] Wichtig ist die Unterscheidung auch deshalb, weil Abstandskonstruktionen entweder bei Objekten oder bei Merkmalswerten ansetzen können.

Ein Beispiel wurde bereits in Abschnitt 1.1 (§ 3) angeführt: Schulabschlüsse. In diesem Beispiel kann man einerseits Abstände zwischen Schulabschlüssen definieren; dann konstruiert man eine Abstandsfunktion für einen Merkmalsraum (dessen Elemente Schulabschlüsse sind). Andererseits kann man auch unter Bezugnahme auf Schulabschlüsse Abstände zwischen Menschen definieren; dann konstruiert man eine Abstandsfunktion für Objekte (Menschen).

2. Definitionen für Merkmalsräume. Bei der Bildung von Abstandsfunktionen für Merkmalsräume kann man auf zwei wesentlich unterschiedliche Weisen vorgehen.

a) Man kann die Bildung einer Abstandsfunktion für einen Merkmalsraum als eine theoretische Aufgabe ansehen, die grundsätzlich unabhängig von Daten erfolgen sollte (was natürlich empirische Bezüge nicht ausschließt). Abstandsdefinitionen sollten dann insbesondere nicht von den Häufigkeiten abhängen, mit denen die Merkmalswerte in irgendwelchen empirisch fixierbaren Gesamtheiten vorkommen. Folgt man dieser Vorgehensweise, sollten zum Beispiel Abstände zwischen Schulabschlüssen unabhängig davon definiert werden, mit welchen Häufigkeiten die unterschiedlichen Schulabschlüsse vorkommen.

b) Andererseits gibt es daten- bzw. verteilungsabhängige Verfahren der Abstandskonstruktion für Merkmalsräume. Die resultierenden Abstandsfunktionen hängen dann auf kontingente Weise von den Daten

[2] Wir versuchen in diesem Text, auch sprachlich zwischen Objekten und Merkmalswerten zu unterscheiden. Mit dem Objektbegriff beziehen wir uns meistens auf eine der folgenden Arten:

a) *Individuelle Objekte*, die empirisch identifiziert werden können; zum Beispiel: Menschen, Tiere, Häuser.

b) *Institutionelle Objekte*, zum Beispiel: Haushalte, Unternehmen, Länder.

c) *Statistisch konstruierte Objekte*, zum Beispiel: Berufe, statistische Verteilungen.

In der Kategorie (c) gibt es scheinbar eine Überschneidung mit Merkmalswerten, denn beispielsweise Berufe können auch als Merkmalswerte verwendet werden. Werden Berufe jedoch als statistisch konstruierte Objekte verwendet, sind jeweils Gesamtheiten von Menschen gemeint, die einen Beruf innehaben bzw. ausüben.

Tabelle 2.1-1 Durch die Kemeny-Metrik definierte Abstände für Rangordnungen mit drei Alternativen.

		1	2	3	4	5	6	7	8	9	10	11	12	13
1	(1,2,3)	0	1	2	3	4	5	6	5	4	3	2	1	3
2	(1,1,3)	1	0	1	2	3	4	5	6	5	4	3	2	2
3	(2,1,3)	2	1	0	1	2	3	4	5	6	5	4	3	3
4	(2,1,2)	3	2	1	0	1	2	3	4	5	6	5	4	2
5	(3,1,2)	4	3	2	1	0	1	2	3	4	5	6	5	3
6	(3,2,2)	5	4	3	2	1	0	1	2	3	4	5	6	2
7	(3,2,1)	6	5	4	3	2	1	0	1	2	3	4	5	3
8	(2,2,1)	5	6	5	4	3	2	1	0	1	2	3	4	2
9	(2,3.1)	4	5	6	5	4	3	2	1	0	1	2	3	3
10	(2,3,2)	3	4	5	6	5	4	3	2	1	0	1	2	2
11	(1,3,2)	2	3	4	5	6	5	4	3	2	1	0	1	3
12	(1,2,2)	1	2	3	4	5	6	5	4	3	2	1	0	2
13	(2,2,2)	3	2	3	2	3	2	3	2	3	2	3	2	0

ab, die für die Konstruktion verwendet werden. Als ein Beispiel besprechen wir in Abschnitt 6.1 eine Skalierungsmethode, bei der Informationen über Häufigkeiten in einer Kontingenztabelle verwendet werden, um quantitative Scores und Abstände zwischen den jeweils verwendeten Kategorien zu berechnen.

Wir sprechen im ersten Fall von *verteilungsunabhängigen*, im zweiten Fall von *verteilungsabhängigen* Verfahren der Abstandskonstruktion.

3. Abstände für Rangordnungen. Zur Verdeutlichung einer verteilungsunabhängigen Abstandskonstruktion beziehen wir uns auf einen Merkmalsraum für Rangordnungen mit drei Alternativen. Es handelt sich um einen qualitativen Merkmalsraum. Gleichwohl ist es möglich, eine Abstandsfunktion zu definieren. Dafür kann die sog. *Kemeny-Metrik* verwendet werden.[3] Zum Beispiel gibt es bei drei Alternativen insgesamt 13 Rangordnungen, zwischen denen mit der Kemeny-Metrik Abstände berechnet werden können. Abbildung 2.1-1 zeigt für dieses Beispiel die Abstandsmatrix.

Offenbar handelt es sich um eine verteilungsunabhängige Abstandsdefinition, denn es wird nur auf Eigenschaften der Rangordnungen Bezug genommen, vollständig unabhängig davon, wie häufig die Rangordnungen in irgendeinem Anwendungsfall auftreten.

[3] Vgl. Rohwer und Pötter (2002a: Kap. 11).

2.2 Abstände zwischen Objekten

1. Verwendung von Merkmalswerten. Abstände zwischen Objekten können entweder direkt empirisch ermittelt oder aus Merkmalswerten der Objekte konstruiert werden. Hier beschäftigen wir uns mit der zweiten Möglichkeit. Gibt es bereits eine Abstandsfunktion für den Merkmalsraum, kann man für die Objekte eine induzierte Abstandsfunktion verwenden (vgl. Abschnitt 1.1, § 4). Oft besteht die Aufgabe jedoch darin, zur Bildung von Abständen gleichzeitig mehrere Merkmalswerte heranzuziehen; und selbst wenn es für jeden einzelnen Merkmalsraum eine Abstandsfunktion gibt, ergibt sich daraus nicht ohne weiteres eine Abstandsfunktion für den gemeinsamen Merkmalsraum.

Hier setzen Vorschläge an, wie man unter Verwendung mehrerer Merkmalswerte Abstandsfunktionen definieren kann. Es gibt sehr viele solcher Vorschläge, deren Anwendungsmöglichkeiten auch davon abhängen, von welcher Art die beteiligten Merkmalsräume sind.[4]

2. Notationen für Datenmatrizen. Um einige der Standarddefinitionen zu erläutern, beziehen wir uns auf eine statistische Variable mit m Komponenten:

$$(X_1, \ldots, X_m) : \Omega \longrightarrow \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$$

Wenn die Elemente von Ω durch $i = 1, \ldots, n$ numeriert werden, können die Daten durch folgende Datenmatrix erfasst werden:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nm} \end{pmatrix}$$

Jede Zeile entspricht einem Objekt; x_{ik} ist der Merkmalswert des Objekts ω_i bei der Variablen X_k bzw. im Merkmalsraum $\mathcal{X}_k$.[5]

Die Aufgabe besteht darin, Abstände für die Zeilen der Matrix $\mathbf{X}$ zu definieren.[6] Jede Zeile kann als ein Vektor aufgefasst werden, der aus m Zahlen besteht. Da wir Vektoren stets als Spaltenvektoren betrachten, verwenden wir folgende Notation:

$$\mathbf{x}_i := (x_{i1}, \ldots, x_{im})' \qquad (i = 1, \ldots, n)$$

Um eine Abstandsfunktion zu definieren, muss eine Funktion angegeben

[4] Übersichten findet man beispielsweise bei Bock (1974: Teil I); Fox (1982); Cox und Cox (1994: 8ff.); Bacher (1994: 198ff.); Batagelj und Bren (1995).

[5] Es sei an die Annahme erinnert, dass es für Merkmalswerte stets eine numerische Repräsentation gibt. Bei den Koeffizienten der Datenmatrix $\mathbf{X}$ handelt es sich also um Zahlen.

[6] Eine formal analoge, aber inhaltlich andere Fragestellung betrifft Zusammenhänge zwischen den Spalten einer Datenmatrix. Ein Beispiel wird in § 6 besprochen.

werden, die für alle Paare von Zeilen $\mathbf{x}_i$ und $\mathbf{x}_j$ eine Zahl $d(\mathbf{x}_i, \mathbf{x}_j)$ angibt, die als ein Abstand interpretiert werden kann und den in Abschnitt 1.1 (§ 2) angegebenen Bedingungen genügt.

3. Abstände für Datenmatrizen. Aus der großen Anzahl der in der Literatur vorgeschlagenen Definitionen von Abstandsfunktionen für statistische Datenmatrizen beschränken wir uns hier auf einige der am häufigsten verwendeten. Wir verwenden die eben erläuterte Notation und beziehen uns auf zwei Zeilen $\mathbf{x}_i$ und $\mathbf{x}_j$ der Datenmatrix $\mathbf{X}$.

a) Der *euklidische Abstand*

$$\| \mathbf{x}_i - \mathbf{x}_j \| := \left(\sum\nolimits_{k=1,m} (x_{ik} - x_{jk})^2 \right)^{1/2}$$

Da diese Definition die Bedingungen für eine Metrik erfüllt, wird auch von einer *euklidischen Metrik* gesprochen.

b) Die *Summe der absoluten Differenzen*

$$d^a(\mathbf{x}_i, \mathbf{x}_j) := \sum\nolimits_{k=1,m} |x_{ik} - x_{jk}|$$

Auch in diesem Fall handelt es sich um eine Metrik; sie wird auch *City-Block-Metrik* genannt.

c) Wenn alle beteiligten Variablen nur zwei Merkmalswerte (0 und 1) aufweisen können, wird der City-Block-Abstand auch als *Hamming-Distanz* bezeichnet. Der Abstand erfasst dann einfach die Anzahl der Merkmale, in denen $\mathbf{x}_i$ und $\mathbf{x}_j$ voneinander abweichen.

4. Gruppierte Daten und Abstände. Die Abstände zwischen identischen Zeilen einer Datenmatrix sind offenbar Null, und ihre Abstände zu allen übrigen Zeilen sind identisch. Es ist deshalb sinnvoll, eine Datenmatrix zunächst in eine gruppierte Form zu bringen, bevor aus ihr eine Abstandsmatrix erzeugt wird. Damit ist gemeint, dass eine neue Matrix gebildet wird, die nur unterschiedliche Zeilen enthält und außerdem eine Angabe über die Häufigkeit, mit der jede Zeile in der ursprünglichen Datenmatrix vorkommt. Man spricht dann auch von *gruppierten Daten*.

5. Illustration mit Klausurdaten. Als ein erstes Beispiel verwenden wir die Ergebnisse einer Statistikklausur, an der 46 Personen teilgenommen haben.[7] Box 2.2-1 zeigt die Daten. Die erste Spalte enthält eine fortlaufende Nummer; dann folgen die Punktzahlen für fünf Aufgaben, in denen maximal 10, 10, 5, 10 und 15 Punkte erzielt werden konnten. Wie man sieht,

[7] Diese Daten wurden bereits zur Illustration einiger Methoden der Datenkonstruktion bei Rohwer und Pötter (2002a: 76) verwendet.

Box 2.2-1 Klausurdaten (Datenfile `klaus1.dat`).

1	2	10	0	8	15	24	10	10	2	6	10
2	2	10	5	3	15	25	10	10	2	6	10
3	4	10	0	5	0	26	10	10	3	6	7
4	8	5	0	4	1	27	10	10	3	6	10
5	8	7	0	10	7	28	10	10	5	4	9
6	8	10	0	9	15	29	10	10	5	4	12
7	8	10	5	6	0	30	10	10	5	8	13
8	10	0	5	1	15	31	10	10	5	8	15
9	10	5	5	0	2	32	10	10	5	8	15
10	10	5	5	5	15	33	10	10	5	8	15
11	10	5	5	5	15	34	10	10	5	9	9
12	10	10	0	0	13	35	10	10	5	9	9
13	10	10	0	0	14	36	10	10	5	10	0
14	10	10	0	4	12	37	10	10	5	10	13
15	10	10	0	4	12	38	10	10	5	10	14
16	10	10	0	6	14	39	10	10	5	10	15
17	10	10	0	8	15	40	10	10	5	10	15
18	10	10	0	9	15	41	10	10	5	10	15
19	10	10	0	10	7	42	10	10	5	10	15
20	10	10	0	10	15	43	10	10	5	10	15
21	10	10	0	10	15	44	10	10	5	10	15
22	10	10	1	10	9	45	10	10	5	10	15
23	10	10	1	10	9	46	10	10	5	10	15

treten einige Zeilen mehrfach auf; insgesamt gibt es 31 unterschiedliche Zeilen.

Wie kann man Abstände zwischen den Zeilen definieren? In diesem Beispiel liegt es nahe, einen additiven Index zu verwenden, also in jeder Zeile die Summe der Punkte zu berechnen und dann deren Differenzen als Abstände zu nehmen. Dann wäre es auch einfach, das Ergebnis darzustellen; man könnte einfach die Verteilung der Indexwerte durch eine Häufigkeitstabelle oder eine Verteilungsfunktion darstellen.

Allerdings ist dies nur eine von vielen möglichen Abstandskonstruktionen. Wenn man die Unterschiede in der Bearbeitung der fünf Klausuraufgaben erfassen möchte, könnte der City-Block-Abstand verwendet werden. Bereits beim Vergleich der ersten beiden Zeilen zeigt sich dann ein Unterschied. Der additive Index liefert in beiden Fällen 35 Punkte, so dass der Abstand Null ist. Verwendet man dagegen die City-Block-Metrik beträgt der Abstand 10.

Als ein Beispiel, das gelegentlich in späteren Kapiteln verwendet wird, erzeugen wir eine Abstandsmatrix mit den City-Block-Abständen. Es handelt sich um eine Matrix mit 31 Zeilen und Spalten; wir nennen sie die *City-Block-Abstandsmatrix für die Klausurdaten*.[8]

Zu beachten ist, dass es in dieser Abstandsmatrix viele *Bindungen* gibt,

[8] Das Datenfile wird `kl6.dat` genannt. Es wurde aus den gruppierten Klausurdaten (`klaus1g.dat`) mit dem Skript `kl5.cf` erzeugt.

Tabelle 2.2-1 Häufigkeiten der Abstände im unteren Dreieck der aus den gruppierten Klausurdaten gebildeten City-Block-Abstandsmatrix.

Abstand	Häufigkeit	Abstand	Häufigkeit	Abstand	Häufigkeit
1	6	13	27	25	10
2	6	14	16	26	7
3	8	15	25	27	4
4	5	16	22	28	6
5	12	17	22	29	8
6	15	18	19	30	6
7	21	19	18	31	6
8	21	20	14	32	4
9	19	21	19	33	1
10	33	22	10	34	2
11	18	23	13	39	1
12	26	24	14	40	1

d.h. viele Abstände kommen mehrfach vor. Tabelle 2.2-1 zeigt eine Häufigkeitsverteilung der Abstände.[9]

6. Abstände zwischen Variablen. Statt Abstände zwischen den Zeilen kann man auch Abstände zwischen den Spalten (Variablen) einer Datenmatrix betrachten. In unserem Beispiel interessiert man sich dann für Abstände zwischen den Klausuraufgaben. Wiederum können solche Abstände auf viele unterschiedliche Weisen definiert werden.

Oft werden Korrelationskoeffizienten verwendet; Tabelle 2.2-2 zeigt sie für unser Beispiel. Man erkennt zum Beispiel, dass eine relative hohe Korrelation zwischen den Aufgaben 2 und 4 besteht. Korrelationskoeffizienten sind jedoch inhaltlich schwer zu interpretieren und liefern auch nicht unmittelbar Abstände. Ein großer Vorteil des Abstandsbegriffs liegt gerade darin, dass man Abstandsdefinitionen unter inhaltlichen Gesichtspunkten vornehmen kann; das wurde ja bereits im vorangegangenen Paragraphen deutlich. Dies trifft auch zu, wenn man sich für Unterschiede in der Art der Bearbeitung der Aufgaben interessiert. Man kann zum Beispiel aus den Klausurdaten in Box 2.2-1 folgende Kontingenztabelle bilden, die für jede Aufgabe angibt, wie gut sie gelöst worden ist:

$$\begin{pmatrix} 39 & 4 & 0 & 1 & 2 \\ 40 & 1 & 4 & 0 & 1 \\ 25 & 0 & 2 & 2 & 17 \\ 21 & 6 & 9 & 6 & 4 \\ 27 & 6 & 8 & 0 & 5 \end{pmatrix} \tag{2.1}$$

Die Zeilen $i = 1, \ldots, 5$ entsprechen den Aufgaben, die Spalten $j = 1, \ldots, 5$ geben an, wie gut eine Aufgabe gelöst wurde: 1 (sehr gut) bis 5 (sehr schlecht). Zum Beispiel wurde die dritte Aufgabe in 25 Fällen sehr gut, in

[9] Erzeugt mit den Skripten `kl6.cf` und `kl7.cf`.

Tabelle 2.2-2 Korrelationen zwischen den Spalten (Aufgaben) des Klausurdatenfiles in Box 2.2-1.

Aufgabe 1	1.0000	-0.0320	0.1985	0.1458	0.1521
Aufgabe 2	-0.0320	1.0000	-0.1254	0.4434	0.1055
Aufgabe 3	0.1985	-0.1254	1.0000	0.1229	0.1175
Aufgabe 4	0.1458	0.4434	0.1229	1.0000	0.1962
Aufgabe 5	0.1521	0.1055	0.1175	0.1962	1.0000

17 Fällen sehr schlecht (oder gar nicht) gelöst.

Zu überlegen bleibt, wie Abstände zwischen den Zeilen dieser Matrix definiert werden können. Unmittelbar euklidische oder City-Block-Abstände zu verwenden, wäre problematisch, denn die Zeilen der Matrix enthalten Häufigkeitsverteilungen. Es sollte deshalb überlegt werden, wie Abstände zwischen Verteilungen konzipiert werden können. Eine Möglichkeit wird im nächsten Abschnitt besprochen.

2.3 Abstände zwischen Verteilungen

In diesem Abschnitt erläutern wir zunächst Notationen für Kontingenztabellen und besprechen dann eine Möglichkeit, Abstände zwischen Verteilungen zu definieren. Zur Illustration werden zuerst Berufsstrukturdaten verwendet, dann werden auch die Klausurdaten aus dem vorangegangenen Abschnitt noch einmal aufgegriffen.

1. Notationen für Kontingenztabellen. Unter einer *Kontingenztabelle* verstehen wir die Darstellung der Verteilung einer p-dimensionalen statistischen Variablen $(X_1, \ldots, X_p) : \Omega \longrightarrow \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ $(p \geq 2)$ in Form einer Häufigkeitstabelle; bei den Komponenten kann es sich um qualitative oder quantitative Variablen handeln. Der gesamte Merkmalsraum besteht aus allen möglichen Kombinationen von Merkmalswerten:

$$\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_p = \{(x_1, \ldots, x_p) \,|\, x_1 \in \mathcal{X}_1, \ldots, x_p \in \mathcal{X}_p\}$$

Werden die Anzahlen der Merkmalswerte in den Komponenten durch $m_j := |\mathcal{X}_j|$ erfasst, gibt es insgesamt $m := m_1 \cdots m_k$ kombinierte Merkmalswerte. Die Kontingenztabelle liefert für jeden kombinierten Merkmalswert eine Häufigkeit $h_{x_1,\ldots,x_p} := |\{\omega \in \Omega \,|\, X_1(\omega) = x_1, \ldots, X_p(\omega) = x_p\}|$, die natürlich auch Null sein kann. Dies sind absolute Häufigkeiten; ihre Summe entspricht der Anzahl der Elemente von Ω. Stattdessen kann man auch relative Häufigkeiten

$$p_{x_1,\ldots,x_k} := \frac{1}{n} h_{x_1,\ldots,x_k}$$

betrachten (n ist hier die Anzahl der Elemente von Ω).

Eine zweidimensionale Kontingenztabelle ($p = 2$) kann man am übersichtlichsten in Form einer rechteckigen Matrix darstellen. Bei mehr als zwei Dimensionen ist das nicht möglich und man verwendet ein Schema der folgenden Art:

X_1	$\cdots$	X_p	Häufigkeit
x_1	$\cdots$	x_p	$h_{x_1,\ldots,x_p}$
$\vdots$		$\vdots$	$\vdots$

(2.2)

Jede Zeile gibt für eine Merkmalskombination die Häufigkeit an. Natürlich genügt es, Zeilen anzuführen, die mit einer positiven Häufigkeit vorkommen.

2. Berufsstrukturdaten. Als Beispiel verwenden wir eine dreidimensionale Kontingenztabelle mit Berufsstrukturdaten.[10] Den formalen Rahmen bildet eine dreidimensionale Variable $(X, Y, Z) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. X erfasst

[10] Die Daten wurden aus der Arbeit von Charles und Grusky (1995: 964) übernommen.

Box 2.3-1 Berufsstrukturdaten (Datenfile `bs1.dat`).

x	y	z	h(x,y,z)	x	y	z	h(x,y,z)
1	1	1	580983	5	1	1	9528
1	1	2	244868	5	1	2	6496
1	2	1	148629	5	2	1	3336
1	2	2	8310	5	2	2	660
1	3	1	437380	5	3	1	8048
1	3	2	209217	5	3	2	12408
1	4	1	709755	5	4	1	4236
1	4	2	30132	5	4	2	5232
1	5	1	833713	5	5	1	5444
1	5	2	65342	5	5	2	6652
1	6	1	3675554	5	6	1	30752
1	6	2	247207	5	6	2	5488
2	1	1	20029	6	1	1	6006
2	1	2	12440	6	1	2	7183
2	2	1	5296	6	2	1	869
2	2	2	789	6	2	2	231
2	3	1	18311	6	3	1	1089
2	3	2	14310	6	3	2	4675
2	4	1	21688	6	4	1	2057
2	4	2	6055	6	4	2	1793
2	5	1	18061	6	5	1	1628
2	5	2	8498	6	5	2	4994
2	6	1	93755	6	6	1	11407
2	6	2	14744	6	6	2	2453
3	1	1	290252	7	1	1	69007
3	1	2	177659	7	1	2	65988
3	2	1	69673	7	2	1	62813
3	2	2	3921	7	2	2	34936
3	3	1	294564	7	3	1	28041
3	3	2	330847	7	3	2	112601
3	4	1	110684	7	4	1	53110
3	4	2	141404	7	4	2	50809
3	5	1	115560	7	5	1	48578
3	5	2	235065	7	5	2	74607
3	6	1	913171	7	6	1	211708
3	6	2	151017	7	6	2	47986
4	1	1	2497820	8	1	1	28710
4	1	2	1639970	8	1	2	23661
4	2	1	1787150	8	2	1	19503
4	2	2	524560	8	2	2	1386
4	3	1	1011550	8	3	1	44649
4	3	2	2933700	8	3	2	50787
4	4	1	571510	8	4	1	51381
4	4	2	832620	8	4	2	23661
4	5	1	1004260	8	5	1	21780
4	5	2	2058480	8	5	2	21780
4	6	1	6796060	8	6	1	147312
4	6	2	1208680	8	6	2	55440

Tabelle 2.3-1 Anzahlen der im Datenfile `bs1.dat` (Box 2.3-1) erfassten Männer und Frauen.

	Land	Männer	Frauen	Insgesamt
1	Türkei	6386014	805076	7191090
2	Griechenland	177140	56836	233976
3	Schweiz	1793904	1039913	2833817
4	Grossbritannien	13668350	9198010	22866360
5	Deutschland	61344	36936	98280
6	Schweden	23056	21329	44385
7	USA	473257	386927	860184
8	Japan	313335	176715	490050
	Insgesamt	22896400	11721742	34618142

das Land und kann folgende Werte annehmen:

$$X = \begin{cases} 1 & \text{Türkei} \\ 2 & \text{Griechenland} \\ 3 & \text{Schweiz} \\ 4 & \text{Grossbritannien} \\ 5 & \text{Deutschland} \\ 6 & \text{Schweden} \\ 7 & \text{USA} \\ 8 & \text{Japan} \end{cases}$$

Y erfasst die Berufsgruppe und kann folgende Werte annehmen:

$$Y = \begin{cases} 1 & \text{Professional} \\ 2 & \text{Managerial} \\ 3 & \text{Clerical} \\ 4 & \text{Sales} \\ 5 & \text{Service} \\ 6 & \text{Production} \end{cases}$$

Z erfasst das Geschlecht und kann die Werte 1 (männlich) oder 2 (weiblich) annehmen. Box 2.3-1 zeigt die Daten. Die Darstellung entspricht dem Schema (2.2); es gibt insgesamt $8 \cdot 6 \cdot 2 = 96$ Zeilen (= Merkmalskombinationen).

3. Unterschiedliche Fragestellungen. Kontingenztabellen dienen zunächst zur Erfassung statistischer Verteilungen. Darüber hinaus können unterschiedliche Fragestellungen verfolgt werden, insbesondere:

a) Man kann untersuchen, wie Verteilungen einzelner Variablen von Werten anderer Variablen abhängen. Dafür werden meistens Methoden der Regressionsrechnung verwendet.

b) Man kann (durch Werte von Variablen bedingte) Verteilungen im Hinblick auf ihre Ähnlichkeit vergleichen. Bei der Kontingenztabelle aus § 2 kann man beispielsweise für jedes Land die Verteilung von Männern und Frauen auf die sechs Berufsgruppen vergleichen, um das unterschiedliche Ausmaß der geschlechtsspezifischen beruflichen Segregation zu erfassen.

c) Man kann die in der Kontingenztabelle erfassten Häufigkeiten verwenden, um die beteiligten Merkmalsräume zu charakterisieren. Die Idee ist, sich auf Abstände zwischen den Merkmalswerten entsprechenden Verteilungen zu beziehen. Beispiele werden in den Paragraphen 5–7 angegeben.

Tabelle 2.3-2 Nach Ländern (1 – 8) differenzierte Verteilungen der Männer und Frauen auf die Berufsgruppen (in %). Berechnet aus den Daten in Box 2.3-1.

y	z	1	2	3	4	5	6	7	8
1	1	8.08	8.56	10.24	10.92	9.69	13.53	8.02	5.86
1	2	3.41	5.32	6.27	7.17	6.61	16.18	7.67	4.83
2	1	2.07	2.26	2.46	7.82	3.39	1.96	7.30	3.98
2	2	0.12	0.34	0.14	2.29	0.67	0.52	4.06	0.28
3	1	6.08	7.83	10.39	4.42	8.19	2.45	3.26	9.11
3	2	2.91	6.12	11.67	12.83	12.63	10.53	13.09	10.36
4	1	9.87	9.27	3.91	2.50	4.31	4.63	6.17	10.48
4	2	0.42	2.59	4.99	3.64	5.32	4.04	5.91	4.83
5	1	11.59	7.72	4.08	4.39	5.54	3.67	5.65	4.44
5	2	0.91	3.63	8.29	9.00	6.77	11.25	8.67	4.44
6	1	51.11	40.07	32.22	29.72	31.29	25.70	24.61	30.06
6	2	3.44	6.30	5.33	5.29	5.58	5.53	5.58	11.31

4. Der Dissimilaritätsindex. Zuvor besprechen wir eine einfache Möglichkeit zur Definition von Abständen zwischen Verteilungen mit dem *Dissimilaritätsindex*. Sind zwei Verteilungen

$$\mathbf{p}_i = (p_{i1}, \ldots, p_{im})' \quad \text{und} \quad \mathbf{p}_j = (p_{j1}, \ldots, p_{jm})'$$

mit relativen Häufigkeiten ($\Sigma_k p_{ik} = \Sigma_k p_{jk} = 1$) gegeben, ist der Dissimilaritätsindex folgendermaßen definiert:

$$\mathrm{DI}(\mathbf{p}_i, \mathbf{p}_j) := \frac{1}{2} \sum\nolimits_{k=1,m} |p_{ik} - p_{jk}| \tag{2.3}$$

Offenbar enstpricht dieser Index gerade der Hälfte des City-Block-Abstandes zwischen $\mathbf{p}_i$ und $\mathbf{p}_j$.

Unterstellt man für die beiden Verteilungen eine gemeinsame Referenzmenge, kann der Dissimilaritätsindex als Anteil derjenigen Objekte aus der Referenzmenge aufgefasst werden, die umverteilt (einem anderen Merkmalswert zugeordnet) werden müssten, um die beiden Verteilungen in Übereinstimmung zu bringen. Es handelt sich insofern um eine einfache Variante einer Substitutionsmetrik.

5. Länderspezifische Berufsstrukturen. Zur Illustration berechnen wir aus

Tabelle 2.3-3 Abstandsmatrix mit Dissimilaritätsindizes (Datenfile: `bs3.dat`), berechnet aus den Verteilungen in Tabelle 2.3-2.

0.0000	0.1551	0.3235	0.3761	0.3142	0.4230	0.3901	0.3040
0.1551	0.0000	0.1801	0.2486	0.1663	0.2950	0.2645	0.1653
0.3235	0.1801	0.0000	0.1127	0.0520	0.1746	0.1696	0.1458
0.3761	0.2486	0.1127	0.0000	0.1027	0.1664	0.1002	0.2027
0.3142	0.1663	0.0520	0.1027	0.0000	0.1821	0.1328	0.1342
0.4230	0.2950	0.1746	0.1664	0.1821	0.0000	0.1769	0.2623
0.3901	0.2645	0.1696	0.1002	0.1328	0.1769	0.0000	0.2134
0.3040	0.1653	0.1458	0.2027	0.1342	0.2623	0.2134	0.0000

Tabelle 2.3-4 Die Berufsstrukturdaten aus Box 2.3-1 in Form einer zweidimensionalen Kontingenztabelle (Datenfile: `bs4.dat`). Die Spalten entsprechen den sechs Berufsgruppen.

		Profess.	Manag.	Clerical	Sales	Service	Product.
M1	Türkei	580983	148629	437380	709755	833713	3675554
M2	Griechenland	20029	5296	18311	21688	18061	93755
M3	Schweiz	290252	69673	294564	110684	115560	913171
M4	Grossbritannien	2497820	1787150	1011550	571510	1004260	6796060
M5	Deutschland	9528	3336	8048	4236	5444	30752
M6	Schweden	6006	869	1089	2057	1628	11407
M7	USA	69007	62813	28041	53110	48578	211708
M8	Japan	28710	19503	44649	51381	21780	147312
F1	Türkei	244868	8310	209217	30132	65342	247207
F2	Griechenland	12440	789	14310	6055	8498	14744
F3	Schweiz	177659	3921	330847	141404	235065	151017
F4	Grossbritannien	1639970	524560	2933700	832620	2058480	1208680
F5	Deutschland	6496	660	12408	5232	6652	5488
F6	Schweden	7183	231	4675	1793	4994	2453
F7	USA	65988	34936	112601	50809	74607	47986
F8	Japan	23661	1386	50787	23661	21780	55440

den Berufsstrukturdaten eine Abstandsmatrix mit dem Dissimilaritätsindex. In einem ersten Schritt wird für jedes Land eine Verteilung der Männer und Frauen auf die Berufsgruppen ermittelt (Tabelle 2.3-2). Dann werden mit dem Dissimilaritätsindex Abstände zwischen den Verteilungen berechnet; Tabelle 2.3-3 zeigt die Abstandsmatrix.[11] Sie wird später mit unterschiedlichen Methoden genauer analysiert.

6. Geschlechtsspezifische Verteilungen. Eine andere interessante Möglichkeit besteht darin, die geschlechtsspezifischen Verteilungen auf die Berufsgruppen zu vergleichen. Als Ausgangspunkt dient in diesem Fall eine aus den Daten in Box 2.3-1 erzeugte zweidimensionale Kontingenztabelle, wie sie in Tabelle 2.3-4 gezeigt wird.[12] Jede Zeile repräsentiert eine für ein

[11] Das Datenfile wird `bs3.dat` genannt; es wurde mit dem Skript `bs3.cf` erzeugt.

[12] Das Datenfile wird `bs4.dat` genannt.

Tabelle 2.3-5 Abstandsmatrix mit Dissimilaritätsindizes für die Klausuraufgaben (Datenfile: `ka1b.dat`), berechnet aus den Zeilen der Matrix (2.1) in Abschnitt 2.2 (§ 6).

Aufgabe 1	0.0000	0.1087	0.3913	0.3913	0.2826
Aufgabe 2	0.1087	0.0000	0.3913	0.4130	0.2826
Aufgabe 3	0.3913	0.3913	0.0000	0.3696	0.3043
Aufgabe 4	0.3913	0.4130	0.3696	0.0000	0.1522
Aufgabe 5	0.2826	0.2826	0.3043	0.1522	0.0000

Land und ein Geschlecht spezifische Verteilung auf die sechs Berufsgruppen. Werden dann mithilfe des Dissimilaritätsindex Abstände zwischen den Zeilen berechnet, erhält man eine Abstandsmatrix mit 16 Zeilen und Spalten.[13] Auch diese Abstandsmatrix wird in späteren Abschnitten mit unterschiedlichen Methoden genauer untersucht.

7. Abstände zwischen Klausuraufgaben. Der Dissimilaritätsindex kann auch verwendet werden, um Abstände für die Zeilen der Matrix (2.1), die in Abschnitt 2.2 (§ 6) definiert wurde, zu gewinnen. Tabelle 2.3-5 zeigt die Abstandsmatrix.[14] Sie wird in den Abschnitten 4.2 (§ 4) und **??** verwendet. Weitere Analysen der Datenmatrix (2.1) erfolgen in Abschnitt 3.3.

[13] Das Datenfile wird `bs4b.dat` genannt; es wurde mit `bs4.cf` erzeugt.

[14] Sie wird `ka1b.dat` genannt; berechnet mit dem Skript `kl9.cf`.

Kapitel 3

Räumliche Bilder

3.1 Streuungsdiagramme

1. Einfache Streuungsdiagramme.
2. Abstände in Streuungsdiagrammen.
3. Sich überlagernde Punkte.
4. Regressionslinien.
5. Zusätzliche Unterscheidungen.
6. Serien von Boxplots.

3.2 Projektionsverfahren

1. Projektion von Datenmatrizen.
2. Die Singularwertzerlegung.
3. Optimale lineare Projektionen.
4. Projektion der Berufsstrukturdaten.
5. Beurteilung der Projektionsgüte.
6. Projektionen großer Datenmengen.

3.3 Korrespondenzanalyse

1. Der theoretische Ansatz.
2. Illustration des Rechenverfahrens.
3. Unterschiedliche Abstandskonzeptionen.
4. Beurteilung der Projektionsgüte.
5. Korrespondenzanalyse der Berufsstrukturdaten.

In diesem Kapitel beginnen wir, uns mit Methoden der Datenrepräsentation zu beschäftigen, die räumliche Darstellungen intendieren. Damit sind Darstellungen gemeint, bei denen sich durch räumliche Abstände in einem Bild Hinweise auf Eigenschaften der für die Darstellung verwendeten Daten gewinnen lassen.

Die Methoden, um solche räumlichen Darstellungen zu erzeugen, sind sehr vielfältig. In diesem Kapitel beginnen wir mit Streuungsdiagrammen. Dann werden einige Methoden besprochen, die räumliche Bilder durch Projektionen erzeugen, die mit Hilfsmitteln der linearen Algebra konstruiert werden. Zunächst besprechen wir in Abschnitt 3.2 den allgemeinen Ansatz, dann in Abschnitt 3.3 eine spezielle Variante, die unter dem Namen Korrespondenzanalyse bekanntgeworden ist.

3.1 Streuungsdiagramme

1. Einfache Streuungsdiagramme. Hier die Grundidee erläutern und anhand eines einfachen Beispiels illustrieren.

Offenbar können mit einem Streuungsdiagramm Daten immer nur im Hinblick auf jeweils zwei Dimensionen sichtbar gemacht werden. Weitere Probleme können aus der Art der Merkmalsräume resultieren. Bei qualitativen Merkmalsräumen ist nicht nur die Reihenfolge der Merkmalswerte beliebig, es gibt auch oft nur wenige Merkmalsausprägungen, so dass sich die Punkte in einem Streuungsdiagramm in vielen Fällen überlagern können. Hat man zum Beispiel Daten in Form einer zweidimensionalen Kontingenztabelle, ist es meistens nicht sinnvoll, ein Streuungsdiagramm zu erzeugen.

2. Abstände in Streuungsdiagrammen. Von besonderer Bedeutung ist, ob bzw. wie sich Abstände zwischen Punkten in Streuungsdiagrammen interpretieren lassen. Dies hängt sowohl von der Richtung der Vergleiche als auch von der Art der den Achsen zugeordneten Merkmalsräume ab.

a) *Vergleiche in einer Achsenrichtung:*

   Wenn es sich um einen qualitativen Merkmalsraum handelt, haben Abstände keine Bedeutung.

   Wenn es sich um einen ordinalen Merkmalsraum handelt (quantitativ, ohne eine Metrik zu unterstellen), können ordinale Abstandsvergleiche vorgenommen werden.

   Wenn es sich um einen quantitativen und metrischen Merkmalsraum handelt, können Abstandsvergleich entsprechend der verwendeten Metrik interpretiert werden.

b) *Vergleiche in diagonaler Richtung.*

   Möglichkeiten, Abstände zu vergleichen, hängen dann von der Art der Merkmalsräume für beide Achsebn des Streuungsdiagramms ab. Euklidische Abstände zwischen Punkten können nur sinnvoll interpretiert werden, wenn es in beiden Achsen einen quantitativen metrischen Merkmalsraum gibt. Aber auch dann hängt die Bedeutung der Abstände auch noch von den jeweils gewählten Skalierungen der Achsen ab.

3. Sich überlagernde Punkte. Um Streuungsdiagramme trotz einer Überlagerung von Punkten informativ zu gestalten, sind manchmal folgende Möglichkeiten zweckmäßig:

a) Kleine zufällige Verschiebungen der Punkte, um Überlagerungen zu vermeiden.

b) Sunflower-Diagramme.

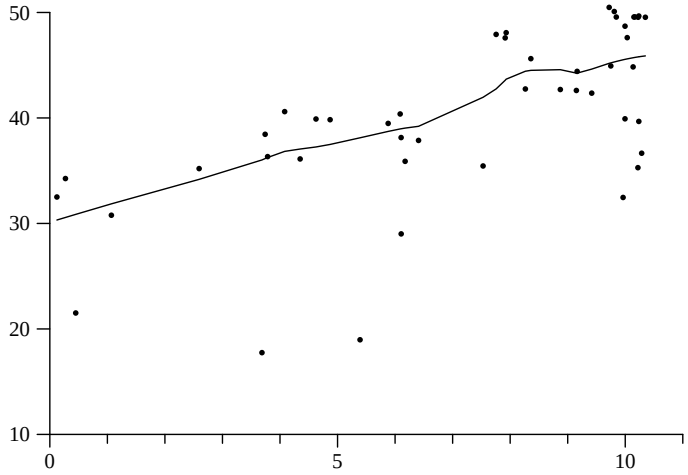


Abb. 3.1-1 Streuungsdiagramm für die Klausurdaten aus Abschnitt 2.2. X-Achse: Punktzahl bei der vierten Aufgabe, Y-Achse: Gesamtpunktzahl. Koordinaten mit einer zufälligen Streuung versehen. Zusätzlich ist eine Regressionslinie eingezeichnet (vgl. § 3).

Zur Illustration verwenden wir die Klausurdaten aus Abschnitt 2.2. Durch ein Streuungsdiagramm soll der Zusammenhang zwischen den in der vierten Aufgabe erzielten Punkten und der Gesamtpunktzahl sichtbar gemacht werden. In diesem Fall überlagern sich viele Punkte, was durch ein einfaches Streuungsdiagramm nicht sichtbar wird. Abbildung 3.1-1 zeigt ein Streuungsdiagramm, bei dem die Koordinaten mit einer zufälligen Streuung versehen wurden;[1] Abbildung 3.1-2 zeigt das Streuungsdiagramm mithilfe von Sunflower-Symbolen.[2]

4. Regressionslinien. Manchmal kann es informativ sein, zusätzlich Regressionslinien in ein Streuungsdiagramm einzuzeichnen. Dafür stehen verschiedene Verfahren zur Verfügung. Oft bieten sich lokale (nichtparametrische) Verfahren an.

Zur Illustration verwenden wir noch einmal die Klausurdaten. In Abbildung 3.1-1 haben wir eine Regressionslinie eingezeichnet, die mit einem nichtparametrischen Regressionsverfahren erzeugt wurde.

5. Zusätzliche Unterscheidungen. Manchmal gehören die in einem Streuungsdiagramm dargestellten Punkte unterschiedlichen Gruppen an und es

[1] Zu den X- und Y-Koordinaten wurden jeweils Werte einer im Intervall von -0.5 bis $+0.5$ gleichverteilten Zufallsvariablen addiert.

[2] Die TDA-Skripte sind `klplot1.cf` und `klplot2.cf`. Verwendet wurde die Prozedur `scplot` für Streuungsdiagramme.

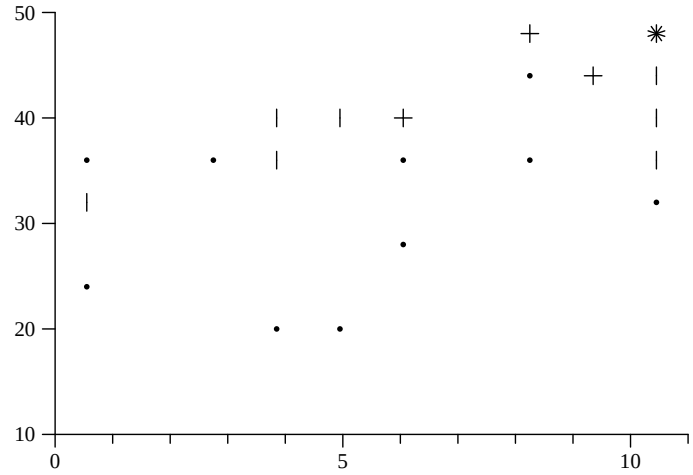


Abb. 3.1-2 Streuungsdiagramm für die Klausurdaten aus Abschnitt 2.2. X-Achse: Punktzahl bei der vierten Aufgabe, Y-Achse: Gesamtpunktzahl. Dargestellt mithilfe von Sunflower-Symbolen.

kann informativ sein, dies durch unterschiedliche Symbole graphisch sichtbar zu machen. Als Beispiel kann man sich Wähler von Parteien in einem Streuungsdiagramm vorstellen, dessen Achsen Bildung und Einkommen repräsentieren. Wenn dann für jede Partei ein unterschiedliches Symbol verwendet wird, erkennt man ihre Verteilungen in dem zweidimensionalen Merkmalsraum.

6. Serien von Boxplots. Ein weiteres Hilfsmittel sind Boxplots. Sie können auch innerhalb eines Diagramms verwendet werden, in dem für jeden Wert auf der X-Achse ein separater Boxplot für die Verteilung der zugeordneten Y-Werte dargestellt wird.

3.2 Projektionsverfahren

1. Projektion von Datenmatrizen. Jetzt beziehen wir uns auf eine Datenmatrix, die für n Objekte Werte für m Variablen liefert; zur Notation verwenden wir

$$\mathbf{X} := \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

Mithilfe von Streuungsdiagrammen könnten immer nur jeweils zwei Variablen (Spalten von $\mathbf{X}$) dargestellt werden. Ein grundsätzlich anderer Ansatz geht von der Vorstellung aus, dass die Zeilen von $\mathbf{X}$, für die wir die Notation

$$\mathbf{x}_i := (x_{i1}, \ldots, x_{im})' \qquad (\text{für } i = 1, \ldots, n)$$

verwenden, auch als Elemente eines m-dimensionalen Zahlenraums aufgefasst werden können: $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbf{R}^m$. Dann kann man die Idee verfolgen, diese Vektoren auf einen Zahlenraum mit wenigen Dimensionen zu projizieren, um Zusammenhänge sichtbar zu machen. In diesem Abschnitt besprechen wir eine Möglichkeit, solche Projektionen zu berechnen.

2. Die Singularwertzerlegung. Ein wichtiges Hilfsmittel zur Berechnung linearer Projektionen ist die *Singularwertzerlegung*. Damit ist gemeint, dass für jede (n, m)-Matrix $\mathbf{X}$ eine Darstellung

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' \tag{3.1}$$

gefunden werden kann, so dass gilt (wir setzen hier für die Notation voraus, dass $n \geq m$ ist):

a) $\mathbf{V}$ ist eine orthogonale (m, m)-Matrix: $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_m$.[3]

b) $\mathbf{U}$ ist eine (n, m)-Matrix, für die gilt: $\mathbf{U}'\mathbf{U} = \mathbf{I}_m$; jedoch ist im allgemeinen $\mathbf{U}\mathbf{U}' \neq \mathbf{I}_n$.

c) $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m)$ ist eine (m, m)-Diagonalmatrix. Die Werte $\lambda_1, \ldots, \lambda_m$ heißen die *Singularwerte* der Matrix $\mathbf{X}$. Es gilt, dass die Anzahl der von Null verschiedenen Singularwerte mit dem Rang der Matrix $\mathbf{X}$ identisch ist.

Zur Illustration betrachten wir folgende Matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \\ 1 & 4 & 8 \end{pmatrix} \tag{3.2}$$

[3] $\mathbf{I}_m$ bezeichnet eine Einheitsmatrix mit m Zeilen und Spalten.

Sie hat den Rang 2, da die letzten beiden Spalten linear abhängig sind. Eine Singularwertzerlegung liefert:

$$\mathbf{\Lambda} = \begin{pmatrix} 12.3834 & 0.0000 & 0.0000 \\ 0.0000 & 0.8075 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} -0.1905 & 0.8147 & -0.4236 \\ -0.3691 & 0.4047 & 0.3060 \\ -0.5477 & -0.0053 & 0.6588 \\ -0.7263 & -0.4154 & -0.5412 \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} -0.1481 & 0.9890 & 0.0000 \\ -0.4423 & -0.0662 & 0.8944 \\ -0.8846 & -0.1324 & -0.4472 \end{pmatrix}$$

Dass die Matrix $\mathbf{X}$ den Rang 2 hat, kann man daran sehen, dass nur zwei Singularwerte ungleich Null sind. Wenn man sich die Mühe macht, kann man feststellen: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$. Außerdem: $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_3$ und $\mathbf{U}'\mathbf{U} = \mathbf{I}_3$. Jedoch ist $\mathbf{U}\mathbf{U}' \neq \mathbf{I}_4$.[4]

3. Optimale lineare Projektionen. Eine Singularwertzerlegung der Datenmatrix $\mathbf{X}$ kann zur Konstruktion von Projektionen verwendet werden. Zunächst wird im Anschluss an (3.1) eine Koordinatentransformation

$$\mathbf{P} := \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Lambda} \tag{3.3}$$

durchgeführt. D.h. anstelle der Zeilen von $\mathbf{X}$ werden die Zeilen von $\mathbf{P}$ betrachtet. Wichtig ist, dass sich dadurch die euklidischen Abstände nicht verändern. Verwendet man nämlich $\mathbf{p}_i := (p_{i1}, \ldots, x_{im})'$ für die Zeilen von $\mathbf{P}$, gilt $\mathbf{p}_i' = \mathbf{x}_i'\mathbf{V}$, und man findet für die Abstände zwischen jeweils zwei Zeilen:

$$\| \mathbf{p}_i' - \mathbf{p}_j' \| = \| \mathbf{x}_i'\mathbf{V} - \mathbf{x}_j'\mathbf{V} \| = \| \mathbf{x}_i' - \mathbf{x}_j' \| \tag{3.4}$$

weil $\mathbf{V}$ eine orthogonale Matrix ist. Die zweite Überlegung betrifft die Auswahl von Dimensionen für die Projektion. Die Zeilen von $\mathbf{P}$ haben ebenso wie die von $\mathbf{X}$ m Dimensionen. Um sie beispielsweise in einer Ebene sichtbar zu machen, können nur zwei davon verwendet werden. Welche? Nehmen wir an, dass die Dimensionen j_1 und j_2 ausgewählt werden, d.h. dass die Punkte

$$\mathbf{p}_{i(j_1,j_2)} := (p_{ij_1}, p_{ij_2})'$$

[4] Für die Berechnungen wurde das Skript `pv1.cf` verwendet. Die Matrix (3.2) befindet sich im Datenfile `pv1.dat`.

für die graphische Darstellung verwendet werden. Da für die quadrierte Länge

$$\| \mathbf{p}'_i \|^2 = \sum_{j=1,m} (u_{ij}\lambda_j)^2$$

gilt, hat der durch die Auswahl der Dimensionen j_1 und j_2 sichtbare Teil die quadrierte Länge

$$\| \mathbf{p}'_{i(j_1,j_2)} \|^2 = u^2_{ij_1}\lambda^2_{j_1} + u^2_{ij_2}\lambda^2_{j_2}$$

Ein plausibles Kriterium für die Auswahl der Dimensionen ist, dass die sichtbare Länge im Durchschnitt aller Punkte möglichst groß wird, also

$$\sum_{i=1,n} \| \mathbf{p}'_{i(j_1,j_2)} \|^2 \longrightarrow \max$$

Nun gilt jedoch wegen der Orthogonalität von $\mathbf{U}$

$$\sum_{i=1,n} \| \mathbf{p}'_{i(j_1,j_2)} \|^2 = \lambda^2_{j_1} + \lambda^2_{j_2} \tag{3.5}$$

Also wird man für die Projektion diejenigen Dimensionen auswählen, die zu den beiden größten Singularwerten von $\mathbf{X}$ gehören.

4. Projektion der Berufsstrukturdaten. Zur Illustration verwenden wir die Berufsstrukturdaten aus Abschnitt 2.3, und zwar in der in Tabelle 2.3-4 angegebenen Form. Jede Zeile enthält eine für ein Land und ein Geschlecht spezifische Verteilung auf die sechs Berufsgruppen. Die gesamte Tabelle kann als eine $(16, 6)$-Matrix betrachtet werden, die wir $\mathbf{B}$ nennen.

Offenbar muss zunächst überlegt werden, welche Vektoren bzw. Abstände bildlich dargestellt werden sollen. Wegen der für die Länder stark unterschiedlichen absoluten Häufigkeiten, verwenden wir die zeilenspezifischen Verteilungen auf die sechs Berufsgruppen, betrachten also eine Matrix $\mathbf{X} = (x_{ij})$ mit den Elementen $x_{ij} := b_{ij}/b_{i.}$.[5] Die so gebildeten Zeilen von $\mathbf{X}$ werden auch als *Zeilenprofile* (von $\mathbf{B}$) bezeichnet. Ganz analog kann man auch von den *Spaltenprofilen* einer Matrix sprechen.

Für die Matrix $\mathbf{X}$ wird nun eine Singularwertzerlegung durchgeführt: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$.[6] Es gibt folgende Singularwerte:

$$\lambda_1 = 1.898,\ \lambda_2 = 0.756,\ \lambda_3 = 0.280,\ \lambda_4 = 0.181,\ \lambda_5 = 0.142,\ \lambda_6 = 0.095$$

Für eine zweidimensionale Projektion sollten also die ersten beiden Dimensionen verwendet werden. Box 3.2-1 zeigt die Koordinaten,[7] Abbildung 3.2-1 zeigt die mit diesen Koordinaten gezeichneten Punkte.[8]

[5] $b_{i.}$ ist die Summe der Elemente in der $i$ten Zeile von $\mathbf{B}$.

[6] Wir haben dafür die TDA-Prozedur `dma` (Option 7) verwendet. Das Skript heißt `ka3.cf`, die Datenmatrix aus Tabelle 2.3-4 wurde `bs4.dat` genannt.

[7] Diese Koordinaten, die hier unverändert aus der Singularwertzerlegung resultieren, können natürlich noch beliebigen linearen Transformationen unterzogen werden, ohne dass sich die euklidischen Abstände verändern.

[8] Die Abbildung wurde mit dem Skript `kaplot3b.cf` erzeugt.

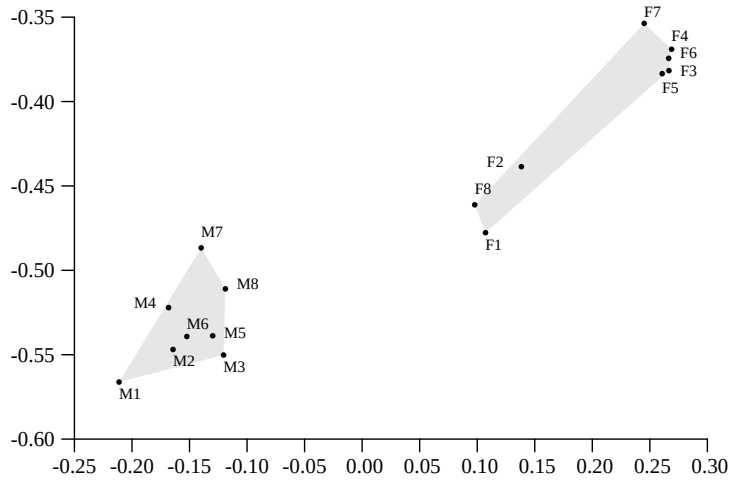


Abb. 3.2-1 Projektion der Zeilen der Matrix $\mathbf{X}$ mithilfe der Koordinaten aus Box 3.2-1 ($\mathbf{X}\mathbf{v}_2$ für die X-Achse, $\mathbf{X}\mathbf{v}_1$ für die Y-Achse).

In der Abbildung sind Skalierungen für die beiden Achsen eingezeichnet, um deutlich zu machen, wie die Koordinaten der Punkte verwendet worden sind. Diese Skalierungen haben jedoch keine für die Interpretation relevante Bedeutung, da die Konfiguration der Punkte beispielsweise verschoben werden kann, ohne dass sich an den Abständen etwas verändert. (Das wird genauer in Abschnitt 4.1 besprochen.) Bei der Interpretation der Abbildung ist außerdem zu beachten, dass die Maßeinheiten auf den beiden Achsen unterschiedlich sind. Die Bedeutung der sichtbaren euklidischen Abstände hängt also von der Richtung ab, in der man die Vergleiche anstellt.

5. Beurteilung der Projektionsgüte. Um das Ausmaß der Verzerrungen durch die Projektion einzuschätzen, kann man sich zunächst an den Singularwerten orientieren. Wie in § 3 ausgeführt wurde, liefert bei einer zweidimensionalen Projektion der Anteil der beiden größten Singularwerte $(\lambda_1^2 + \lambda_2^2)$ an der Summe aller quadrierten Singularwerte einen ersten Hinweis. In unserem Beispiel: $4.174/4.314 = 96.8\,\%$.

Eine andere Möglichkeit besteht darin, alle Abstände mit ihren Projektionen zu vergleichen und dann Aussagen über die Verteilung der Differenzen zu machen.

6. Projektionen großer Datenmengen. Bei den bisherigen Beispielen war die Anzahl der Objekte klein, so dass überschaubare Projektionen resultierten. Außerdem, für Interpretationsmöglichkeiten noch wichtiger, konnten die Objekte identifiziert werden; zum Beispiel als bestimmte Länder

Box 3.2-1 Aus der Singularwertzerlegung von $\mathbf{X}$ berechnete Koordinaten für die Zeilen von $\mathbf{X}$.

$$\mathbf{X}\mathbf{v}_1 = \begin{pmatrix} -0.5662 \\ -0.5469 \\ -0.5502 \\ -0.5221 \\ -0.5388 \\ -0.5392 \\ -0.4867 \\ -0.5110 \\ -0.4777 \\ -0.4386 \\ -0.3817 \\ -0.3690 \\ -0.3835 \\ -0.3744 \\ -0.3537 \\ -0.4612 \end{pmatrix} \qquad \mathbf{X}\mathbf{v}_2 = \begin{pmatrix} -0.2112 \\ -0.1643 \\ -0.1204 \\ -0.1682 \\ -0.1298 \\ -0.1523 \\ -0.1399 \\ -0.1188 \\ 0.1072 \\ 0.1384 \\ 0.2666 \\ 0.2689 \\ 0.2607 \\ 0.2664 \\ 0.2451 \\ 0.0978 \end{pmatrix}$$

oder Berufsgruppen. Wenn man es mit größeren Datenmengen zu tun hat, ist dies normalerweise nicht möglich, und ihre Projektionen liefern oft keine brauchbaren Informationen.

3.3 Korrespondenzanalyse

1. Der theoretische Ansatz. Das im vorangegangenen Abschnitt beschriebene Projektionsverfahren kann in vielen Varianten verwendet werden. Eine der Varianten ist unter dem Namen *Korrespondenzanalyse* bekanntgeworden.[9] Abgesehen von teilweise speziellen Bezeichnungen gibt es hauptsächlich drei Besonderheiten:

a) Die Korrespondenzanalyse bezieht sich in ihren Standardanwendungen auf eine zweidimensionale Kontingenztabelle[10] und setzt Häufigkeiten voraus.[11]

b) Es werden simultan Abstände zwischen den Zeilen und Abstände zwischen den Spalten einer Kontingenztabelle dargestellt.

c) Es werden keine euklidischen Abstände verwendet, sondern die Abstände werden aus einer statistischen Deutung der Kontingenztabelle gewonnen.

Um den theoretischen Ansatz zu erklären, beziehen wir uns auf eine zweidimensionale Kontingenztabelle $\mathbf{F} = (f_{ij})$ mit n Zeilen und m Spalten. Wir nehmen an, dass es sich bei den Elementen f_{ij} um relative Häufigkeiten handelt, so dass die Summe der Elemente $f_{..} = 1$ ist.[12] Aus der Tabelle wird eine neue Matrix $\mathbf{A} = (a_{ij})$ gebildet, deren Koeffizienten durch

$$a_{ij} := \frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.j}}} \tag{3.6}$$

definiert sind. Diese Koeffizienten werden als *standardisierte Residuen* (der Kontingenztabelle $\mathbf{F}$) bezeichnet. Sie zeigen in gewisser Weise, wie die Tabelle von einer statistischen Unabhängigkeit abweicht.[13]

---

[9] Man vgl. beispielsweise Greenacre (1993), Clausen (1998), Blasius (2001).

[10] Um höherdimensionale Tabellen zu verwenden, muss man mehrere Merkmalsräume formal in einen kombinierten eindimensionalen Merkmalsraum transformieren. Man spricht dann manchmal von *multipler* Korrespondenzanalyse. Vgl. das Beispiel in § 5.

[11] Dazu heißt es bei Greenacre (1993: 8): „The concept of a set of relative frequencies, or a profile, is fundamental to correspondence analysis.“ Und Blasius (2001: 81) kommentiert: Es werden stets „Profile interpretiert und nicht die absoluten Werte; d.h. die Interpretationen von Spalten- und Zeilenmerkmalen sind immer relative Aussagen.“ Diese Betrachtungsweise setzt offenbar Daten in Gestalt statistischer Verteilungen voraus.

[12] Wir verwenden die üblichen Notationen für Tabellen:

$$f_{i.} := \sum_j f_{ij}, \quad f_{.j} := \sum_i f_{ij}, \quad f_{..} := \sum_i \sum_j f_{ij}$$

Wenn f_{ij} zunächst absolute Häufigkeiten erfasst, erhält man durch $f_{ij}/f_{..}$ eine Tabelle mit relativen Häufigkeiten, deren Gesamtsumme = 1 ist.

[13] Werden durch f_{ij} relative Häufigkeiten erfasst, kann die hier gemeinte statistische Unabhängigkeit näherungsweise durch die Bedingung $f_{ij} \approx f_{i.}f_{.j}$ definiert werden.

Die Matrix $\mathbf{A}$ mit den standardisierten Residuen bildet nun den Ausgangspunkt für eine Singularwertzerlegung:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' \tag{3.7}$$

Um eine optimale zweidimensionale Projektion der Zeilen von $\mathbf{A}$ zu erreichen, kann man also (wie in Abschnitt 3.2 erklärt wurde) die Koordinaten $\mathbf{A}\mathbf{v}_1$ und $\mathbf{A}\mathbf{v}_2$ verwenden.[14] Indem man von $\mathbf{A}' = \mathbf{V}\mathbf{\Lambda}\mathbf{U}'$ ausgeht, findet man ganz analog, dass $\mathbf{u}_1'\mathbf{A}$ und $\mathbf{u}_2'\mathbf{A}$ Koordinaten für eine optimale Projektion der Spalten von $\mathbf{A}$ liefern.

Allerdings werden bei der Korrespondenzanalyse diese Koordinaten noch durch die Wurzeln der Zeilen- bzw. Spaltensummen dividiert, so dass schließlich für die graphische Darstellung folgende sogenannte *Hauptkoordinaten* verwendet werden. *Hauptkoordinaten für die Zeilen:*

$$\frac{(\mathbf{A}\mathbf{v}_1)_1}{\sqrt{f_{1\cdot}}}, \ldots, \frac{(\mathbf{A}\mathbf{v}_1)_n}{\sqrt{f_{n\cdot}}} \quad \text{und} \quad \frac{(\mathbf{A}\mathbf{v}_2)_1}{\sqrt{f_{1\cdot}}}, \ldots, \frac{(\mathbf{A}\mathbf{v}_2)_n}{\sqrt{f_{n\cdot}}}$$

und *Hauptkoordinaten für die Spalten:*

$$\frac{(\mathbf{u}_1'\mathbf{A})_1}{\sqrt{f_{\cdot 1}}}, \ldots, \frac{(\mathbf{u}_1'\mathbf{A})_n}{\sqrt{f_{\cdot m}}} \quad \text{und} \quad \frac{(\mathbf{u}_2'\mathbf{A})_1}{\sqrt{f_{\cdot 1}}}, \ldots, \frac{(\mathbf{u}_2'\mathbf{A})_n}{\sqrt{f_{\cdot m}}}$$

2. Illustration des Rechenverfahrens. Um das Rechenverfahren und die graphische Darstellung zu illustrieren, verwenden wir die Klausurdaten aus Abschnitt 2.2. Aus diesen Daten haben wir folgende Kontingenztabelle gebildet:[15]

$$\mathbf{F}^* = \begin{pmatrix} 39 & 4 & 0 & 1 & 2 \\ 40 & 1 & 4 & 0 & 1 \\ 25 & 0 & 2 & 2 & 17 \\ 21 & 6 & 9 & 6 & 4 \\ 27 & 6 & 8 & 0 & 5 \end{pmatrix} \tag{3.8}$$

Die Zeilen $i = 1, \ldots, 5$ entsprechen den Aufgaben, die Spalten $j = 1, \ldots, 5$ geben an, wie gut eine Aufgabe gelöst wurde: 1 (sehr gut) bis 5 (sehr schlecht). Zum Beispiel wurde die dritte Aufgabe in 25 Fällen sehr gut, in 17 Fällen sehr schlecht (oder gar nicht) gelöst. Aus $\mathbf{F}^*$ wird nun zunächst durch Division mit der Summe der Elemente ($n = 230$) eine Matrix $\mathbf{F}$ mit

[14] Wir nehmen an, dass die Singularwerte und entsprechend auch die Spalten von $\mathbf{V}$ der Größe nach geordnet sind.

[15] Identisch mit den Daten in Abschnitt 2.2, § 6.

relativen Häufigkeiten gebildet:

$$\mathbf{F} = \begin{pmatrix} 0.1696 & 0.0174 & 0.0000 & 0.0043 & 0.0087 \\ 0.1739 & 0.0043 & 0.0174 & 0.0000 & 0.0043 \\ 0.1087 & 0.0000 & 0.0087 & 0.0087 & 0.0739 \\ 0.0913 & 0.0261 & 0.0391 & 0.0261 & 0.0174 \\ 0.1174 & 0.0261 & 0.0348 & 0.0000 & 0.0217 \end{pmatrix} \quad \begin{pmatrix} 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \end{pmatrix}$$

$$\begin{pmatrix} 0.6609 & 0.0739 & 0.1000 & 0.0391 & 0.1261 \end{pmatrix}$$

(Hier haben wir außerdem die Zeilen- und Spaltensummen angegeben.) Dann wird aus $\mathbf{F}$ die Matrix $\mathbf{A}$ mit den standardisierten Residuen erzeugt:[16]

$$\mathbf{A} = \begin{pmatrix} 0.1028 & 0.0215 & -0.1414 & -0.0393 & -0.1040 \\ 0.1148 & -0.0858 & -0.0184 & -0.0885 & -0.1314 \\ -0.0646 & -0.1216 & -0.0799 & 0.0098 & 0.3066 \\ -0.1124 & 0.0930 & 0.1353 & 0.2064 & -0.0493 \\ -0.0407 & 0.0930 & 0.1045 & -0.0885 & -0.0219 \end{pmatrix}$$

Die Singularwertzerlegung $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$ liefert die Singularwerte

$$\lambda_1 = 0.3893, \;\; \lambda_2 = 0.3570, \;\; \lambda_3 = 0.1761, \;\; \lambda_4 = 0.1141, \;\; \lambda_5 = 0.0000$$

Zu den beiden größten Singularwerten gehören jeweils die ersten beiden Spalten von $\mathbf{U}$ und $\mathbf{V}$:

$$(\mathbf{u}_1, \mathbf{u}_2) = \begin{pmatrix} -0.3209 & 0.3695 \\ -0.3749 & 0.3755 \\ 0.8588 & 0.2446 \\ -0.0282 & -0.7889 \\ -0.1348 & -0.2006 \end{pmatrix} \quad (\mathbf{v}_1, \mathbf{v}_2) = \begin{pmatrix} -0.3157 & 0.4543 \\ -0.2418 & -0.4095 \\ -0.0880 & -0.5774 \\ 0.1555 & -0.5342 \\ 0.9000 & 0.0851 \end{pmatrix}$$

Schließlich kann man die Hauptkoordinaten für die Zeilenprofile $(\mathbf{x}_1^r, \mathbf{x}_2^r)$ und für die Spaltenprofile $(\mathbf{x}_1^c, \mathbf{x}_2^c)$ berechnen:

$$(\mathbf{x}_1^r, \mathbf{x}_2^r) = \begin{pmatrix} -0.2782 & 0.2961 \\ -0.3248 & 0.3010 \\ 0.7485 & 0.1921 \\ -0.0280 & -0.6297 \\ -0.1175 & -0.1596 \end{pmatrix} \quad (\mathbf{x}_1^c, \mathbf{x}_2^c) = \begin{pmatrix} -0.1502 & 0.2002 \\ -0.3491 & -0.5355 \\ -0.1110 & -0.6523 \\ 0.3000 & -0.9642 \\ 0.9871 & 0.0814 \end{pmatrix}$$

Mit diesen Koordinaten wurde die Abbildung 3.3-1 erzeugt, wobei $\mathrm{R}_1, \ldots, \mathrm{R}_5$ auf die Zeilen- und $\mathrm{C}_1, \ldots, \mathrm{C}_5$ auf die Spaltenprofile verweisen.

Auch bei dieser Abbildung haben wir Skalierungen für die beiden Achsen eingezeichnet, um deutlich zu machen, wie die Koordinaten der Punkte

[16] Für die praktischen Berechnungen haben wir die TDA-Prozedur `dma` (Option 6 für Korrespondenzanalyse) verwendet. Das Skript wurde `ka1.cf` genannt; die Tabelle (3.8) befindet sich im Datenfile `ka1.dat`.

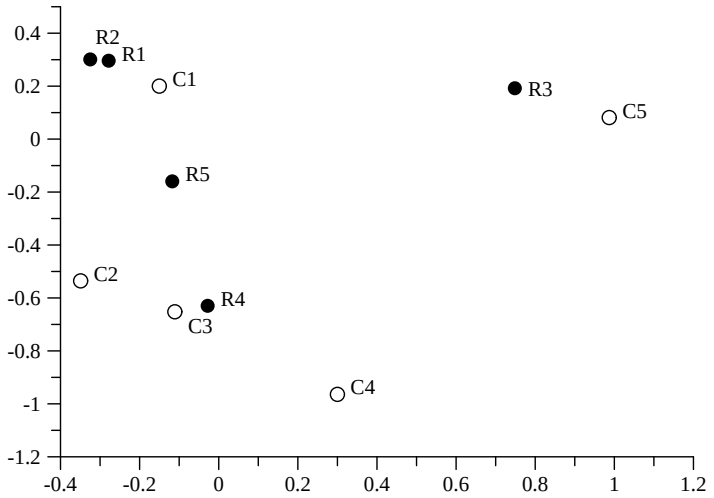


Abb. 3.3-1 Graphische Darstellung des Ergebnisses der Korrespondenzanalyse der Kontingenztabelle (3.8).

verwendet worden sind. Wie bereits bei der Abbildung 3.2-1 ist auch hier zu betonen, dass diese Skalierungen keine für die Interpretation relevante Bedeutung haben, da die Konfiguration der Punkte beispielsweise verschoben werden kann, ohne dass sich an den Abständen etwas verändert. Bestenfalls können also Abstände zwischen den Punkten interpretiert werden.

3. Unterschiedliche Abstandskonzeptionen. Allerdings sind die Abstände in der Abbildung 3.3-1 schwer zu interpretieren, denn sie beziehen sich in gewichteter Form auf die Zeilen bzw. Spalten der Matrix **A**, also auf die standardisierten Residuen der zunächst gegebenen Kontingenztabelle.

Es ist deshalb informativ, auch noch andere Abstandsdefinitionen in Betracht zu ziehen. Beispielsweise kann man die Notenverteilungen bei den fünf Aufgaben vergleichen. Man berechnet dann aus $\mathbf{F}^*$ eine neue Matrix, die in jeder Zeile die entsprechende Notenverteilung enthält. Dann kann man die Zeilen dieser Matrix (also die Zeilenprofile von $\mathbf{F}^*$) mit der in Abschnitt 3.2 besprochenen Projektionsmethode graphisch repräsentieren; Abbildung 3.3-2 zeigt das Ergebnis.[17] Die jetzt sichtbaren Abstände entsprechen den euklidischen Abständen zwischen den Verteilungen in den Zeilen von $\mathbf{F}^*$. Wenn man sie direkt berechnet, erhält man die Abstandsmatrix[18]

$$\mathbf{D} := \begin{pmatrix} 0.00 & 0.12 & 0.46 & 0.45 & 0.32 \\ 0.12 & 0.00 & 0.48 & 0.46 & 0.33 \\ 0.46 & 0.48 & 0.00 & 0.37 & 0.33 \\ 0.45 & 0.46 & 0.37 & 0.00 & 0.19 \\ 0.32 & 0.33 & 0.33 & 0.19 & 0.00 \end{pmatrix} \tag{3.9}$$

Die in Abbildung 3.3-2 sichtbaren Abstände entsprechen infolge der Projektionsverluste natürlich nur grob den Abständen in dieser Matrix.

4. Beurteilung der Projektionsgüte. Um das Ausmaß der Verzerrungen durch die Projektion einzuschätzen, kann man sich zunächst an den Singularwerten orientieren. Entsprechend der in Abschnitt 3.2 (§ 3) skizzierten Überlegung liefert bei einer zweidimensionalen Projektion der Anteil der beiden größten Singularwerte $(\lambda_1^2+\lambda_2^2)$ an der Summe aller quadrierten Singularwerte einen ersten Hinweis. In unserem Beispiel: $0.279/0.323 = 86\,\%$.

Eine andere Möglichkeit besteht darin, alle Abstände mit ihren Projektionen zu vergleichen und dann Aussagen über die Verteilung der Differenzen zu machen.

5. Korrespondenzanalyse der Berufsstrukturdaten. Für ein weiteres Beispiel verwenden wir die Berufsstrukturdaten aus Tabelle 2.3-4. Abbildung

[17] Die Daten für die Abbildung wurden mit dem Skript `ka1c.cf` erzeugt; für die Abbildung wurde das Skript `kaplot2.cf` verwendet.

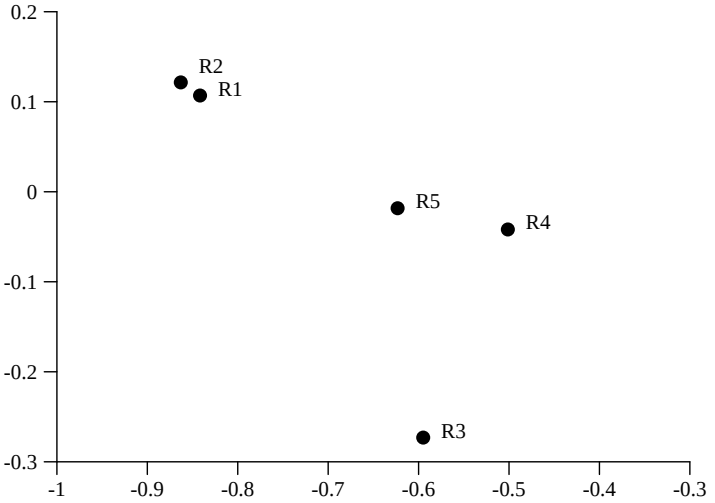


Abb. 3.3-2 Direkte lineare Projektionen der Zeilenprofile der Kontingenztabelle (3.8).

[18] Das Datenfile wird `kl5.dat` genannt und später in Abschnitt 4.2 für einen Vergleich verwendet.

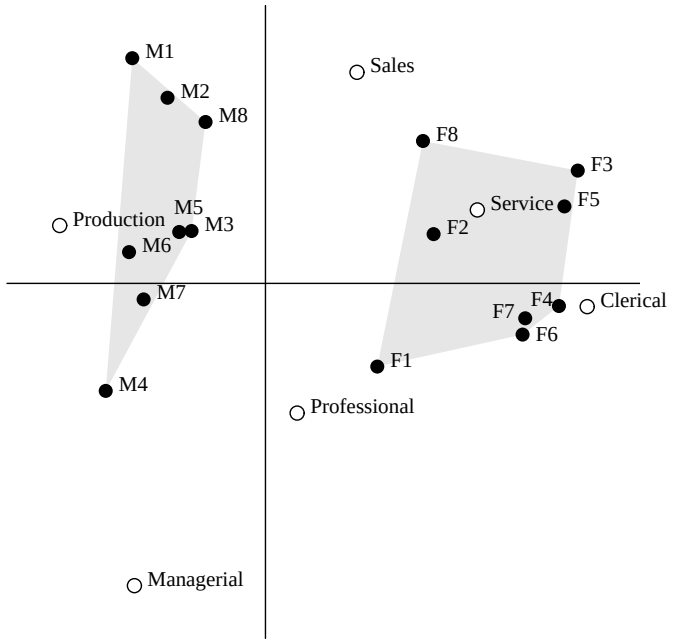


Abb. 3.3-3 Korrespondenzanalyse der Berufsstrukturdaten (Tab. 3.2-1).

3.3-3 zeigt das Ergebnis.[19]

[19] Die Berechnung wurde mit dem Skript `ka8.cf` durchgeführt; die Abbildung wurde mit dem Skript `kaplot4.cf` erzeugt.

Kapitel 4

Multidimensionale Skalierung

4.1 Konfigurationen
1. Konfigurationen und Abstände.
2. Ein zweidimensionales Beispiel.
3. Translationen und Rotationen.
4. Prokrustes-Rotation.

4.2 MDS mit Hauptkoordinaten
1. Die Problemstellung.
2. Überlegungen zum Einbettungsproblem.
3. Illustration der Berechnung.
4. Vergleich mit direkter Projektion.
5. Vergleich der Abstandsrepräsentation.
6. Berufsstrukturdaten.
7. Modifikation der Abstandsmatrix.

4.3 Metrische MDS-Verfahren
1. Die Problemstellung.
2. Alternative Problemformulierungen.
3. Rechentechnische Probleme.

4.4 Nichtmetrische MDS-Verfahren
1. Die Problemstellung.
2. Berücksichtigung von Bindungen.
3. Berechnungsmethoden.
4. Vollständige Stressreduktion.
5. Unvollständige Stressreduktion.
6. Das Shepard-Diagramm.
7. Unvollständige Abstandsmatrizen.

4.5 Zusätzliche Merkmalsachsen
1. Ergänzungen der MDS-Bilder.
2. Illustration mit Schulabschlüssen.
3. Konstruktion ergänzender Achsen.

Im vorangegangenen Kapitel wurde besprochen, wie man ausgehend von einer statistischen Datenmatrix oder Kontingenztabelle räumliche Bilder mit Projektionsverfahren erzeugen kann. In diesem Kapitel besprechen wir Methoden zur Erzeugung räumlicher Bilder, die stattdessen von einer Abstandsmatrix ausgehen. *Multidimensionale Skalierung* (MDS) wird als Sammelbegriff für diese Verfahren verwendet, die allgemein dem Zweck

dienen, Abstände, die durch eine Abstandsmatrix gegeben sind, durch räumliche Abstände zu repräsentieren.[1] Meistens wird ein Zahlenraum mit einer euklidischen Metrik verwendet; indem man sich dann auf zwei (oder maximal drei) Dimensionen beschränkt, können räumliche Bilder erzeugt werden. In diesem Kapitel besprechen wir drei Ansätze:

a) Einen Ansatz, manchmal „klassische“ MDS genannt, der auf einer Berechnung von Hauptkoordinaten beruht, die aus Eigenvektoren gewonnen werden.

b) Metrische MDS, bei der versucht wird, die Punkte im Zahlenraum so zu bestimmen, dass ihre Abstände möglichst weitgehend der vorausgesetzten Abstandsmatrix entsprechen.

c) Nichtmetrische MDS, bei der versucht wird, Abstände so zu konstruieren, dass ihre ordinalen Beziehungen möglichst weitgehend denjenigen in der vorausgesetzten Abstandsmatrix entsprechen.

Da in allen Ansätzen das Ziel darin besteht, eine *Konfiguration* von Punkten in einem Zahlenraum zu finden, beginnen wir mit einigen Bemerkungen zu diesem Begriff. In den dann folgenden Abschnitten orientieren wir uns an der Frage, wie zweidimensionale Bilder erzeugt werden können. Eindimensionale Skalierung bildet in gewisser Weise einen Sonderfall und wird erst im nächsten Kapitel behandelt.

4.1 Konfigurationen

1. Konfigurationen und Abstände. Unter einer *Konfiguration* verstehen wir eine Menge von n Punkten in einem Zahlenraum. Als Zahlenraum wird im Folgenden stets $\mathbf{R}^p$ mit einer beliebigen Dimension p verwendet. Eine Konfiguration besteht dann aus n Vektoren: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$. Wir folgen der Konvention, Vektoren stets als Spaltenvektoren aufzufassen, also kann jeder Vektor in der Form $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ geschrieben werden. Die aus den n Vektoren bestehende Konfiguration kann auch in einer (n, p)-Matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

dargestellt werden. Die Zeilen enthalten die Punkte der Konfiguration.

Für die Punkte einer Konfiguration können auf viele unterschiedliche Weisen Abstände definiert werden. Insbesondere können euklidische

[1] Es gibt eine umfangreiche Literatur, u.a. Kruskal und Wish (1978); Young und Hamer (1987); Borg und Lingoes (1987); Cox und Cox (1994). Speziell mit Marketing-Anwendungen beschäftigen sich Green, Carmone und Smith (1989).

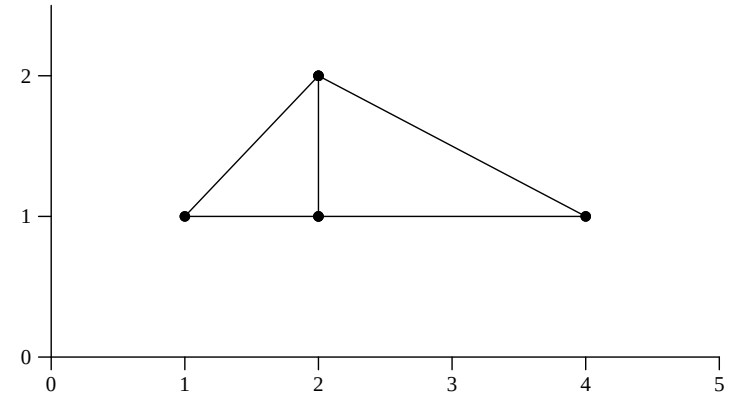


Abb. 4.1-1 Vier Städte in einem zweidimensionalen Koordinatensystem.

Abstände verwendet werden:

$$\| \mathbf{x}_i - \mathbf{x}_j \| = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2}$$

Für die quadrierten Abstände gilt:

$$\| \mathbf{x}_i - \mathbf{x}_j \|^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = \mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j$$

2. Ein zweidimensionales Beispiel. Abbildung 4.1-1 zeigt zur Illustration eine Konfiguration, die aus vier Punkten in einem zweidimensionalen Zahlenraum besteht. Sie kann durch eine Matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 2 & 2 \end{pmatrix} \tag{4.1}$$

erfasst werden. Verwendet man euklidische Abstände, erhält man die Abstandsmatrix

$$\mathbf{D} := \begin{pmatrix} 0 & 1 & 3 & \sqrt{2} \\ 1 & 0 & 2 & 1 \\ 3 & 2 & 0 & \sqrt{5} \\ \sqrt{2} & 1 & \sqrt{5} & 0 \end{pmatrix} \tag{4.2}$$

3. Translationen und Rotationen. Eine Konfiguration kann einer Reihe von Transformationen unterzogen werden, ohne dass sich an den euklidischen Abständen zwischen ihren Punkten etwas ändert.

a) Unter einer *Translation* versteht man, dass zu allen Punkten einer Konfiguration der gleiche Vektor addiert wird. Offenbar gilt

$$\| (\mathbf{x}_i + \mathbf{a}) - (\mathbf{x}_j + \mathbf{a}) \| = \| \mathbf{x}_i - \mathbf{x}_j \|$$

wobei $\mathbf{a}$ ein beliebiger Vektor ist.

b) Wenn man alle Punkte einer Konfiguration mit einer Zahl multipliziert, gilt für die euklidischen Abstände:

$$\| (\alpha\mathbf{x}_i) - (\alpha\mathbf{x}_j) \| = |\alpha| \, \| \mathbf{x}_i - \mathbf{x}_j \|$$

Offenbar verändern sich die Abstände nicht, wenn $\alpha = -1$ ist; man spricht dann von einer *Reflexion*.

c) Schließlich verändern sich die euklidischen Abstände auch dann nicht, wenn man eine Konfiguration rotiert. Mathematisch kann dass durch die Multiplikation der Punkte mit einer orthogonalen Matrix ausgedrückt werden: $\mathbf{x}_i^* := \mathbf{R}\mathbf{x}_i$.[2] Man erkennt:

$$\begin{aligned} \| \mathbf{x}_i^* - \mathbf{x}_j^* \|^2 = \| \mathbf{R} (\mathbf{x}_i - \mathbf{x}_j) \|^2 &= (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{R}' \mathbf{R} (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\mathbf{x}_i - \mathbf{x}_j)' (\mathbf{x}_i - \mathbf{x}_j) = \| \mathbf{x}_i - \mathbf{x}_j \|^2 \end{aligned}$$

Zur Illustration transformieren wir die vier Punkte der Konfiguration aus § 2 durch $\mathbf{x}_i^* = \mathbf{R}\mathbf{x}_i + \mathbf{a}$, wobei

$$\mathbf{R} = \begin{pmatrix} 0.866 & -0.500 \\ 0.500 & 0.866 \end{pmatrix} \quad \text{und} \quad \mathbf{a} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

ist. Durch Nachrechnen erkennt man, dass $\mathbf{R}$ orthogonal ist; diese Matrix entspricht einer Drehung um 30°.[3] Abbildung 4.1-2 zeigt gestrichelt die transformierte Konfiguration.

4. Prokrustes-Rotation. Als Ergebnis kann festgehalten werden, dass Konfigurationen, die ausgehend von Abständen konstruiert werden, nicht eindeutig bestimmt sind. Orientiert man sich an euklidischen Abständen, können die konstruierten Konfigurationen beliebigen Translationen, Rotationen und Reflexionen unterzogen werden.

Umgekehrt kann man sich fragen, wie man eine Konfiguration durch Translationen und Rotationen zu einer vorgegebenen Konfiguration

[2] Eine quadratische Matrix $\mathbf{R}$ wird *orthogonal* genannt, wenn gilt: $\mathbf{R}'\mathbf{R} = \mathbf{I}$, wenn also die transponierte gleich der inversen Matrix ist.

[3] $\mathbf{R}$ ist eine Beispiel für eine zweidimensionale Rotationsmatrix, die allgemein die Form

$$\mathbf{R} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

hat. In unserem Beispiel ist $\phi = 30°$.

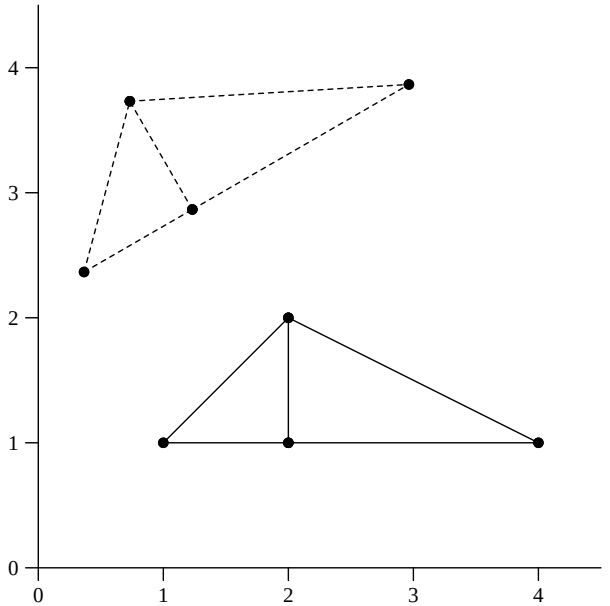


Abb. 4.1-2 Gedrehte und verschobene Konfiguration aus Abb. 4.1-1.

möglichst ähnlich machen kann. Diesem Zweck dient die sogenannte *Prokrustes-Rotation*. Ausgangspunkt sind zwei Konfigurationen, $\mathbf{X}$ und $\mathbf{Y}$, mit jeweils n Zeilen und m Spalten.

Es gibt verschiedene Varianten des Verfahrens.[4] In einer ersten Variante sind eine orthogonale (m, m)-Matrix $\mathbf{R}$ und eine (n, m)-Translationsmatrix $\mathbf{T}$ gesucht, so dass sich $\mathbf{X}$ und

$$\mathbf{Y}_{(\mathbf{R},\mathbf{T})} := \mathbf{Y}\mathbf{R} + \mathbf{T}$$

möglichst ähnlich sind.[5] Als Kriterium wird

$$\| \mathbf{X} - \mathbf{Y}_{(\mathbf{R},\mathbf{T})} \| \longrightarrow \min$$

verwendet.[6] In einer zweiten Variante wird außerdem eine variable Skalie-

[4] Eine ausführliche Diskussion gibt Commandeur (1991).

[5] Unter einer *Translationsmatrix* verstehen wir eine Matrix, deren Zeilen identisch (= dem oben so genannten Translationsvektor) sind.

[6] Für eine beliebige Matrix $\mathbf{A} = (a_{ij})$ ist der Ausdruck $\| \mathbf{A} \|$ durch

$$\| \mathbf{A} \| = (\Sigma_i \Sigma_j a_{ij}^2)^{1/2}$$

definiert. Bei Vektoren entspricht er ihrer euklidischen Länge.

rung zugelassen, d.h. es werden eine orthogonale Matrix $\mathbf{R}$, eine Translationsmatrix $\mathbf{T}$ und ein Skalierungsfaktor α gesucht, so dass

$$\| \mathbf{X} - \mathbf{Y}_{(\alpha,\mathbf{R},\mathbf{T})} \| \longrightarrow \min$$

wobei jetzt $\mathbf{Y}_{(\alpha,\mathbf{R},\mathbf{T})} := \alpha\mathbf{YR} + \mathbf{T}$ ist.[7]

Zur Illustration sei angenommen, dass man die in (4.1) definierte Matrix $\mathbf{X}$ kennt und die Koordinaten der in Abbildung 4.1-2 gestrichelt gezeichneten Konfiguration:

$$\mathbf{Y} = \begin{pmatrix} 0.3660 & 2.3660 \\ 1.2320 & 2.8660 \\ 2.9640 & 3.8660 \\ 0.7320 & 3.7320 \end{pmatrix}$$

Gesucht sind jetzt α, $\mathbf{R}$ und $\mathbf{T}$, so dass $\| \mathbf{Y} - (\alpha\mathbf{XR} + \mathbf{T}) \|$ minimal wird. Man findet:[8]

$$\alpha = 1.0 \qquad \mathbf{R} = \begin{pmatrix} 0.8660 & 0.5000 \\ -0.5000 & 0.8660 \end{pmatrix} \qquad \mathbf{T} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

In diesem Beispiel lässt sich eine perfekte Übereinstimmung erzielen, d.h. der Ausdruck $\| \mathbf{Y} - (\alpha\mathbf{XR} + \mathbf{T}) \|$ wird Null. Im Allgemeinen ist nur eine Annäherung möglich; Beispiele folgen in späteren Abschnitten.

4.2 MDS mit Hauptkoordinaten

In diesem Abschnitt besprechen wir eine Variante der multidimensionalen Skalierung (MDS), die auf einer Berechnung von Hauptkoordinaten beruht. Der Ansatz, manchmal „klassische“ MDS genannt, wurde zuerst von Young und Householder (1938) vorgechlagen und dann von Torgerson (1952, 1958) für (zunächst hauptsächlich psychometrische) Anwendungen ausgearbeitet.[9]

1. Die Problemstellung. Wir nehmen an, dass eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ für n Objekte $\omega_1, \ldots, \omega_n$ gegeben ist. Die leitende Idee besteht darin, die Objekte durch Punkte (Vektoren) in einem Zahlenraum $\mathbf{R}^p$ zu repräsentieren. Gesucht sind also n Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$. Jeder dieser Vektoren ist ein Spaltenvektor mit p Komponenten: $\mathbf{x}_i = (x_{i1,}, \ldots, x_{ip})'$.

[7] Die mathematischen Hintergründe des Verfahrens sind kompliziert und sollen hier nicht besprochen werden; man vgl. beispielsweise Mardia, Kent und Bibby (1979: 416ff.).

[8] Für die Berechnung wurde das Skript `pr1.cf` verwendet.

[9] Darstellungen findet man u.a. bei Mardia, Kent und Bibby (1979: 397); Cox und Cox (1994: 22ff.); Falk, Becker und Marohn (1995: 266ff.).

Um von ihren Abständen sprechen zu können, wird eine euklidische Metrik verwendet: $\| \mathbf{x}_i - \mathbf{x}_j \| = (\Sigma_k (x_{ik} - x_{jk})^2)^{1/2}$. Jetzt können zwei Varianten der Problemstellung formuliert werden:

a) Kann man eine minimale Dimension p und Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ finden, so dass gilt:

$$\| \mathbf{x}_i - \mathbf{x}_j \| = d_{ij} \quad \text{(für alle } 1 \le i < j \le n) \tag{4.3}$$

Wir nennen dies das *Einbettungsproblem* (für $\mathbf{D}$).

b) Wie kann man Vektoren $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^* \in \mathbf{R}^2$ finden, so dass deren Abstände $\| \mathbf{x}_i^* - \mathbf{x}_j^* \|$ näherungsweise den vorgegebenen Abständen d_{ij} entsprechen, und wie kann man Aussagen über die Güte der Approximation machen? Wir nennen dies das *Darstellungsproblem* (für $\mathbf{D}$).

2. Überlegungen zum Einbettungsproblem. Die Überlegung beginnt damit, eine neue (n,n)-Matrix $\mathbf{A} = (a_{ij})$ zu betrachten, deren Elemente folgendermaßen definiert sind:

$$a_{ij} := -\frac{1}{2} (d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2) \tag{4.4}$$

wobei zur Abkürzung verwendet wird:

$$d_{i\cdot}^2 := \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2, \quad d_{\cdot j}^2 := \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2, \quad d_{\cdot\cdot}^2 := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2$$

Die Matrix $\mathbf{A}$, oft eine *doppelt zentrierte Matrix* genannt, ist also genau wie $\mathbf{D}$ eine symmetrische Matrix. Man kann sich durch Nachrechnen davon überzeugen, dass folgende Beziehung gilt:

$$d_{ij}^2 = a_{ii} - a_{ij} - a_{ji} + a_{jj}$$

Ebenso kann man nachrechnen, dass die Zeilen- und Spaltensummen der Matrix $\mathbf{A}$ stets Null sind. Ihr Rang ist also höchstens $n-1$. Das Konstruktionsverfahren besteht nun darin, die Eigenwerte und Eigenvektoren der Matrix $\mathbf{A}$ zu untersuchen, also Lösungen der Gleichung $\mathbf{Az} = \lambda\,\mathbf{z}$. Da $\mathbf{A}$ eine symmetrische Matrix ist, sind ihre Eigenwerte reelle Zahlen; und da der Rang von $\mathbf{A}$ höchstens $n-1$ ist, gibt es höchstens $n-1$ von Null verschiedene Eigenwerte. Wir setzen außerdem voraus, dass $\mathbf{A}$ positiv-semidefinit ist, so dass alle Eigenwerte größer oder gleich Null sind.[10] Es kann dann angenommen werden, dass es p von Null verschiedene Eigenwerte gibt und dass man sie der Größe nach ordnen kann:

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p > 0 \quad \text{und} \quad \lambda_{p+1} = \cdots = \lambda_n = 0$$

[10] Wie vorgegangen werden kann, wenn das nicht der Fall ist, wird später besprochen.

Die zugehörigen Eigenvektoren bezeichnen wir mit $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Es sind Spaltenvektoren mit jeweils n Komponenten. Die Eigenvektoren können zu einer (n, n)-Matrix

$$\mathbf{Z} := (\mathbf{z}_1, \ldots, \mathbf{z}_p, \mathbf{z}_{p+1}, \ldots, \mathbf{z}_n)$$

zusammengefasst und so normiert werden, dass gilt: $\mathbf{Z}\mathbf{Z}' = \mathbf{I}_n$. Fasst man weiterhin die Eigenwerte zu einer Diagonalmatrix $\mathbf{\Lambda} := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ zusammen, können die Eigenwertgleichungen zusammengefasst werden: $\mathbf{A}\mathbf{Z} = \mathbf{Z}\mathbf{\Lambda}$. Es folgt wegen $\mathbf{Z}\mathbf{Z}' = \mathbf{I}_n$:

$$\mathbf{A} = \mathbf{A}\mathbf{Z}\mathbf{Z}' = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}' \tag{4.5}$$

Jetzt definieren wir eine (n, p)-Matrix

$$\mathbf{X} := \mathbf{Z}\mathbf{\Lambda}^{1/2} = (\sqrt{\lambda_1}\,\mathbf{z}_1, \ldots, \sqrt{\lambda_p}\,\mathbf{z}_p) \tag{4.6}$$

Sie besteht aus den ersten p Spalten von $\mathbf{Z}$, wobei jedoch jeder dieser Spaltenvektoren mit der Wurzel des entsprechenden Eigenwerts multipliziert wird. Da die letzten $n - p$ Diagonalelemente von $\mathbf{\Lambda}$ Null sind, folgt aus dieser Definition $\mathbf{Z}\mathbf{\Lambda}\mathbf{Z}' = \mathbf{X}\mathbf{X}'$, so dass man schließlich die Darstellung

$$\mathbf{A} = \mathbf{X}\mathbf{X}' \tag{4.7}$$

erhält. Die Zeilen von $\mathbf{X}$, also $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, liefern nun die Koordinaten in $\mathbf{R}^p$, um die Abstände zwischen den Objekten räumlich zu repräsentieren. Man sieht das, wenn man für je zwei dieser Vektoren, $\mathbf{x}_i$ und $\mathbf{x}_j$, ihren euklidischen Abstand berechnet. Dann findet man nämlich (zunächst für die quadrierten Abstände):

$$\| \mathbf{x}_i - \mathbf{x}_j \|^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) =$$
$$\mathbf{x}_i'\mathbf{x}_i - \mathbf{x}_i'\mathbf{x}_j - \mathbf{x}_j'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j = a_{ii} - a_{ij} - a_{ji} + a_{jj} = d_{ij}^2$$

Die euklidischen Abstände zwischen den Zeilen von $\mathbf{X}$ sind also mit den vorgegebenen Abständen identisch.

3. Illustration der Berechnung. Ein einfaches Beispiel soll die Rechenschritte illustrieren. Wir nehmen an, dass vier Objekte (Städte) in einem zweidimensionalen Koordinatensystem gegeben sind (Abbildung 4.2-1) und dass ihre Abstände durch eine euklidische Metrik erfasst werden:

$$\mathbf{D} := \begin{pmatrix} 0 & 1 & 3 & \sqrt{2} \\ 1 & 0 & 2 & 1 \\ 3 & 2 & 0 & \sqrt{5} \\ \sqrt{2} & 1 & \sqrt{5} & 0 \end{pmatrix}$$

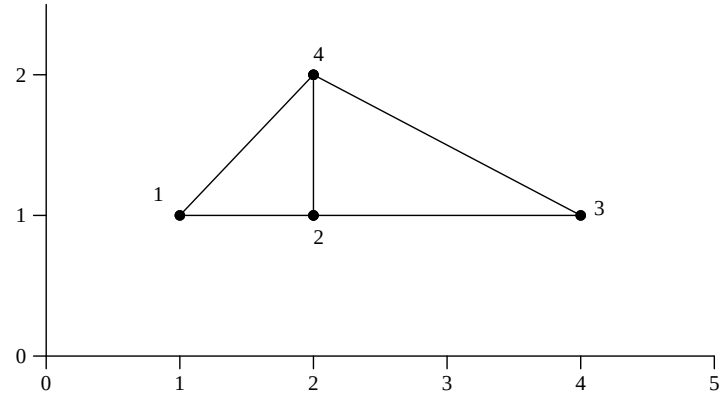


Abb. 4.2-1 Vier Städte in einem zweidimensionalen Koordinatensystem.

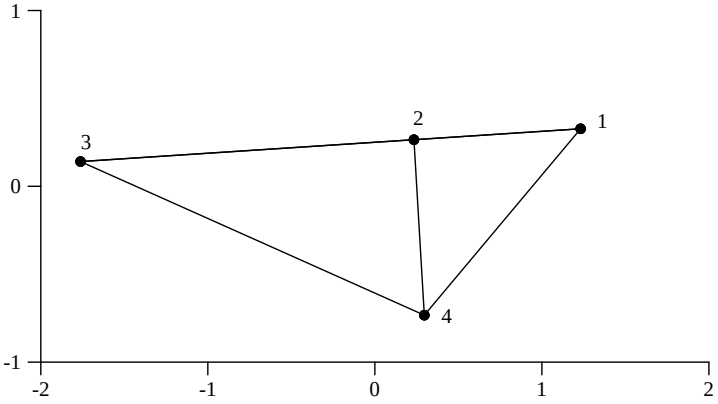


Abb. 4.2-2 Darstellung der vier Punkte mit den in (4.8) angegebenene Koordinaten.

Daraus gewinnt man entsprechend (4.4) die Matrix[11]

$$\mathbf{A} = \begin{pmatrix} 1.6250 & 0.3750 & -2.1250 & 0.1250 \\ 0.3750 & 0.1250 & -0.3750 & -0.1250 \\ -2.1250 & -0.3750 & 3.1250 & -0.6250 \\ 0.1250 & -0.1250 & -0.6250 & 0.6250 \end{pmatrix}$$

[11] Die folgenden Berechnungen wurden mit der TDA-Prozedur `mdsc` durchgeführt, das Skript ist `dma3.cf`.

Diese Matrix liefert die Eigenwerte

$$\lambda_1 = 4.7656, \;\; \lambda_2 = 0.7343, \;\; \lambda_3 = 0.0000, \;\; \lambda_4 = 0.0000$$

und die zugehörigen Eigenvektoren

$$\mathbf{Z} = \begin{pmatrix} 0.5644 & 0.3818 & -0.5000 & 0.5345 \\ 0.1072 & 0.3093 & -0.5000 & -0.8018 \\ -0.8072 & 0.1643 & -0.5000 & 0.2673 \\ 0.1357 & -0.8553 & -0.5000 & 0.0000 \end{pmatrix}$$

Da nur zwei Eigenwerte ungleich 0 sind, können die Entfernungen zwischen den Städten in einem zweidimensionalen Raum dargestellt werden. Die entsprechend (4.6) berechneten Koordinaten sind

$$\begin{aligned} \mathbf{x}_1 &= (1.2320, 0.2340, -1.7622, 0.2961)' \\ \mathbf{x}_2 &= (0.3272, 0.2650, 0.1408, -0.7330)' \end{aligned} \tag{4.8}$$

Berechnet man die Abstände zwischen diesen vier Punkten, findet man die durch $\mathbf{D}$ vorgegebenen Distanzen. Abbildung 4.2-2 gibt eine graphische Darstellung der vier Punkte. Ein Vergleich mit Abbildung 4.2-1 zeigt, dass eine Drehung und Translation stattgefunden hat; zwischen den Punkten in Abbildung 4.2-2 gibt es jedoch die gleichen Abstände wie zwischen den Städten in Abbildung 4.2-1.

4. Vergleich mit direkter Projektion. Es mag von Interesse sein, die MDS mit Hauptkoordinaten mit der Methode der direkten Projektion aus Abschnitt 3.2 zu vergleichen. Dafür verwenden wir die Klausurdaten, mit denen in Abschnitt 3.3 (§ 3) bereits eine direkte Projektion durchgeführt wurde. Jetzt dient als Ausgangspunkt die dort erzeugte Abstandsmatrix $\mathbf{D}$.[12] Eine Hauptkoordinaten-MDS mit dieser Abstandsmatrix liefert die beiden größten Eigenwerte $\lambda_1 = 0.1814$ und $\lambda_2 = 0.0761$, die zusammen 94 % der Summe aller Eigenwerte bilden, und die Koordinaten:[13]

$$\mathbf{X} = \begin{pmatrix} 0.2178 & -0.0213 \\ 0.2346 & -0.0068 \\ -0.1966 & -0.2105 \\ -0.1896 & 0.1569 \\ -0.0660 & 0.0817 \end{pmatrix} \tag{4.9}$$

Abbildung 4.2-3 zeigt zunächst diese Konfiguration.[14] Außerdem wird ein Vergleich mit der durch direkte Projektion gewonnenen Konfiguration vorgenommen. Die folgende Matrix $\mathbf{Y}$ zeigt die für Abbildung 3.3-2 verwendete Konfiguration:

$$\mathbf{Y} = \begin{pmatrix} -0.8417 & 0.1069 \\ -0.8630 & 0.1215 \\ -0.5949 & -0.2731 \\ -0.5013 & -0.0419 \\ -0.6232 & -0.0183 \end{pmatrix} \qquad \mathbf{Y}^* = \begin{pmatrix} 0.2232 & 0.0090 \\ 0.2516 & 0.0078 \\ -0.2462 & -0.1630 \\ -0.1758 & 0.1030 \\ -0.0526 & 0.0432 \end{pmatrix} \tag{4.10}$$

Die Matrix $\mathbf{Y}^*$ ist die mit dem Verfahren der Prokrustes-Rotation gewonnene optimale Anpassung von $\mathbf{Y}$ an $\mathbf{X}$;[15] ihre Zeilen sind in Abbildung 4.2-3 als Kreuze eingetragen.

5. Vergleich der Abstandsrepräsentation. In diesem Beispiel kann der Vergleich noch etwas fortgesetzt werden, weil beide Verfahren – die MDS mit Hauptkoordinaten und die direkte Projektion – eine Repräsentation der euklidischen Abstände zwischen den Zeilen der Matrix $\mathbf{F}^*$ (aus Abschnitt 3.3) anstreben. Hierfür wird aus der jeweils ermittelten Konfiguration eine Matrix der euklidischen Abstände berechnet und dann mit der Abstandsmatrix $\mathbf{D}$ aus Abschnitt 3.3 verglichen.[16]

[12] Vgl. (3.9) in Abschnitt 3.3; das Datenfile für diese Abstandsmatrix ist `kl5.dat`.

[13] Die Berechnungen wurden mit dem Skript `mdsc1.cf` durchgeführt.

[14] Die Abbildung wurde mit dem Skript `mdsplot4.cf` erzeugt.

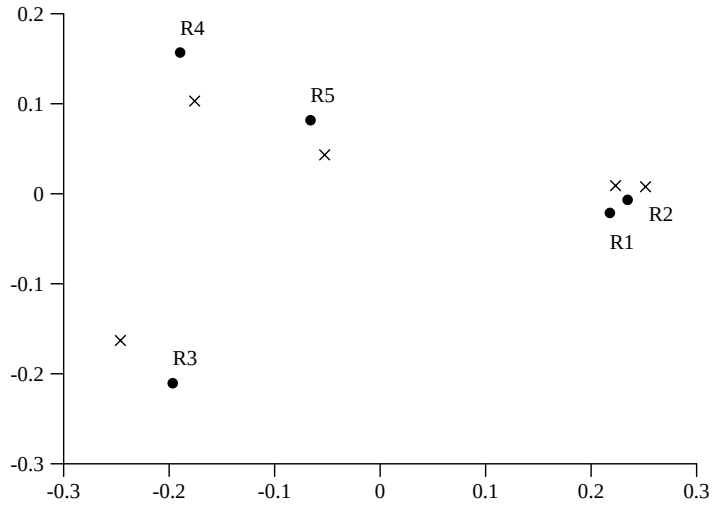


Abb. 4.2-3 Durch eine Hauptkoordinaten-MDS erzeugte Konfiguration zur Repräsentation der Abstandsmatrix $\mathbf{D}$ in (3.9). Außerdem in Form von Kreuzen eine mittels Prokrustes-Rotation angepasste Variante der Konfiguration aus Abbildung 3.3-2.

[15] Es wurde das Skript `pr2.cf` verwendet. Es ist bemerkenswert, dass das Verfahren in diesem Fall auch eine Skalierung impliziert: $\alpha = 1.1$.

[16] Für die Berechnungen wurden die Skripte `mdsc2.cf` und `mdsc3.cf` verwendet.

a) Die aus $\mathbf{X}$ in (4.9) gebildete Abstandsmatrix wird $\mathbf{D}_x$ genannt. Dann findet man: $\|\mathbf{D} - \mathbf{D}_x\| = 0.1575$.

b) Die aus $\mathbf{Y}$ in (4.10) gebildete Abstandsmatrix wird $\mathbf{D}_y$ genannt. Dann findet man: $\|\mathbf{D} - \mathbf{D}_y\| = 0.3186$.

c) Die aus $\mathbf{Y}^*$ in (4.10) gebildete Abstandsmatrix wird $\mathbf{D}_{y^*}$ genannt. Dann findet man: $\|\mathbf{D} - \mathbf{D}_{y^*}\| = 0.2474$.

In diesem Beispiel liefert also die MDS mit Hauptkoordinaten die vergleichweise beste Repräsentation der Abstände.

6. Berufsstrukturdaten. Die MDS mit Hauptkoordinaten setzt normalerweise voraus, dass die aus der Abstandsmatrix gebildete doppelt zentrierte Matrix positiv-semidefinit ist; denn nur dann sind alle Eigenwerte nichtnegativ und lässt sich die Abstandsmatrix in einen euklidischen Zahlenraum einbetten. Diese Bedingung ist jedoch oft nicht erfüllt. Um das Problem zu illustrieren, verwenden wir die Berufsstrukturdaten, mit denen in Abschnitt 3.3 (§ 5) eine Korrespondenzanalyse durchgeführt wurde. Jetzt gehen wir anders vor und erzeugen zuerst eine Abstandsmatrix, die dann den Ausgangspunkt für die MDS bildet.

Anders als bei der Korrespondenzanalyse ist man jetzt frei, die Abstände, die man sichtbar machen möchte, aufgrund expliziter Überlegungen zu definieren. Für die gegenwärtige Illustration verwenden wir City-Block-Abstände zwischen den Zeilenprofilen der Daten in Tabelle 3.2-1; dies ist äquivalent zu einer Verwendung von Dissimilaritätsindizes.

Führt man mit der resultierenden $(16, 16)$-Abstandsmatrix eine Hauptkoordinaten-MDS durch,[17] findet man, dass 6 der 16 Eigenwerte negativ sind. Die Abstandsmatrix ist also nicht in einen euklidischen Zahlenraum einbettbar. Was kann man in einer solchen Situation tun? Es gibt zwei Möglichkeiten.

Eine Möglichkeit besteht darin, die Abstandsmatrix durch Addition einer Konstanten so zu modifizieren, dass die aus ihr gebildete doppelt zentrierte Matrix positiv-semidefinit wird. Wenn, was fast immer der Fall ist, die beiden größten Eigenwerte positiv sind, kann man aber auch einfach die zu ihnen gehörenden Eigenvektoren verwenden, um daraus Koordinaten für eine Konfiguration zu gewinnen. Wenn man zunächst dieser Möglichkeit folgt, findet man die in Abbildung 4.2-4 gezeigte Konfiguration.[18]

7. Modifikation der Abstandsmatrix. Die andere der beiden erwähnten Möglichkeiten beruht darauf, dass man aus der zunächst gegebenen Abstandsmatrix $\mathbf{D} = (d_{ij})$ durch Hinzufügen einer Konstanten eine neue

[17] Verwendet wurde `mdsc4.cf`.

[18] Die Abbildung wurde mit `mdsplot5.cf` erzeugt.

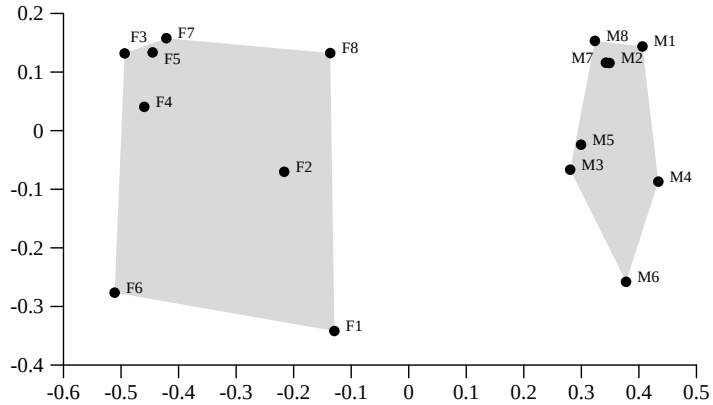


Abb. 4.2-4 Durch eine Hauptkoordinaten-MDS erzeugte Konfiguration zur Repräsentation der City-Block-Abstände zwischen den Zeilenprofilen der Berufsstrukturdaten in Tabelle 3.2-1.

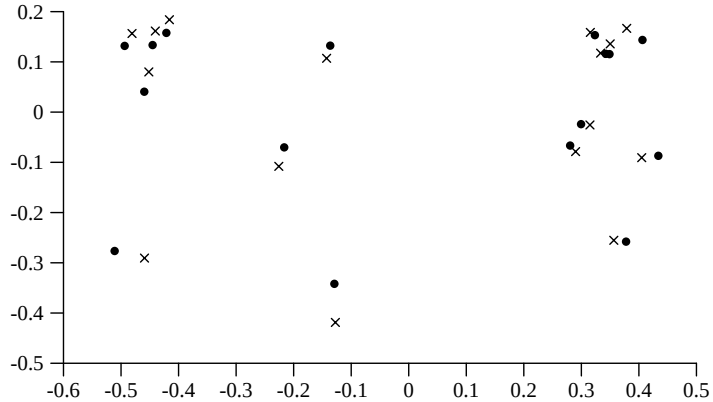


Abb. 4.2-5 Vergleich der Konfiguration aus Abbildung 4.2-4 mit einer durch Kreuze dargestellten Konfiguration, für deren Bildung die Abstandsmatrix durch eine additive Konstante ($\alpha = 0.7336$) modifiziert wurde.

Abstandsmatrix

$$\mathbf{D}^\alpha = (d_{ij}^\alpha) \qquad \text{mit} \quad d_{ij}^\alpha := \begin{cases} d_{ij} + \alpha & \text{wenn } i \neq j \\ 0 & \text{wenn } i = j \end{cases} \tag{4.11}$$

bildet.[19] Sei nämlich λ_m der kleinste (negative) Eigenwert. Es gilt dann:

[19] Einen allgemeineren Ansatz besprechen Bénasséni, Dosse und Joly (2007).

Wenn $\alpha \geq \sqrt{4\lambda_m^2 - 2\lambda_m} - 2\lambda_m$ (und außerdem nichtnegativ) ist, ist die aus $\mathbf{D}^\alpha$ gebildete doppelt zentrierte Matrix positiv-semidefinit.[20]

Wendet man diese Methode für das Beispiel mit den Berufsstrukturdaten an, findet man $\alpha = 0.7336$.[21] Modifiziert man die Abstandsmatrix (mit den Abständen zwischen den Zeilenprofilen der Berufsstrukturdaten) entsprechend (4.11) mit dieser Konstanten, sind alle Eigenwerte der resultierenden doppelt zentrierten Matrix größer oder gleich Null. Abbildung 4.2-5 vergleicht die beiden Konfigurationen.[22]

Welche der beiden Möglichkeiten verwendet werden sollte, hängt vom Verwendungszweck ab. Geht es darum, die gegebene Abstandsmatrix möglichst gut darzustellen, ist das Hinzufügen einer additiven Konstanten meistens nicht zweckmäßig; denn dadurch wird die Differenz zwischen der gegebenen und der durch die Konfiguration erzeugten Abstandsmatrix normalerweise größer. Dies ist auch in unserem Beispiel der Fall.[23]

4.3 Metrische MDS-Verfahren

1. Die Problemstellung. Ausgangspunkt ist wieder eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ für n Objekte. Außerdem wird jetzt der Zahlenraum für die räumliche Darstellung vorgegeben (so dass sich kein Einbettungs- und Projektionsproblem stellt). Grundsätzlich kann ein Zahlenraum $\mathbf{R}^p$ mit einer beliebigen Dimension p verwendet werden. Wenn man an graphischen Darstellungen interessiert ist, verwendet man meistens den zweidimensionalen Zahlenraum $\mathbf{R}^2$; das wird im Folgenden angenommen.

Gesucht ist nun eine Konfiguration $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$, so dass die durch ihre Punkte gegebenen Abstände $\| \mathbf{x}_i - \mathbf{x}_j \|$ möglichst den vorgegebenen Abständen d_{ij} entsprechen. Dafür kann folgendes Kriterium verwendet werden:

$$s(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \sum\nolimits_{j<i} w_{ij} \left(d_{ij} - \| \mathbf{x}_i - \mathbf{x}_j \|\right)^2 \tag{4.12}$$

Dabei sind w_{ij} nichtnegative Gewichte, die dem Zweck dienen, eine etwas allgemeinere Problemformulierung zu erreichen. Man kann beispielsweise $w_{ij} = 0$ setzen, wenn ein Abstandswert d_{ij} nicht bekannt ist. Bei vollständig bekannten Abstandsmatrizen wird man meistens für alle Gewichte $w_{ij} = 1$ annehmen, so dass man sie ignorieren kann. Die Funktion

[20] Einen ausführlichen Nachweis findet man bespielsweise bei Falk, Becker und Marohn (1995: 272).

[21] Verwendet wurde das Skript `mdsc4a.cf`.

[22] Für die Prokrustes-Rotation wurde `pr4.cf` verwendet; die Abbildung wurde mit `mdsplot5a.cf` erzeugt.

[23] Analog zur Vorgehensweise in § 5 können zum Nachrechnen die Skripte `mdsc5.cf` und `mds6.cf` verwendet werden.

s wird *Stressfunktion* genannt. Die Aufgabe besteht darin, eine Konfiguration zu finden, die den Wert dieser Funktion minimal macht.

2. Alternative Problemformulierungen. Für die metrische MDS wird meistens die Stressfunktion (4.12) verwendet. Varianten können bei zwei Aspekten ansetzen. Einerseits kann man anstelle der euklidischen Abstände $\| \mathbf{x}_i - \mathbf{x}_j \|$ andere Abstandsdefinitionen verwenden. Vorgeschlagen wurde insbesondere die City-Block-Metrik; die Stressfunktion nimmt dann folgende Form an:

$$s^c(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \sum\nolimits_{j<i} w_{ij} \left(d_{ij} - d^c(\mathbf{x}_i, \mathbf{x}_j)\right)^2 \tag{4.13}$$

wobei $d^c(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$ ist.[24] Andererseits kann man die Art des Vergleichs zwischen den vorgegebenen und den zu konstruierenden Abständen verändern. Hier sind zwei Varianten:

a) Man verwendet das Kriterium

$$a(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \sum\nolimits_{j<i} w_{ij} \left| d_{ij} - \| \mathbf{x}_i - \mathbf{x}_j \| \right| \tag{4.14}$$

also absolute anstelle der quadrierten Abweichungen.[25]

b) Man verwendet das Kriterium

$$m(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \max\{ \left| d_{ij} - \| \mathbf{x}_i - \mathbf{x}_j \| \right| \Big| 1 \leq i < j \leq n\} \tag{4.15}$$

d.h. man versucht, die maximale Abweichung zwischen den vorgegebenen und den zu konstruierenden Abständen zu minimieren.

3. Rechentechnische Probleme. In allen Varianten treten rechentechnische Probleme auf. Dies gilt auch für die Standardvariante mit der Stressfunktion (4.12), an der wir uns im Folgenden orientieren. Das Hauptproblem besteht darin, dass die als Kriterium verwendete Funktion meistens zahlreiche (und unter Umständen sehr viele) lokale Minima aufweist und alle üblichen Minimierungsalgorithmen nur solche lokalen Minima finden können. Infolgedessen findet man nicht unbedingt auch ein globales Minimum der Zielfunktion.

Bei der praktischen Durchführung einer metrischen MDS ist es deshalb sinnvoll, den Minimierungsalgorithmus ausgehend von unterschiedlichen Startkonfigurationen sehr oft zu wiederholen. So kann man sich einen gewissen Überblick über lokale Minima verschaffen und schließlich das beste der bisher gefundenen Ergebnisse auswählen.

[24] Vgl. Pliner (1986); Hubert, Arabie und Hesson-Mcinnis (1992); Groenen, Heiser und Meulman (1998).

[25] Vgl. Heiser (1988).

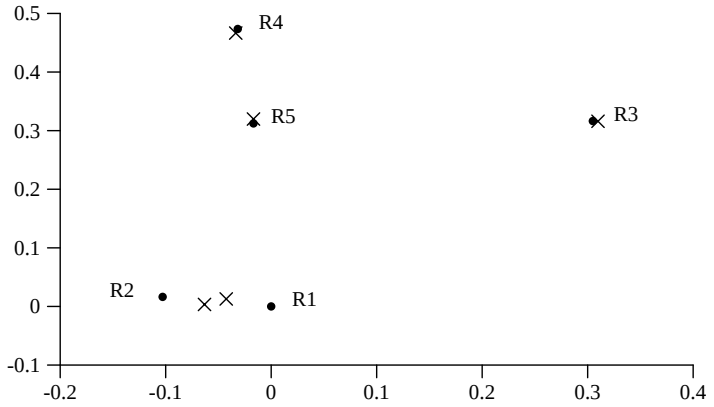


Abb. 4.3-1 Metrische MDS mit der Abstandsmatrix (3.9) für die Klausurdaten. Außerdem in Form von Kreuzen die in Abschnitt 4.2 (§ 4) mit der Hauptkoordinaten-MDS gefundene Konfiguration.

Zur Illustration des Verfahrens verwenden wir wieder die Klausurdaten, und zwar gehen wir wie in Abschnitt 3.3 von der Abstandsmatrix $\mathbf{D}$ in (3.9) aus. Bei 100 Wiederholungen eines Minimierungsalgorithmus für das Kriterium (4.12), die mit der TDA-Prozedur `mdsm` durchgeführt wurden,[26] wurde in 63 Fällen das relativ beste lokale Minimum gefunden. Abbildung 4.3-1 zeigt die diesem Minimum entsprechende Konfiguration.[27] Die Abbildung zeigt außerdem die in Abschnitt 4.2 (§ 4) mit der Hauptkoordinaten-MDS gefundene Konfiguration.[28] Da die metrische MDS den Abstand zur vorgegebenen Abstandsmatrix direkt minimiert, ist die Anpassung natürlich besser; in diesem Beispiel beträgt der Abstand bei der metrischen MDS 0.0820, bei der Hauptkoordinaten-MDS 0.1575.[29]

[26] Das Skript ist `mdsm1.cf`.

[27] Die Abbildung wurde mit dem Skript `mdsplot6.cf` erzeugt.

[28] Die Prokrustes-Rotation wurde mit dem Skript `pr3.cf` durchgeführt.

[29] Für die analog zur Vorgehensweise in Abschnitt 4.2 (§ 5) durchgeführten Berechnungen wurden die Skripte `mdsm1a.cf` und `mdsm1b.cf` verwendet.

4.4 Nichtmetrische MDS-Verfahren

1. Die Problemstellung. In diesem Abschnitt besprechen wir das Verfahren der *nichtmetrischen MDS*. Das Verfahren wurde Mitte der 1960er Jahre von J. B. Kruskal vorgeschlagen;[30] seitdem gibt es viele weitere Beiträge.[31] Ausgangspunkt ist wiederum eine Abstandsmatrix $\mathbf{D} = (d_{ij})$, und gesucht ist eine Konfiguration $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$, deren Abstände die vorgegebenen Abstände d_{ij} möglichst gut repräsentieren. Wir beschränken uns wie bisher auf den Fall $p = 2$ und nehmen an, dass für die Konfiguration euklidische Abstände verwendet werden. Im Unterschied zum Vorgehen bei der metrischen MDS wird jetzt jedoch nur gefordert, dass die Ordnungen der vorgegebenen und der durch die Konfiguration gebildeten Abstände sich möglichst gut entsprechen sollen.

Um das zu präzisieren, nehmen wir zunächst an, dass es in der Abstandsmatrix $\mathbf{D}$ keine Bindungen gibt. Dann können die Abstände d_{ij} in eine streng aufsteigende Reihenfolge gebracht werden, und mit passend gewählten Indizes kann man einen Vektor

$$\mathbf{d} = (d_1, \ldots, d_q)' \quad \text{mit} \quad d_1 < d_2 < \cdots < d_q$$

bilden, wobei $q := n\,(n-1)/2$ die Anzahl der Abstände im unteren Dreieck der Abstandsmatrix $\mathbf{D}$ ist (da $\mathbf{D}$ symmetrisch ist, genügt es, diese Abstände zu betrachten). Ist nun eine Konfiguration $\mathbf{X}$ gegeben, gibt es korrespondierend zu jedem Abstand d_{ij} einen Abstand zwischen $\mathbf{x}_i$ und $\mathbf{x}_j$, den wir mit d^x_{ij} bezeichnen. Diese Abstände werden analog zu $\mathbf{d}$, also in derselben Reihenfolge, zu einem Vektor

$$\mathbf{d}^x = (d^x_1, \ldots, d^x_q)'$$

zusammengefasst. Gesucht ist schließlich eine Konfiguration, die möglichst gut folgender Bedingung genügt:

$$d_j \le d_k \implies d^x_j \le d^x_k \tag{4.16}$$

und natürlich soll auch $d^x_1 < d^x_q$ sein. Zwar ist nicht sicher, dass man eine solche Konfiguration finden kann; aber man kann jedenfalls die Menge der Vektoren angeben, die zu einer im Sinne des Kriteriums (4.16) perfekten Lösung führen würden, nämlich

$$\mathcal{R}_q := \{(r_1, \ldots, r_q)' \,|\, r_1 \le r_2 \le \cdots \le r_q, r_1 < r_q\} \tag{4.17}$$

Wenn man eine Konfiguration $\mathbf{X}$ finden kann, so dass $\mathbf{d}^x \in \mathcal{R}_q$ ist, hat man eine perfekte Lösung gefunden. Ansonsten ist eine möglichst gute Annäherung gesucht, was durch folgende Forderung präzisiert wird: Gesucht ist

[30] Vgl. Kruskal (1964a, 1864b).

[31] Wir beziehen uns u.a. auf Kruskal und Wish (1978); Cox und Cox (1994: 42ff.).

eine Konfiguration $\mathbf{X}^*$, so dass

$$\min_{\mathbf{r}\in\mathcal{R}_q} \| \mathbf{d}^{x^*} - \mathbf{r} \| \;\leq\; \min_{\mathbf{r}\in\mathcal{R}_q} \| \mathbf{d}^{x} - \mathbf{r} \| \tag{4.18}$$

für alle möglichen Konfigurationen $\mathbf{X}$ ist. Eine äquivalente Formulierung verwendet eine explizite Bezeichnung eines Vektors $\check{\mathbf{d}}^x \in \mathcal{R}_q$, der zu $\mathbf{d}^x$ einen minimalen Abstand hat, für den also

$$\| \check{\mathbf{d}}^x - \mathbf{d}^x \| \leq \| \mathbf{r} - \mathbf{d}^x \| \quad \text{(für alle } \mathbf{r} \in \mathcal{R}_q)$$

gilt.[32] Mit dieser Bezeichnung kann folgende *Stressfunktion* definiert werden:

$$s(\mathbf{X}) := \frac{\| \check{\mathbf{d}}^x - \mathbf{d}^x \|}{\| \mathbf{d}^x \|} = \sqrt{\frac{\sum_{k=1}^q (\check{d}_k^x - d_k^x)^2}{\sum_{k=1}^q (d_k^x)^2}} \tag{4.19}$$

und die Aufgabe besteht darin, eine Konfiguration $\mathbf{X}$ zu finden, die diese Stressfunktion minimiert.[33]

2. Berücksichtigung von Bindungen. Um die Stressfunktion (4.19) zu minimieren, ist es insbesondere erforderlich, dass man zu einem Vektor $\mathbf{d}^x$ (für irgendeine Konfiguration $\mathbf{X}$) die Projektion $\check{\mathbf{d}}^x$ berechnen kann.

Bisher wurde angenommen, dass es in der Abstandsmatrix $\mathbf{D}$ keine Bindungen gibt.

3. Berechnungsmethoden. Wir beginnen mit einem Algorithmus, der zuerst von Kruskal angegeben wurde.[34]

4. Vollständige Stressreduktion.

5. Unvollständige Stressreduktion. Bereits bei Kruskal (1964b: 116).

6. Das Shepard-Diagramm. Hilfsmittel.[35]

[32] Dieser Vektor ist eindeutig bestimmt und wird auch als Projektion von $\mathbf{d}^x$ auf $\mathcal{R}_q$ bezeichnet; vgl. Rohwer und Pötter (2002a: 179).

[33] In der Literatur findet man auch noch andere Varianten der Stressfunktion; vgl. Kruskal und Wish (1978: 26).

[34] Vgl. Kruskal (1964b); eine Beschreibung findet man auch bei Cox und Cox (1994: 50ff.).

[35] Diagramme dieser Art wurden zuerst von Shepard (1962) verwendet. Vgl. auch Kruskal und Wish (1978: 19).

4.5 Zusätzliche Merkmalsachsen

1. Ergänzungen der MDS-Bilder. Bei bildlichen Darstellungen von MDS-Konfigurationen können (bestenfalls) Abstände zwischen den dargestellten Punkten interpretiert werden. Da die Konfigurationen beliebig verschoben und gedreht werden können, haben Richtungen keine Bedeutung. Somit stellt sich die Frage, ob noch weitere Informationen, die vielleicht über die dargestellten Objekte verfügbar sind, in den Bildern dargestellt werden können.

Diese Frage stellt sich natürlich nur dann, wenn die Punkte in einem MDS-Bild nicht bereits identifizierbaren Objekten entsprechen (wie beispielsweise in Abbildung 4.2-4) oder bestimmten Gruppen zugeordnet werden können, die dann durch jeweils besondere Symbole kenntlich gemacht werden können. Dann kann man an folgende Möglichkeiten denken.

a) Eine einfache Möglichkeit entsteht, wenn es für die dargestellten Objekte eine Reihenfolge gibt, beispielsweise eine zeitliche Reihenfolge. Dann können die Punkte in der MDS-Konfiguration entsprechend ihrer bekannten Reihenfolge durch Linien verbunden werden.

b) Eine andere Möglichkeit entsteht, wenn man für die dargestellten Objekte außer der Abstandsmatrix auch noch Werte einer oder mehrerer quantitativer (nicht unbedingt metrischer) Variablen kennt. Dann kann man für jede dieser Variablen eine Achse einzeichnen, so dass die Projektionen der Punkte auf diese Achse mit den Variablenwerten möglichst gut korrelieren.

Im Folgenden illustrieren wir die beiden Möglichkeiten mit einem einfachen Beispiel.

2. Illustration mit Schulabschlüssen. Als Beispiel verwenden wir Daten über Schulabschlüsse aus dem ALLBUS. Der kumulierte ALLBUS (1980–2002) erlaubt, sowohl nach dem Gechlecht als auch nach Geburtskohorten zu differenzieren. Tabelle 4.5-1 zeigt die Daten für insgesamt 14108 Männer und 15618 Frauen.[36] Für die gegenwärtige Illustration verwenden wir nur die Daten für Frauen und bilden mit ihnen eine $(14, 14)$-Abstandsmatrix. Zur Berechnung von Abständen zwischen den Verteilungen wird der Dissimilaritätsindex verwendet.[37]

Mit dieser Abstandsmatrix wird eine metrische MDS durchgeführt.[38] Bei 100 Wiederholungen mit zufälligen Anfangskonfigurationen beträgt der minimale Stresswert 0.0055 und wird in 28 der 100 Wiederholungen erreicht. Abbildung 4.5-1 zeigt die resultierende Konfiguration. Die Punkte wurden entsprechend der zeitlichen Reihenfolge der Geburtskohorten

[36] Datenfiles: `bi1.dat` (insgesamt), `bi1m.dat` (nur Männer), `bi1f.dat` (nur Frauen).

[37] Berechnet mit dem Skript `bi2f.cf`; die Abstandsmatrix wird `bi2f.dat` genannt.

[38] Das Skript ist `mdsm3f.cf`.

Tabelle 4.5-1 Verteilungen (in %) der Schulabschlüsse, differenziert nach Geburtskohorten und Geschlecht. 1 = ohne Abschluss, 2 = Hauptschulabschluss, 3 = Realschulabschluss, 4 = Fachhochschulreife, 5 = Abitur. Quelle: Kumulierter ALLBUS 1980–2002.

	Männer					Frauen				
Geburtskohorte	1	2	3	4	5	1	2	3	4	5
1908-1912	2.7	70.5	14.1	2.5	10.2	3.3	77.8	14.2	0.8	3.9
1913-1917	1.0	68.6	17.1	2.4	10.9	3.6	71.7	17.8	1.4	5.4
1918-1922	1.9	65.6	16.2	4.0	12.4	3.9	72.9	15.7	1.6	5.8
1923-1927	1.0	67.0	14.3	3.8	13.9	3.0	68.2	16.9	3.0	8.9
1928-1932	3.0	65.8	16.1	4.3	10.9	4.4	68.9	18.2	2.3	6.2
1933-1937	1.8	65.8	16.6	5.2	10.6	2.7	71.0	18.3	2.1	5.9
1938-1942	0.6	57.9	21.8	6.1	13.7	1.7	62.1	25.4	2.4	8.5
1943-1947	0.9	51.6	24.1	6.4	17.1	1.3	56.4	29.2	3.3	9.8
1948-1952	1.1	49.1	22.0	7.4	20.4	0.7	54.5	28.0	4.5	12.3
1953-1957	1.1	42.5	21.6	10.4	24.4	1.2	43.3	31.6	5.7	18.2
1958-1962	1.2	36.5	24.9	9.1	28.4	1.3	32.7	37.2	6.3	22.6
1963-1967	1.5	31.2	27.2	8.2	31.9	1.2	26.7	38.0	7.3	26.9
1968-1972	1.1	31.1	27.2	9.0	31.7	0.8	24.3	39.9	8.1	27.0
1973-1977	0.9	19.6	30.9	9.1	39.5	3.6	16.3	33.7	6.6	39.8

verbunden; die Richtung wird durch den Pfeil angegeben.

3. Konstruktion ergänzender Achsen. Jetzt verfolgen wir die zweite der eingangs erwähnten Möglichkeiten.[39] Die MDS-Konfiguration sei durch die Koordinaten (x_i, y_i) für $i = 1, \ldots, n$ gegeben; außerdem gebe es für die Punkte Werte $v_1, \ldots, v_n$ einer quantitativen (nicht unbedingt auch metrischen) Variablen.

Die Achsenkonstruktion verläuft folgendermaßen. Wenn man die Konfiguration um einen Winkel ϕ (im Uhrzeigersinn) dreht, gewinnt man die neuen X-Koordinaten durch

$$x_i^\phi = x_i \cos(\phi) + y_i \sin(\phi)$$

Also kann man einen Winkel ϕ bestimmen, so dass die Rangkorrelation zwischen

$$(x_1^\phi, \ldots, x_n^\phi) \quad \text{und} \quad (v_1, \ldots, v_n)$$

maximal wird.[40] Dann wird eine Achse verwendet, die mit der X-Achse des Koordinatensystems den Winkel ϕ bildet. Sie hat die Eigenschaft, dass die Projektionen der Punkte der Konfiguration auf diese Achse mit den

[39] Überlegungen hierzu in dem bei Green, Carmone und Smith (1989: 318ff.) abgedruckten Beitrag zum Programm PROFIT von Chang und Carroll. Hinweise geben auch Jones und Koehly (1993: 110ff.).

[40] Wir verwenden Kendalls Rangkorrelation; vgl. Rohwer und Pötter (2002a: 164). Auch andere Korrelationsmaße könnten verwendet werden, zum Beispiel der gewöhnliche Korrelationskoeffizient; vgl. Holtmann (1975).

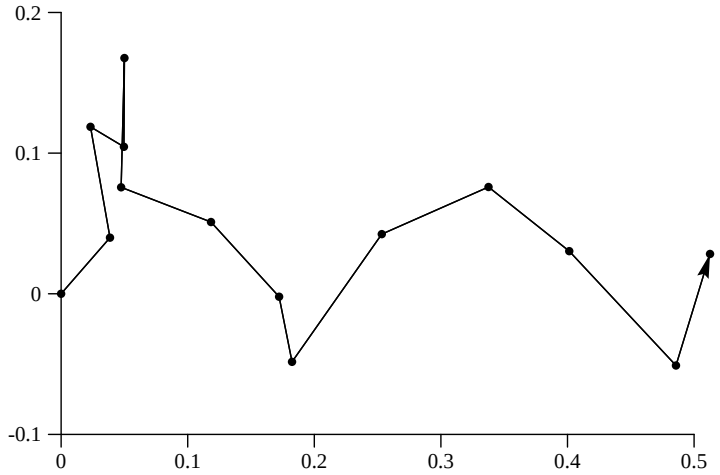


Abb. 4.5-1 MDS-Konstellation der Abstände zwischen den 14 Schulabschlussverteilungen für Frauen in Tabelle 4.5-1.

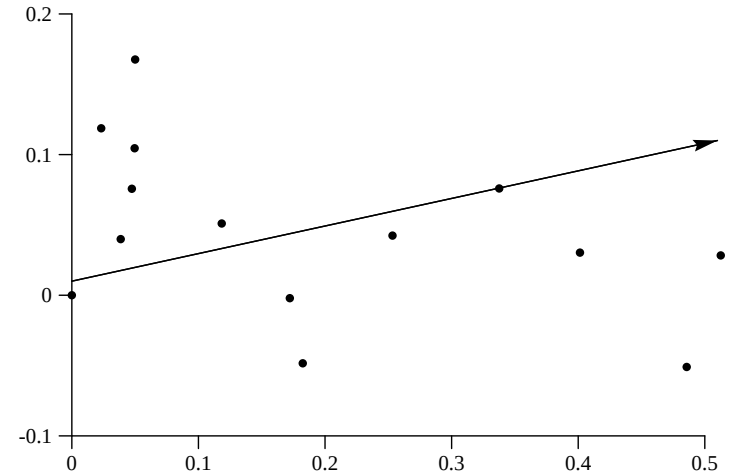


Abb. 4.5-2 MDS-Konstellation der Abstände zwischen den 14 Schulabschlussverteilungen für Frauen mit einer zusätzlichen Achse für die zeitliche Richtung der Geburtskohorten.

auf dieser Achse vorstellbaren Werten $v_1, \ldots, v_n$ am besten korrelieren. Natürlich entsteht eine informative Achse nur dann, wenn eine hohe Korrelation erzielt wird.

Abbildung 4.5-2 zeigt das Ergebnis für unser Beispiel, wobei als ergänzende Variable die zeitliche Reihenfolge der Geburtskohorten verwendet wird. Als optimalen Winkel findet man $\phi = 0.2$ ($= 11.5°$); die Rangkorrelation hat dann den Wert 0.956.[41]

[41] Für die Berechnungen wurde die TDA-Prozedur `mdsr` verwendet; das Skript ist `mdsr1f.cf`. Die Achse wurde so eingezeichnet, dass sie durch den Mittelpunkt der Konfiguration geht.

Kapitel 5

Reihenfolgen und Relationen

In gewisser Weise bildet die eindimensionalen Skalierung nur einen Spezialfall der multidimensionalen Skalierung. Anstatt sich in erster Linie für räumliche Bilder zu interessieren, kann man sich jedoch im eindimensionalen Fall auch noch an zwei anderen Fragen orientieren. Erstens kann man sich auf die Frage beziehen, wie man unter Verwendung von Abstandsinformationen eine Menge von Objekten am besten in einer Reihenfolge anordnen kann. Diese Variante der Fragestellung erscheint insbesondere dann sinnvoll, wenn man aus theoretischen Gründen die Existenz einer Reihenfolge unterstellen kann; beispielweise in der Archäologie, wo es darum geht, für eine Menge gefundener Artefakte eine zeitliche Reihenfolge zu bestimmen. Zweitens kann man eindimensionale Skalierung als eine Methode der Quantifizierung auffassen. Man versucht dann, die den Objekten durch das Skalierungsverfahren zugerechneten Zahlen als quantitativ interpretierbare Scores aufzufassen.

In diesem Kapitel verfolgen wir die erste Fragestellung. Im ersten Abschnitt werden zwei Varianten des Seriationsproblems besprochen; im zweiten Abschnitt werden einige andere Verfahren zur Konstruktion von Ordnungen behandelt. Ansätze, die eindimensionale Skalierung als Verfahren

zur Quantifizierung interpretieren, werden im nächsten Kapitel besprochen.

5.1 Seriation und Skalierung

1. Unterschiedliche Problemformulierungen. Wir beziehen uns auf n Objekte, die durch eine Zahlenmenge $\mathcal{N} = \{1, \ldots, n\}$ repräsentiert werden und für die eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist.

a) In einer ersten Problemformulierung geht es nur darum, für die Objekte eine Reihenfolge zu bestimmen. Wir sprechen in diesem Fall von einem (nichtmetrischen) *Seriationsproblem*.[1]

b) In einer zweiten Problemformulierung geht es darum, korrespondierend zu den n Objekten reelle Zahlen $x_1, \ldots, x_n$ zu finden, so dass die (euklidischen) Abstände zwischen diesen Zahlen möglichst den vorgegebenen Abständen entsprechen, d.h. die Zahlen sollen aus einer Minimierung des folgenden Kriteriums gewonnen werden:[2]

$$f(x_1, \ldots, x_n) = \sum\nolimits_{j<i} (d_{ij} - |x_i - x_j|)^2 \tag{5.1}$$

Wir sprechen in diesem Fall von einem Problem der (metrischen) *eindimensionalen Skalierung*.

2. Ein Beispiel aus der Archäologie. Seriationsprobleme treten zum Beispiel in der Archäologie auf, wenn es sich darum handelt, Informationen über Ähnlichkeiten zwischen Artefakten zur Begründung einer zeitlichen Reihenfolge auszunutzen.[3] Zur Illustration übernehmen wir ein Beispiel von P. Ihm (1978: 483), das auf Daten von V. Elisseeff (1968) beruht. Tabelle 5.1-1 zeigt die Daten.[4] Es handelt sich um chinesische Bronzegefäße, die in 17 Typen eingeteilt wurden. Für jeden Typ gibt es Werte von acht 0-1-Variablen, deren Bedeutung am Ende der Tabelle angegeben ist.

Um aus den Daten eine Abstandsmatrix zu erzeugen, verwenden wir die Hamming-Distanz

$$d_{ij} := \sum_{k=1}^{8} |x_{ik} - x_{jk}|$$

[1] Der Ausdruck 'Seriation' wurde von D. G. Kendall (1971) eingeführt.

[2] Die Notation $\Sigma_{j<i}$ soll bedeuten, dass über alle Elemente unterhalb der Hauptdiagonalen der Abstandsmatrix summiert wird.

[3] Zur Diskussion von Seriationsproblemen in der Archäologie vgl. Laxton (1997).

[4] Ein Eintrag für den Typ Z wurde aufgrund der Angaben bei Elisseeff (1968: 109) verändert.

Tabelle 5.1-1 Werte von acht Variablen für 17 Typen chinesischer Bronzegefäße (`arch1.dat`). Quelle: Ihm (1978: 483), Elisseeff (1968: 109).

	Typ	Anzahl	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	A	14	1	1	1	1	1	1	1	1
2	B	1	1	1	1	0	1	1	1	1
3	C	5	1	0	1	1	1	1	1	1
4	D	18	1	0	0	1	1	1	1	1
5	F	1	1	1	1	1	0	1	1	1
6	H	1	1	0	1	1	0	1	1	1
7	J	11	1	0	0	1	0	1	1	1
8	K	1	1	0	0	0	0	1	1	1
9	M	1	1	0	0	0	0	0	1	1
10	N	14	0	0	0	1	0	1	0	0
11	P	1	0	0	0	1	0	0	0	0
12	R	6	1	0	0	0	1	1	1	1
13	S	1	1	0	0	0	1	0	1	1
14	T	1	1	0	0	0	1	0	1	0
15	V	32	1	0	0	0	1	1	0	0
16	X	2	1	0	0	0	1	0	0	0
17	Z	2	1	0	0	0	0	0	0	0

X_1 Position des Bügels: lateral (1), transversal (0)
X_2 Querschnitt des Bügels: gedreht (1), flach (0)
X_3 Bügelaufhängung: Ring (1), Maske (0)
X_4 Griff des Deckels: Knopf (1), Kuppel (0)
X_5 Kante: vorhanden (1), fehlernd (0)
X_6 Profil des Deckels: mit Hals (1), ohne Hals (0)
X_7 Ränder des Deckels: mit Vorsprung (1), ohne Vorsprung (0)
X_8 Form des Gefäßes: rund (1), unten ausgebaucht (0)

durch die erfasst wird, in wievielen Variablen die Objekte i und j unterschiedliche Werte aufweisen. Tabelle 5.1-2 zeigt die Abstandsmatrix.[5]

Die Fragestellung lautet nun: Kann man gestützt auf diese Daten eine zeitliche Reihenfolge der Gefäßtypen bestimmen? Dabei soll die Annahme verwendet werden, dass es eine Entsprechung zwischen den Abständen in der Datenmatrix und den zeitlichen Abständen im Auftreten der Gefäßtypen gibt.

3. Bestimmung einer Reihenfolge. In allgemeiner Form kann die Aufgabe folgendermaßen formuliert werden: Gesucht ist eine Permutation

$$\pi : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$$

so dass $\pi(i)$ die Position des Objekts i in der Reihenfolge angibt und die Entfernung der Objekte in der Reihenfolge möglichst gut den vorgegebenen

[5] Erzeugt mit dem Skript `arch2.cf`; das Datenfile mit der Abstandsmatrix wird `arch2.dat` genannt.

Tabelle 5.1-2 Aus den Daten in Tabelle 5.1-2 mit der Hamming-Distanz erzeugte Abstandsmatrix (`arch2.dat`).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	1	1	2	1	2	3	4	5	6	7	3	4	5	5	6	7
2	1	0	2	3	2	3	4	3	4	7	8	2	3	4	4	5	6
3	1	2	0	1	2	1	2	3	4	5	6	2	3	4	4	5	6
4	2	3	1	0	3	2	1	2	3	4	5	1	2	3	3	4	5
5	1	2	2	3	0	1	2	3	4	5	6	4	5	6	6	7	6
6	2	3	1	2	1	0	1	2	3	4	5	3	4	5	5	6	5
7	3	4	2	1	2	1	0	1	2	3	4	2	3	4	4	5	4
8	4	3	3	2	3	2	1	0	1	4	5	1	2	3	3	4	3
9	5	4	4	3	4	3	2	1	0	5	4	2	1	2	4	3	2
10	6	7	5	4	5	4	3	4	5	0	1	5	6	5	3	4	3
11	7	8	6	5	6	5	4	5	4	1	0	6	5	4	4	3	2
12	3	2	2	1	4	3	2	1	2	5	6	0	1	2	2	3	4
13	4	3	3	2	5	4	3	2	1	6	5	1	0	1	3	2	3
14	5	4	4	3	6	5	4	3	2	5	4	2	1	0	2	1	2
15	5	4	4	3	6	5	4	3	4	3	4	2	3	2	0	1	2
16	6	5	5	4	7	6	5	4	3	4	3	3	2	1	1	0	1
17	7	6	6	5	6	5	4	3	2	3	2	4	3	2	2	1	0

Abständen entspricht. Die Entfernung von zwei Objekten i und j in der durch π gegebenen Reihenfolge kann durch

$$d_{ij}^{\pi} := |\pi(i) - \pi(j)|$$

erfasst werden. Unter Berücksichtigung der Möglichkeit, dass es in der Abstandsmatrix **D** Bindungen geben kann, liefern folgende Bedingungen ein Kriterium dafür, dass die Abstände in der Reihenfolge den vorgegebenen Abständen vollständig (nicht-metrisch) entsprechen:[6]

$$d_{ij} < d_{kl} \implies d_{ij}^{\pi} \le d_{kl}^{\pi} \quad \text{und} \quad d_{ij} > d_{kl} \implies d_{ij}^{x} \ge d_{kl}^{\pi} \tag{5.2}$$

In den meisten Anwendungsfällen kann man natürlich nur erreichen, dass diese Bedingungen möglichst gut erfüllt werden, d.h. dass die Anzahl der Abstandsvergleiche, bei denen eine der Bedingungen verletzt wird, möglichst klein wird.

Um eine optimale Permutation zu finden, können kombinatorische Methoden verwendet werden; das wird im Anhang A.1 näher erläutert. Für unser Beispiel verwenden wir die TDA-Prozedur `uds`, die eine Näherungslösung liefert.[7] Tabelle 5.1-3 zeigt das Ergebnis. Natürlich kann die

[6] Es genügt, alle Elemente im unteren Dreieck der Abstandsmatrix zu betrachten. Also alle d_{ij} mit $i = 2, \ldots, n$ und $j = 1, \ldots, i-1$. Dann werden zu jedem dieser d_{ij}-Abstände alle d_{kl}-Abstände betrachtet, für die gilt: $k = i$ und $l = j, \ldots, k-1$ oder $k > i$ und $l = 0, \ldots, k-1$.

[7] Wir verwenden die Option 2, der ein von D. H. West (1983) entwickelter Algorithmus zur approximativen Lösung des QA-Problems zugrunde liegt. Für die Berechnung wurde das Skript `uds2.cf` verwendet.

Tabelle 5.1-3 Ergebnisse eindimensionaler Skalierungen der Abstandsmatrix in Tabelle 5.1-2.

Nicht-metrische Lösung		Metrische Lösung	
11	P	4.41	P
10	N	4.00	N
17	Z	3.00	Z
16	X	2.59	X
15	V	2.06	V
14	T	1.47	T
9	M	0.53	M
13	S	0.35	S
8	K	-0.35	K
12	R	-0.53	R
7	J	-1.12	J
4	D	-1.29	D
6	H	-2.24	H
3	C	-2.41	C
2	B	-3.24	B
1	A	-3.53	A
5	F	-3.71	F

chronologische Richtung mit den Abständen allein nicht bestimmt werden.

Prüft man, wie gut in diesem Fall die Bedingungen (5.2) erfüllt werden, findet man, dass bei 668 von 9180 Abstandsvergleichen eine der Bedingungen verletzt wird.

4. Metrische eindimensionale Skalierung. Jetzt besprechen wir die metrische eindimensionale Skalierung, bei der nicht nur eine Reihenfolge, sondern außerdem Positionen auf der Zahlengeraden bestimmt werden sollen. Mathematisch betrachtet geht es darum, Zahlen zu finden, die die Funktion (5.1) minimieren. Das Problem ist kompliziert, weil diese Funktion weder stetig differenzierbar noch global konvex ist.[8] Man muss deshalb Permutationsverfahren, (Gradienten-)Verfahren zur Funktionsminimierung und/oder Verfahren der linearen Programmierung kombinieren.[9]

Für die praktische Durchführung der Berechnungen kann wiederum die TDA-Prozedur `uds` verwendet werden. Für die metrische eindimensionale Skalierung verwendet sie einen von Defays (1978) vorgeschlagenen Algorithmus, der durch eine Kombination unterschiedlicher Verfahren ein globales Minimum des Kriteriums (5.1) findet. Dieses Verfahren ist allerdings sehr rechenaufwendig, so dass es nur bis zu einer Anzahl von etwa 20 Objekten praktikabel ist. Für unser Beispiel liefert das Verfahren die in

[8] Darauf gehen wir ausführlicher im Anhang A.2 ein.

[9] Man vgl. hierzu die Beiträge von Defays (1978), Hubert und Arabie (1988), Pliner (1984, 1996), Lau, Leung und Tse (1998); Hubert, Arabie und Meulman (2002); Brusco (2002).

Tabelle 5.1-3 angegebene Lösung.[10] Die Reihenfolge ist offenbar identisch mit derjenigen, die durch das nicht-metrische Verfahren gefunden wurde. Zusätzlich erhält man jetzt zu jedem Objekt i einen Skalenwert x_i, der näherungsweise die Plazierung des Objekts auf der Zahlenachse angibt.

5. Die Qualität der Skalierung. Mit den durch die metrische Skalierung erzeugten x-Werten kann offenbar eine neue Abstandsmatrix $\mathbf{D}^x = (d^x_{ij})$ berechnet werden, wobei $d^x_{ij} := |x_i - x_j|$ ist. Der euklidische Abstand zwischen $\mathbf{D}$ und $\mathbf{D}^x$, also

$$\| \mathbf{D} - \mathbf{D}^x \| = \left(\sum\nolimits_{i \neq j} (d_{ij} - d^x_{ij})^2 \right)^{1/2} \tag{5.3}$$

liefert dann ein Maß für die Qualität der Skalierung. In unserem Beispiel beträgt der Wert 16.79.

Vielleicht informativer ist jedoch die durchschnittliche absolute Differenz zwischen den vorgegebenen und den durch die Skalierung erzeugten Abständen, also

$$\frac{2}{n(n-1)} \sum\nolimits_{j<i} |d_{ij} - d^x_{ij}| \tag{5.4}$$

In unserem Beispiel findet man den Wert 0.81.

6. Skalierung von Berufsgruppen. Wir sollten ein Beispiel mit Berufen bzw. Berufsgruppen einführen, das später auch für die Regression mit Scores verwendet werden kann.

7. Größere Mengen von Objekten. Das in § 4 verwendete Verfahren zur Minimierung des Kriteriums (5.1) ist nur praktikabel, wenn die Anzahl der Objekte klein ist (bis etwa $n = 20$). Bei größeren Anzahlen kann man folgende Möglichkeiten in Betracht ziehen.

a) Man kann zunächst mit dem in § 3 besprochenen approximativen Verfahren eine näherungsweise optimale Reihenfolge ermitteln und dann innerhalb dieser Reihenfolge nach einem Minimum des Kriteriums (5.1) suchen.

b) Man kann ein für die multidimensionale Skalierung konzipiertes Verfahren verwenden. Hierbei stellt sich natürlich erneut das Problem, dass man normalerweise nur lokale Minima findet.

[10] Berechnet mit dem Skript `uds3.cf`.

Tabelle 5.2-1 Dominanzmatrix für Beziehungen zwischen 20 Schimpansen (`dom1.dat`). Quelle: Michaud (1983: 24).

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	David	M		0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1
2	Goliath	M	1		1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
3	Mc Gregor	M	0	0		1	1	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1
4	Flo	F	0	0	0		1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1
5	Fifi	F	0	0	0	0		0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
6	Figan	J	0	0	0	0	1		0	1	1	0	0	0	0	0	0	0	0	0	1	1
7	William	M	0	0	0	0	1	1		0	1	0	0	0	0	0	0	0	1	0	0	1
8	Olly	F	0	0	0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0
9	Flint	J	0	0	0	0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	0
10	Mike	M	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1
11	J. B.	M	0	1	0	1	1	1	1	1	1	0		1	1	0	1	1	1	1	1	1
12	Huxley	M	0	0	0	1	1	1	1	1	1	0	0		0	0	0	0	1	0	1	1
13	Leakey	M	0	0	1	1	1	1	1	1	1	0	0	1		0	0	1	1	1	1	1
14	Hugh	M	0	0	1	1	1	1	1	1	1	0	1	1	1		0	0	1	0	1	1
15	Rodolph	M	1	0	1	1	1	1	1	1	1	0	0	1	1	1		1	1	1	1	1
16	Humphrey	M	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0		1	1	1	1
17	Evered	J	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0		0	0	1
18	Worzle	M	0	0	0	0	1	1	1	1	1	0	0	1	0	1	0	0	1		1	1
19	Faben	J	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0		1
20	Melissa	F	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	

5.2 Dominanzbeziehungen

1. Beispiel einer Dominanzmatrix. Unter einer *Dominanzmatrix* verstehen wir eine (n, n)-Matrix $\mathbf{A} = (a_{ij})$, deren Koeffizienten nur die Werte 0 oder 1 annehmen und folgende Bedeutung haben:

$$a_{ij} = \begin{cases} 1 & \text{wenn der Akteur } i \text{ den Akteur } j \text{ dominiert} \\ 0 & \text{andernfalls} \end{cases}$$

Als Beispiel verwenden wir eine Dominanzmatrix für Beziehungen zwischen 20 Schimpansen, die von P. Michaud (1983: 24) auf der Grundlage der Beobachungen von Jane van Lawick-Goodall (1971) erstellt wurde. Tabelle 5.2-1 zeigt die Daten. Die erste Spalte enthält die von Lawick-Goodall verwendeten Namen, die zweite Spalte gibt an, ob es sich um einen männlichen, weiblichen oder jugendlichen Schimpansen handelt.

2. Konstruktionen linearer Ordnung. Die in Tabelle 5.2-1 angegebenen Daten liefern nicht ohne weiteres eine lineare Ordnung. Die durch sie formulierten Dominanzbeziehungen sind weder in allen Fällen asymmetrisch noch transitiv. Somit stellt sich die Frage, ob sich wenigstens näherungsweise eine lineare Ordnung konstruieren lässt.

Eine naheliegende Idee besteht darin, für die Bildung einer Reihenfolge die Anzahl der jeweils dominierten anderen Schimpansen zu verwenden. Man gelangt dann zu der Reihenfolge, die in der Spalte ganz rechts in

Tabelle 5.2-2 Permutation der Dominanzmatrix aus Tabelle 5.2-1, um die Anzahl der Einsen unterhalb der Hauptdiagonalen zu minimieren.

	10	2	15	1	11	16	14	13	3	4	18	12	19	7	17	6	5	20	9	8	
10		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	10
2	0		1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
15	0	0		1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
1	0	0	0		1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	0	1	1	0		1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	11
16	0	0	0	1	0		1	1	1	1	1	1	1	1	1	1	1	1	1	1	16
14	0	0	0	0	1	0		1	1	1	0	1	1	1	1	1	1	1	1	1	13
13	0	0	0	0	0	1	0		1	1	1	1	1	1	1	1	1	1	1	1	14
3	0	0	0	0	1	0	0	0		1	1	1	0	1	1	1	1	1	1	1	3
4	0	0	0	0	0	0	0	0	0		1	0	1	1	1	1	1	1	1	1	18
18	0	0	0	0	0	0	1	0	0	0		1	1	1	1	1	1	1	1	1	4
12	0	0	0	0	0	0	0	0	0	1	0		1	1	1	1	1	1	1	1	12
19	0	0	0	0	0	0	0	0	1	0	0	0		1	1	0	1	1	1	1	19
7	0	0	0	0	0	0	0	0	0	0	0	0	0		1	1	1	1	1	0	6
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1	1	1	1	1	7
6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		1	1	1	1	17
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1	1	1	5
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	1	8
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		1	9
8	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0		20

Tabelle 5.2-2 dargestellt ist.[11] Da die Daten nicht bereits einer linearen Ordnung entsprechen, führt diese Vorgehensweise jedoch nicht unbedingt zu einem optimalen Ergebnis. Wir besprechen im Folgenden ein Verfahren, bei dem explizit nach einer optimalen Reihenfolge gesucht wird.

3. Ein Permutationsverfahren. Um die Vorgehensweise in allgemeiner Form darstellen zu können, ist es zweckmäßig, eine explizite Notation für Umnummerierungen und Permutationen einzuführen. Dafür beziehen wir uns auf eine Adjazenzmatrix $\mathbf{A}$ für n Objekte. Die Objektmenge wird durch $\mathcal{N} := \{1, \ldots, n\}$ bezeichnet. Eine Umnummerierung besteht nun aus einer Permutation $\pi : \mathcal{N} \longrightarrow \mathcal{N}$, die jeder Nummer $i \in \mathcal{N}$ eine neue Nummer $\pi(i) \in \mathcal{N}$ zuordnet. Jeder Permutation entspricht auch eine neue Adjazenzmatrix, nämlich

$$\mathbf{A}^{\pi} := (a_{\pi(i),\pi(j)})$$

Somit kann die Aufgabe allgemein folgendermaßen beschrieben werden: Man finde aus der Menge aller möglichen Permutationen diejenige Permutation π, bei der in der Adjazenzmatrix $\mathbf{A}^{\pi}$ die Summe der Einsen oberhalb der Hauptdiagonalen maximal und unterhalb der Hauptdiagonalen minimal wird.

Offenbar genügt es, entweder die Summe der Einträge oberhalb der Hauptdiagonalen zu maximieren oder die Summe der Einträge unterhalb

[11] Erstellt mit dem Skript `dom1.cf`.

Box 5.2-1 TDA-Skript `dom4.cf`, um eine optimale Permutation der Dominanzmatrix in Tabelle 5.2-1 zu finden.

```
mfmt = 1.0;
mdeff(A) = dom1.dat;    Einlesen der Dominanzmatrix
mdefc(20,20,0,C);       Nullmatrix (wird nicht benoetigt)
mdefc(20,20,0,U);       Nullmatrix
repeat(n=20,I);         Fuellt das untere Dreieck von U mit Einsen
    repeat(n=20,J);
        if (gt(I,J));
            msetv(1,U(I,J));
        endif;
    endrepeat;
endrepeat;
mqap(U,A,C,Pi);         Quadratic Assignment
mpr(P);                 Permutationsvektor P
mpsym(A,P,B);           A wird mit P permutiert
mpr(B);                 Anzeige der permutierten Matrix B
```

der Hauptdiagonalen zu minimieren. Wir orientieren uns an der zweiten Variante. Zur Formulierung wird eine untere Dreiecksmatrix verwendet, deren Koeffizienten unterhalb der Hauptdiagonalen immer den Wert 1 haben, also $\mathbf{U} = (u_{ij})$, wobei die Koeffizienten durch

$$u_{ij} := \begin{cases} 1 & \text{wenn } i > j \\ 0 & \text{wenn } i \leq j \end{cases}$$

definiert sind. Die Summe der Einträge unterhalb der Hauptdiagonalen in der Adjazenzmatrix $\mathbf{A}^{\pi}$ erhält man dann durch

$$f(\pi) := \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij}\, a_{\pi(i),\pi(j)} \tag{5.5}$$

und die Aufgabe kann folgendermaßen formuliert werden: Man finde aus der Menge aller möglichen Permutationen diejenige Permutation π, die die Funktion $f(\pi)$ minimal macht.

Dies ist eine Variante eines Problems, das in der Literatur als quadratic-assignment-Problem bezeichnet wird, im folgenden abgekürzt QA-Problem. Die Schwierigkeiten seiner Lösung resultieren daraus, dass die Anzahl der möglichen Permutation, nämlich $n! = 1 \cdot 2 \cdots n$, mit wachsendem n schnell sehr groß wird. Beim derzeitigen Entwicklungsstand kann man nur dann optimale Lösungen finden, wenn n nicht größer als 15 bis 20 ist. Bei größeren Matrizen muss man sich mit näherungsweise optimalen Lösungen zufrieden geben.

Um die Berechnung praktisch durchzuführen, verwenden wir das TDA-Skript `dom4.cf` (Box 5.2-1).[12] Tabelle 5.2-2 zeigt die optimale Reihen-

[12] Der hierbei verwendeten TDA-Prozedur `mqap` liegt ein von D. H. West (1983) entwickelter Algorithmus zur approximativen Lösung des QA-Problems zugrunde.

folge und die entsprechend permutierte Dominanzmatrix. Unterhalb der Hauptdiagonalen befinden sich nach der Permutation nur noch 11 Einsen. (Hätte man sich einfach an der Anzahl der Dominierungen orientiert, gäbe es stattdessen die in der Spalte ganz rechts gezeigte Reihenfolge und unterhalb der Hauptdiagonalen befänden sich 17 Einsen.)

HINWEIS: Michaud verwendet ein anderes Verfahren, das wir auch noch besprechen könnten.

4. Konstruktionen partieller Ordnung. Das Ergebnis der eben durchgeführten Permutation ist nicht bereits eine (vollständige) lineare Ordnung. (Es scheint auch keine Zerlegung in Äquivalenzklassen möglich zu sein.)

5. Abstufungen von Dominanzbeziehungen. Eine Verallgemeinerung besteht darin, nicht nur zu unterscheiden, ob Dominanz vorliegt oder nicht, sondern Grade der Dominanz zu definieren, beispielsweise abgestuft von 0 bis 1. Dann ähnlich zu Kapitalverflechtungsmatrizen.

5.3 Relationen

In den beiden vorangegangenen Abschnitten wurden Ansätze zur Konstruktion von Reihenfolgen besprochen. Reihenfolgen sind spezielle Ordnungsstrukturen. Um zu einer allgemeineren Sichtweise zu gelangen, ist es zweckmäßig, von Relationen auszugehen. In diesem Abschnitt wird zuerst erläutert, wie im Weiteren von Relationen gesprochen werden soll; dann wird besprochen, wie ausgehend von relationalen Daten Relationen mit bestimmten Eigenschaften konstruiert werden können.

1. Adjazenzmatrizen und Relationen. Wir beziehen uns wie bisher auf eine Menge von Objekten: $\Omega = \{\omega_1, \ldots, \omega_n\}$. Eine *Adjazenzmatrix* (für diese Objektmenge) ist eine quadratische (n, n)-Matrix $\mathbf{A} = (a_{ij})$, deren Koeffizienten nur die Werte 0 und 1 annehmen und mit folgender Interpretation verbunden werden können: $a_{ij} = 1$, wenn eine Beziehung (einer bestimmten Art) zwischen ω_i und ω_j besteht; andernfalls $a_{ij} = 0$. Unter einer *Relation* verstehen wir im Folgenden eine für eine Objektmenge definierte Adjazenzmatrix.

Relationen können inhaltlich und formal charakterisiert werden. Inhaltliche Charakterisierungen setzen natürlich einen jeweils bestimmten Anwendungsfall voraus. Einige formale Eigenschaften, die Relationen haben können, können allgemein definiert werden:

a) Eine Relation ist *reflexiv*, wenn $a_{ii} = 1$ (für alle i).

b) Eine Relation ist *transitiv*, wenn $a_{ij} = a_{jk} = 1$ impliziert, dass $a_{ik} = 1$ ist; anders formuliert: $a_{ij} + a_{jk} - a_{ik} \leq 1$ (für alle i, j, k).

c) Eine Relation ist *symmetrisch*, wenn $a_{ij} = a_{ji}$ (für alle i, j).

d) Eine Relation ist *antisymmetrisch*, wenn $a_{ij} + a_{ji} \leq 1$ (für alle $i \neq j$).

e) Eine Relation ist *vollständig*, wenn $a_{ij} + a_{ji} \geq 1$ (für alle $i \neq j$).

Eine Relation, die reflexiv, symmetrisch und transitiv ist, wird auch als *Äquivalenzrelation* bezeichnet.

2. Ordnungsrelationen. Einen wichtigen Sonderfall bilden Ordnungsrelationen. Wir unterscheiden hauptsächlich zwei Formen:

a) Eine *Ordnungsrelation* ist reflexiv, transitiv und antisymmetrisch. Ist sie außerdem vollständig, spricht man von einer vollständigen Ordnungsrelation. Als Beispiel kann man an die Relation $\leq$ für Zahlen denken.

b) Eine Relation, die nur reflexiv und transitiv ist, wird als *Präordnungsrelation* bezeichnet. Ist sie außerdem vollständig, spricht man von einer vollständigen Präordnungsrelation. Als Beispiel kann man an Präferenzen für eine Menge von Alternativen denken, bei denen auch Indifferenzen vorkommen.

Tabelle 5.3-1 Adjazenzmatrix der aus X_5 in Tabelle 5.1-1 gebildeten Äquivalenzrelation.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
2	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
3	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
4	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
13	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
14	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
15	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
16	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Man kann sich die Unterscheidung auch anhand von Adjazenzmatrizen verdeutlichen. In beiden Fällen ist die Adjazenzmatrix oberhalb und in der Hauptdiagonalen vollständig mit Einsen besetzt. Bei einer Präordnungsrelation können auch unterhalb der Hauptdiagonalen Einsen vorkommen; dagegen stehen dort bei einer Ordnungsrelation nur Nullen.

3. Mehrfache Relationen. Objekte können auch gleichzeitig durch mehrere Relationen charakterisiert werden. Als Beispiel verwenden wir die in Abschnitt 5.1 (§ 2) erläuterten Daten über chinesische Bronzegefäße. Korrespondierend zu den Variablen $X_1, \ldots, X_8$ in Tabelle 5.1-1 können acht Relationen (Adjazenzmatrizen) $\mathbf{R}_1, \ldots, \mathbf{R}_8$ definiert werden:

$$\mathbf{R}_k = (r_{k,ij}) \quad \text{mit } r_{k,ij} = 1, \text{ wenn } X_k(\omega_i) = X_k(\omega_j) = 1$$

Offenbar handelt es sich um Äquivalenzrelationen. Tabelle 5.3-1 zeigt die Adjazenzmatrix für $\mathbf{R}_5$.

4. Vergleiche von Relationen. Relationen (die für die gleiche Objektmenge definiert sind) können nicht nur gleich oder verschieden, sondern auch mehr oder weniger ähnlich sein. Es gibt zahlreiche Möglichkeiten, um Abstandsfunktionen zu definieren.[13] Sind $\mathbf{R} = (r_{ij})$ und $\mathbf{R}' = (r'_{ij})$ zwei Adjazenzmatrizen der gleichen Ordnung, erhält man eine besonders einfache Abstandsfunktion durch

$$d_r(\mathbf{R}, \mathbf{R}') := \quad \text{Anzahl der Paare } (i,j) \text{ mit } r_{ij} \neq r'_{ij} \tag{5.6}$$

[13] Vgl. etwa Tüshaus (1983: 11ff.).

Zur Berechnung kann auch folgende Formel verwendet werden:

$$d_r(\mathbf{R}, \mathbf{R}') = \sum\nolimits_{i,j} (r_{ij} + r'_{ij} - 2\, r_{ij}\, r'_{ij}) \tag{5.7}$$

5. Anpassung von Relationen. Die Abstandsfunktion d_r kann verwendet werden, um Relationen zu konstruieren, die möglichst gut zu einer oder zu mehreren bereits gegebenen Relationen passen.[14] Angenommen, dass für $k = 1, \ldots, p$ Relationen $R_k = (r_{k,ij})$ gegeben sind. Ein geeignetes Kriterium für die Bildung einer möglichst gut passenden neuen Relation $R' = (r'_{ij})$ ist

$$\sum\nolimits_{k=1,p} d_r(\mathbf{R}_k, \mathbf{R}') \;\longrightarrow\; \min \tag{5.8}$$

Die Eigenschaften, die R erfüllen soll, können als Nebenbedingungen für dieses Minimierungsproblem formuliert werden.

Für die praktische Berechnung ist es nützlicher, das Kriterium auf folgende Weise umzuformulieren:

$$\begin{aligned}\sum\nolimits_{k=1,p} d_r(\mathbf{R}_k, \mathbf{R}') &= \sum\nolimits_k \sum\nolimits_{i,j} r'_{ij} + r_{k,ij} - 2r'_{ij} r_{k,ij} = \\ &\sum\nolimits_k \sum\nolimits_{i,j} r'_{ij} + \sum\nolimits_k \sum\nolimits_{i,j} r_{k,ij} - 2\sum\nolimits_k \sum\nolimits_{i,j} r'_{ij} r'_{ij} r_{k,ij} = \\ &\sum\nolimits_{i,j} \big(p - 2\sum\nolimits_k r_{k,ij}\big) r'_{ij} + \sum\nolimits_k \sum\nolimits_{i,j} r_{k,ij}\end{aligned}$$

Offenbar genügt es, den ersten Termin in der letzten Zeile zu minimieren. Definiert man zur Abkürzung $c_{ij} := p - 2\sum_k r_{k,ij}$, gelangt man also zu folgender Zielfunktion:

$$\sum\nolimits_{i,j} c_{ij}\, r'_{ij} \;\longrightarrow\; \min \tag{5.9}$$

Da die Koeffizientn c_{ij} ganzzahlig sind und für die Variablen r'_{ij} nur die Werte 0 oder 1 zulässig sind, handelt es sich um ein Problem der linearen ganzzahligen Programmierung. Schwierigkeiten für die praktische Berechnung entstehen allerdings daraus, dass die Eigenschaften der zu konstruierenden Relation R' durch Nebenbedingungen angegeben werden müssen, deren Anzahl schnell sehr groß werden kann; insbesondere dann, wenn die zu konstruierende Relation transitiv sein soll.[15]

[14] Wir folgen hier Überlegungen von Tüshaus (1983).

[15] In der Literatur werden deshalb auch Rechenverfahren diskutiert, die nur Näherungslösungen finden; man vgl. Tüshaus (1983) und Schader und Tüshaus (1988).

Kapitel 6

Skalierung als Quantifizierung

Zu Beginn von Kapitel 5 wurde darauf hingewiesen, dass man die eindimensionale Skalierung unter zwei Aspekten betrachten kann: einerseits als ein Verfahren zur Konstruktion von Reihenfolgen, andererseits als ein Verfahren zur Quantifizierung (von Objekten oder Merkmalswerten irgendeiner Art). Dieser zweite Aspekt steht im Mittelpunkt einer Methode der eindimensionalen Skalierung, die – eng verwandt mit der Korrespondenzanalyse – auf einer Verwendung von Eigenvektoren beruht. Wir nennen sie *Skalierung mit Eigenvektoren*. Sie wurde in zahlreichen Publikationen insbesondere von S. Nishisato (1980, 1994) unter der Bezeichnung „dual scaling“ propagiert, und auch einige andere Autoren sehen in ihr ein nützliches Hilfsmittel zur Datenanalyse.[1] In diesem Kapitel besprechen wir im ersten Abschnitt die Grundzüge dieser Skalierungsmethode; dann wird im zweiten Abschnitt auf einen Zusammenhang zur Idee der kanonischen Korrelation (für Kontingenztabellen) hingewiesen; und schließlich wird im dritten Abschnitt gezeigt, wie der Skalierungsansatz mit einem Ansatz der Regressionsrechnung kombiniert werden kann.

[1] Man vgl. etwa Blasius (2001: 67); Greenacre (1993: 48ff.); Faust und Wasserman (1993); Schriever (1983).

6.1 Skalierung mit Eigenvektoren

1. Der Skalierungsansatz. Ausgangspunkt ist wie bei der Korrespondenzanalyse eine Kontingenztabelle $\mathbf{F} = (f_{ij})$ mit n Zeilen und m Spalten. Dabei kann es sich um absolute oder um relative Häufigkeiten handeln. Es wird angenommen, dass sich die m Spalten auf qualitativ unterschiedliche Merkmalswerte beziehen (beispielsweise auf sechs Berufsgruppen wie bei den Daten in Tabelle 2.3-4). Die Aufgabe soll darin bestehen, für diese Merkmalsausprägungen quantitativ interpretierbare Scores $s_1, \ldots, s_m$ zu finden.[2]

Die Überlegung verläuft nun folgendermaßen: Wenn es solche Scores gäbe, könnte man Mittelwerte und Streuungen berechnen; insbesondere den Mittelwert für die gesamte Tabelle:

$$\bar{s} = \sum_i \sum_j f_{ij} s_j / f_{..}$$

und Mittelwerte für jede Zeile:

$$\bar{s}_i = \sum_j f_{ij} s_j / f_{i.}$$

Definiert man nun Streuungen:

$$q(\mathbf{s}) = \sum_i \sum_j f_{ij} (s_j - \bar{s})^2$$

$$q_1(\mathbf{s}) = \sum_i f_{i.} (\bar{s}_i - \bar{s})^2$$

$$q_2(\mathbf{s}) = \sum_i \sum_j f_{ij} (s_j - \bar{s}_i)^2$$

wobei $\mathbf{s} := (s_1, \ldots, s_m)'$ ist, kann man folgende Streuungszerlegung vornehmen:

$$q(\mathbf{s}) = q_1(\mathbf{s}) + q_2(\mathbf{s}) \tag{6.1}$$

Die Gesamtstreuung ist die Summe aus $q_1(\mathbf{s})$, der Streuung zwischen den Zeilen, und $q_2(\mathbf{s})$, der durchschnittlichen Streuung innerhalb der Zeilen. Die Idee besteht nun darin, den Score-Vektor $\mathbf{s}$ so zu bestimmen, dass die Streuung zwischen den Zeilen maximal wird:

$$\frac{q_1(\mathbf{s})}{q(\mathbf{s})} \longrightarrow \max \tag{6.2}$$

Allerdings gibt es keine eindeutige Lösung, denn jede lineare Transformation von $\mathbf{s}$ liefert den gleichen Wert für das Kriterium $q_1(\mathbf{s})/q(\mathbf{s})$. Es wird deshalb eine Normalisierung

$$\sum_i \sum_j f_{ij} s_j = 0 \tag{6.3}$$

[2] Ganz analog kann man dann Scores für die Spalten berechnen; darauf bezieht sich Nishisatos Bezeichnung „dual scaling“.

angenommen. Um nun eine Lösung für das Maximierungsproblem zu finden, wird eine Matrix

$$\mathbf{A} = (a_{ij}) \quad \text{mit Koeffizienten} \quad a_{ij} := \frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}}$$

verwendet. Man kann nämlich zeigen: Wenn $\mathbf{w} = (w_1, \ldots, w_m)'$ ein Eigenvektor von $\mathbf{A}'\mathbf{A}$ ist, erhält man durch

$$\mathbf{s} = (s_1, \ldots, s_m)' = \left(\frac{w_1}{\sqrt{f_{.1}}}, \ldots, \frac{w_m}{\sqrt{f_{.m}}} \right)' \tag{6.4}$$

einen Score-Vektor, der einen Extremwert $q_1(\mathbf{s})/q(\mathbf{s}) = \lambda$ liefert, wobei λ der zu $\mathbf{w}$ gehörige Eigenwert ist.[3] Um Scores zu berechnen, die das Kriterium maximieren, sollte man also den Eigenvektor verwenden, der zum größten Eigenwert von $\mathbf{A}'\mathbf{A}$ gehört.

Allerdings gehört aufgrund der Konstruktion von $\mathbf{A}$ zum größten Eigenwert stets ein Eigenvektor, dessen Komponenten alle gleich 1 sind. Dies würde bedeuten, für alle Spalten der Tabelle den gleichen Score-Wert zu verwenden (was dann die Streuung zwischen den Zeilen der Tabelle maximal macht). Da dies eine „uninteressante“ Lösung ist, wird von Nishisato und anderen Autoren vorgeschlagen, stattdessen den Eigenvektor zu verwenden, der zum zweitgrößten Eigenwert gehört.

2. Berechnung der Score-Werte. Um die Eigenwerte und Eigenvektoren von $\mathbf{A}'\mathbf{A}$ zu berechnen, kann eine Singularwertzerlegung der Matrix $\mathbf{A}$ verwendet werden: $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$. Denn daraus folgt:

$$\mathbf{A}'\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}'\mathbf{U}\mathbf{\Lambda}\mathbf{V}' = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}', \quad \text{also} \quad (\mathbf{A}'\mathbf{A})\mathbf{V} = \mathbf{V}\Lambda^2$$

Sind $\lambda_1, \ldots, \lambda_m$ die Singularwerte und $\mathbf{v}_1, \ldots, \mathbf{v}_m$ die Spalten von $\mathbf{V}$, gewinnt man daraus folgende Gleichungen:

$$(\mathbf{A}'\mathbf{A})\,\mathbf{v}_j = \lambda_j^2\,\mathbf{v}_j \quad (\text{für } j = 1, \ldots, m)$$

Sie zeigen, dass die quadrierten Singularwerte von $\mathbf{A}$ die Eigenwerte von $\mathbf{A}'\mathbf{A}$ und die zugehörigen Eigenvektoren die Spalten der Matrix $\mathbf{V}$ sind.

Zur Illustration der Berechnung verwenden wir ein einfaches Beispiel von Nishisato (1994: 58). Die Daten beziehen sich auf drei Lehrer, deren Unterricht von 29 Schülern bewertet wird:

$$\mathbf{F} = \begin{pmatrix} 1 & 3 & 6 \\ 3 & 5 & 2 \\ 6 & 3 & 0 \end{pmatrix}$$

Die Zeilen entsprechen den Lehrern; die Spalten geben die Anzahlen der

[3] Dies wird genauer im Anhang A.3 erklärt.

Bewertungen „gut“ (1. Spalte), „mittel“ (2. Spalte) und „schlecht“ (3. Spalte) an. Aus diesen Daten gewinnt man die Matrix[4]

$$\mathbf{A} = \begin{pmatrix} 0.1000 & 0.2860 & 0.6708 \\ 0.3000 & 0.4767 & 0.2236 \\ 0.6325 & 0.3015 & 0.0000 \end{pmatrix}$$

Ihre Singularwertzerlegung liefert die Singularwerte

$$\lambda_1 = 1.0000, \quad \lambda_2 = 0.6070, \quad \lambda_3 = 0.1777$$

und die Matrix

$$\mathbf{V} = \begin{pmatrix} -0.5872 & 0.6319 & 0.5059 \\ -0.6159 & 0.0568 & -0.7858 \\ -0.5252 & -0.7730 & 0.3558 \end{pmatrix}$$

Verwendet man die zweite Spalte, die zum zweitgrößten Singularwert gehört, findet man entsprechend (6.4) die Score-Werte

$$s_1 = 0.1998, \quad s_2 = 0.0171, \quad s_3 = -0.2733$$

Wie bereits gesagt wurde, sind beliebige lineare Transformationen möglich, zum Beispiel: $s_j^* = -4.2274\, s_j + 1.8446$. Dann bekommt man die Score-Werte

$$s_1^* = 1, \quad s_2^* = 1.77, \quad s_3^* = 3$$

[4] Die Berechnungen wurden mit den Skripten `ka5.cf` bzw. `ka6.cf` durchgeführt. `ka5.cf` verwendet TDAs Matrix-Befehle, `ka6.cf` verwendet die Option 5 der `dma`-Prozedur für die Duale Skalierung. Die Daten befinden sich im File `ka5.dat`.

6.2 Kanonische Korrelation

1. Der theoretische Ansatz.

6.3 Regression mit Scores

In diesem Abschnitt besprechen wir eine Möglichkeit, die Berechnung von Scores mit einem Regressionsansatz zu kombinieren.[5] Dieser Regressionsansatz kann unter Umständen eine nützliche Alternative zu Logit- und Probit-Modellen sein, wenn die Anzahl der Kategorien der abhängigen Variablen groß ist.

1. Der theoretische Ansatz. Ausgangspunkt ist eine Datenmatrix für n Fälle, die folgende Form hat:

Y	X_1	$\cdots$	X_p
y_1	x_{11}		x_{1p}
$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}		x_{np}

Y ist eine qualitative abhängige Variable mit m Merkmalswerten, $X_1, \ldots, X_p$ sind die unabhängigen Variablen (ggf. einschließlich einer Interzeptspalte). Es sei nun $\mathbf{y}$ der Spaltenvektor mit den Werten von Y und $\mathbf{X}$ die (n, p)-Matrix mit den Werten der unabhängigen Variablen. Wäre Y eine quantitative Variable, hätte ein gewöhnlicher Regressionsansatz folgende Form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

wobei $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)'$ die Regressionskoeffizienten und $\mathbf{e} = (e_1, \ldots, e_n)'$ die Residuen erfasst. Y ist jedoch eine qualitative Variable, und um deren m Ausprägungen numerisch zu repräsentieren, können beliebige Score-Werte verwendet werden.

Angenommen, man hätte die Score-Werte $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m)'$. Definiert man nun eine (n, m)-Matrix $\mathbf{Z} = (z_{ij})$ mit den Elementen

$$z_{ij} := \begin{cases} 1 & \text{wenn } y_i = j \\ 0 & \text{andernfalls} \end{cases}$$

kann ein Regressionsansatz folgendermaßen formuliert werden:

$$\mathbf{Z}\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{6.5}$$

wobei jetzt $\boldsymbol{\alpha}$ und $\boldsymbol{\beta}$ durch eine Minimierung der quadrierten Residuen, also durch

$$\| \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} \|^2 \longrightarrow \min \tag{6.6}$$

zu bestimmen sind.

2. Illustration der Berechnung. Der mathematische Ansatz zur Lösung dieses Problems verläuft analog zu der im vorangegangenen Abschnitt besprochenen Methode und wird im Anhang A.4 besprochen. Hier illustrieren wir die Berechnung anhand eines einfachen Beispiels.

[5] Diese Idee wurde vor längerer Zeit von Rubinfeld (1982) vorgestellt.

Kapitel 7

Ansätze der Clusteranalyse

In diesem Kapitel beginnen wir mit der Diskussion von Methoden der Clusteranalyse bzw. Klassifikation. Zunächst beschäftigen wir uns mit der Frage, wie Cluster definiert werden können. Dann werden partitionierende Verfahren besprochen.

Notationen. $\mathcal{N} = \{1, \ldots, n\}$ repräsentiert die Menge der Objekte, die von beliebiger Art sein können. Es wird angenommen, dass eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist, die für jeweils zwei Elemente $i, j \in \mathcal{N}$ einen Abstand d_{ij} angibt. Für Teilmengen von $\mathcal{N}$, die als Cluster betrachtet werden können, wird meistens der Buchstabe C, für Partitionen wird der Buchstabe P verwendet.

7.1 Unterschiedliche Ansätze

1. Wie können Cluster definiert werden? Clusteranalyse dient hier als Sammelbegriff für eine breite Palette von Verfahren, deren Gemeinsamkeit im Wesentlichen nur darin besteht, dass sie die Objektmenge $\mathcal{N}$ irgendwie in Cluster einteilen. Somit stellt sich zunächst die Frage, wie Cluster definiert werden können. Eine gelegentlich verfolgte Idee für die Definition von Clustern wurde von K. D. Bailey (1983: 255) so formuliert:

„It is axiomatic in cluster analysis, as in all classification, that individuals or variables in a class are considered to be more similar to each other than to individuals or variables not in the class."

Tatsächlich führt diese Idee zu einer sehr engen Clusterdefinition, die sich nur selten realisieren lässt. Um das deutlich zu machen, präzisieren wir zunächst die Formulierung. Eine echte Teilmenge $C \subset \mathcal{N}$ wird ein *separierbares* oder kurz ein *S-Cluster* genannt, wenn C mindestens zwei Elemente hat und folgende Bedingung erfüllt ist:

$$\text{Für alle } i \in C\text{: } \max\{d_{ij} \mid j \in C\} < \min\{d_{ij} \mid j \notin C\} \tag{7.1}$$

$\max\{d_{ij} \mid j \in C\}$ wird auch als *Durchmesser* des Clusters C bezeichnet. Die Bedingung sagt, dass der Clusterdurchmesser kleiner sein sollte als der kleinste Abstand zu einem Objekt außerhalb des Clusters.

Verwendet man das so formulierte Kriterium, lautet die Frage, ob sich $\mathcal{N}$ in zwei oder mehr S-Cluster einteilen lässt. Wie man sehen wird, ist das oft nicht der Fall.

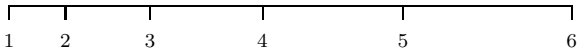
Als Beispiel betrachten wir eine Menge $\mathcal{N} = \{1, \ldots, 5\}$, die fünf Schulabschlüsse repräsentiert: 1 = ohne Hauptschulabschluss, 2 = Hauptschulabschluss, 3 = Realschulabschluss, 4 = Fachhochschulreife, 5 = Abitur. Es wird folgende Abstandsmatrix angenommen:

$$\begin{pmatrix} 0.0 & 2.0 & 3.0 & 4.5 & 5.5 \\ 2.0 & 0.0 & 1.0 & 2.5 & 3.5 \\ 3.0 & 1.0 & 0.0 & 1.5 & 2.5 \\ 4.5 & 2.5 & 1.5 & 0.0 & 1.0 \\ 5.5 & 3.5 & 2.5 & 1.0 & 0.0 \end{pmatrix} \tag{7.2}$$

Man erkennt, dass es kein S-Cluster gibt, das den Schulabschluss 1 als Element enthält (denn wollte man den Abschluss 2 hinzufügen, müsste man auch alle anderen Abschlüsse mit aufnehmen). Es liegt also nahe, vor der Bildung von S-Clustern alle Elemente aus $\mathcal{N}$ zu entfernen, die keine Elemente von S-Clustern sein können. In unserem Beispiel ist das das Element 1. Es bleibt die reduzierte Menge $\{2, 3, 4, 5\}$, die sich in die beiden S-Cluster $C_1 = \{2, 3\}$ und $C_2 = \{4, 5\}$ zerlegen lässt.

Es ist natürlich möglich, dass alle Elemente von $\mathcal{N}$ keine Elemente von S-Clustern sein können. So verhält es sich zum Beispiel bei den folgenden 6

Punkten, deren Abstände durch ihre Lage auf einer Graden gegeben sind:



In diesem Fall müssten beginnend mit Nr. 6 sukzessive alle Punkte entfernt werden, so dass überhaupt keine S-Cluster gebildet werden können.[1]

2. S-Cluster bei Berufsstrukturdaten. Als ein weiteres Beispiel betrachten wir die Berufsstrukturdaten aus Abschnitt 2.3. Geht man von der Abstandsmatrix für die acht Länder aus (Tabelle 2.3-2), findet man, dass die Länder 6 (Schweden) und 8 (Japan) nicht für S-Cluster verwendet werden können. Die restlichen Länder können in drei S-Cluster partitioniert werden: (Griechenland, Türkei), (Deutschland, Schweiz), (Grossbritannien, USA).

Vielversprechender ist vielleicht die in § 6 von Abschnitt 2.3 definierte Abstandsmatrix für die 16 geschlechtsspezifischen Berufsverteilungen. Man kann ihre Struktur folgendermaßen andeuten:

$$\left(\begin{array}{ll|l} \text{M1} & \min = 0.05 & \min = 0.24 \\ \vdots & \max = 0.20 & \max = 0.50 \\ \text{M8} & & \\ \hline \text{F1} & \min = 0.24 & \min = 0.05 \\ \vdots & \max = 0.50 & \max = 0.32 \\ \text{F8} & & \end{array}\right)$$

Man erkennt, dass die Verteilungen der Männer (oberer linker Quadrant) zu einem S-Cluster zusammengefasst werden können, nicht jedoch die Verteilungen der Frauen. F2 (Türkei) und F6 (Schweden) sind isoliert, nur die restlichen F-Verteilungen können zu einem S-Cluster zusammengefasst werden. Der hierarchische Bildungsprozess von S-Clustern kann anhand der Ausgabe der `scla`-Prozedur (Box 7.1-1) rekonstruiert werden.

3. Abgeschwächte Clusterkonzeptionen. In der Literatur findet man eine Fülle unterschiedlicher Clusterkonzeptionen, die im Vergleich zur Idee der S-Cluster oft nur sehr schwache Anforderungen stellen. Hier sind einige Beispiele.

„Classification can be described as the activity of dividing a set of objects into a smaller number of classes in such a way that objects in the same class are similar to one another and dissimilar to objects in other classes.“ (Gordon 1987: 119)

„[. . .] an investigator would like to group together variables so that they are as homogenous as possible within subsets, and as different as possible between subsets.“ (Cliff et al. 1986: 201)

„*Cluster analysis* refers to a wide variety of techniques used to group entities into homogeneous subgroups on the basis of their similarities.“ (Lorr 1983: 1)

[1] Zur Berechnung von S-Clustern kann die TDA-Prozedur `scla` verwendet werden.

Box 7.1-1 Ausgabe der `scla`-Prozedur zur Berechnung von S-Clustern für die Abstandsmatrix der geschlechtsspezifischen Berufsverteilungen.

```
2 isolated points: 9,14
Remaining number of points: 14

Size 2
   1 -: 1 2
   2 -: 3 5
   3 -: 4 7
   4 -: 10 16
   5 -: 11 13
Size 3
Size 4
   1 -: 11 12 13 15
Size 5
Size 6
   1 -: 10 11 12 13 15 16
Size 7
Size 8
   1 -: 1 2 3 4 5 6 7 8
Size 9
Size 10
Size 11
Size 12
```

„Basically, one wants to form groups in such a way that objects in the same group are similar to each other, whereas objects in different groups are as dissimilar as possible.“ (Kaufman und Leonard 1990: 1)

„Roughly speaking, the goal of a clustering algorithm is to group the objects of a database into a set of meaningful subclasses.“ (Ankerst et al. 1999: 49-60)

4. Abstände und Häufigkeiten. Bei vielen Überlegungen zur Clusteranalyse vermischen sich zwei Ideen: Einerseits die Idee, dass Objekte innerhalb desselben Clusters ähnlich sein sollten; und andererseits eine Idee, die mit Häufigkeiten operiert: dass Cluster aus denjenigen Objekten gebildet werden sollten, die „gehäuft“ vorkommen. Zum Beispiel:

„The goal of clustering, in general, is to discover dense and sparse regions in a dataset.“ (Ganti, Gehrke und Ramakrishnan 1999: 73)

Es ist bemerkenswert, dass es keinen wesentlichen Zusammenhang zwischen den beiden Ideen gibt. Überlegungen, die mit Ähnlichkeiten bzw. Abständen argumentieren, sind zunächst von wesentlich anderer Art als Überlegungen, die mit Häufigkeiten argumentieren. Während Häufigkeiten eine Bezugnahme auf Daten voraussetzen, ist das bei einer Betrachtung von Abständen nicht erforderlich. Überlegungen, die mit Abständen argumentieren, können sich beispielsweise auf Merkmalsräume beziehen, ohne dass Daten erforderlich sind. Ein gutes Beispiel ist die Kemeny-Metrik für Rangordnungen (vgl. Abschnitt 2.1, § 3).

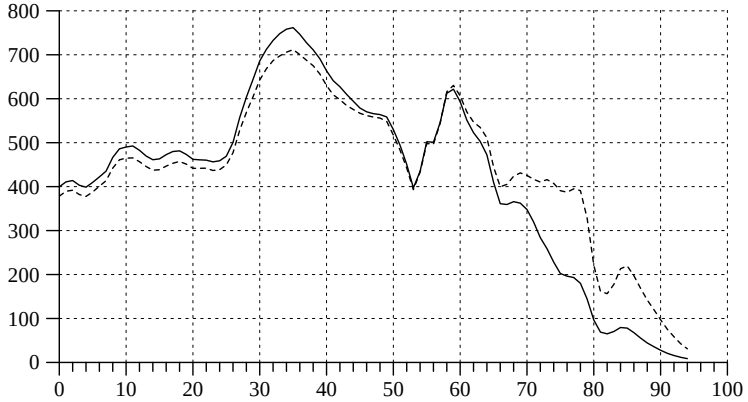


Abb. 7.1-1 Altersverteilung (absolute Häufigkeiten in 1000) der Männer (durchgezogene Linie) und Frauen (gestrichelte Linie) in Deutschland 1999.

Um die Problematik von an Häufigkeiten orientierten Clusteranalysen zu verdeutlichen, genügt bereits eine Betrachtung eindimensionaler Häufigkeitsfunktionen. Als Beispiel betrachte man die Altersverteilungen in Abbildung 7.1-1. Kann man anhand dieser Häufigkeitsfunktionen sinnvolle Cluster abgrenzen?

5. Ein zweidimensionales Beispiel. Es ist nützlich, sich die unterschiedlichen Ansätze auch anhand eines zweidimensionalen Beispiels zu verdeutlichen. Dafür verwenden wir 100 Werte einer zweidimensionalen Normalverteilung, die ersten 50 mit dem Mittelwert $(3, 3)$ die anderen 50 mit dem Mittelwert $(6, 5)$. Abbildung 7.1-2 zeigt die erzeugten Punkte.[2]

Wie könnte in diesem Beispiel ein Ansatz verfolgt werden, der sich an Häufigkeiten orientiert. Abbildung 7.1-3 macht deutlich, dass es jedenfalls nicht genügen würde, nur die eindimensionalen Häufigkeitsprojektionen zu betrachten.

Dagegen führt eine Orientierung an Abständen zu einer anderen Idee. Sie besteht darin, einige Punkte als Clusterzentren auszuwählen und dann alle übrigen Punkte demjenigen Clusterzentrum zuzuordnen, zu dem ihr Abstand am kleinsten ist. Dies ist die Grundidee der sogenannten partitionierenden Verfahren der Clusteranalyse.

6. Eine unvollständige Übersicht. Zum Abschluss dieses Abschnitts geben wir eine kurze Übersicht über die in der Literatur hauptsächlich verfolgten und verwendeten Ansätze der Clusteranalyse.

a) Verfahren der Umgruppierung von Abstandsmatrizen, so dass Objekte, die einen geringen Abstands aufweisen, in möglichst nahe beieinander liegende Zeilen bzw. Spalten plaziert werden.

[2] Die Daten wurden mit dem Skript `cl1.cf`, die Abbildung mit `clplot1.cf` erzeugt.

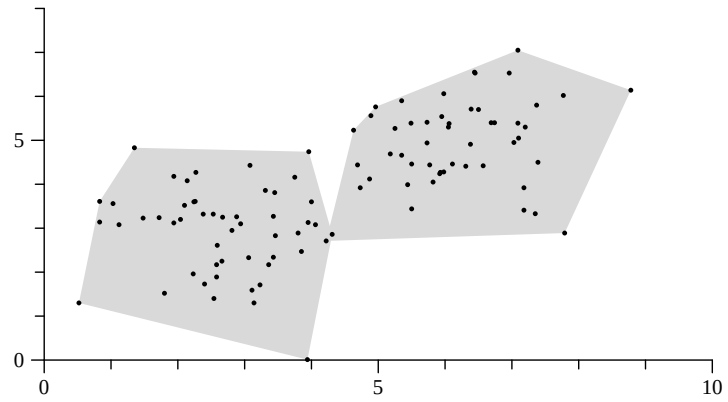


Abb. 7.1-2 100 mit einer zweidimensionalen Normalverteilung erzeugte Punkte; 50 mit dem Mittelwert $(3, 3)$, 50 dem Mittelwert $(6, 5)$.

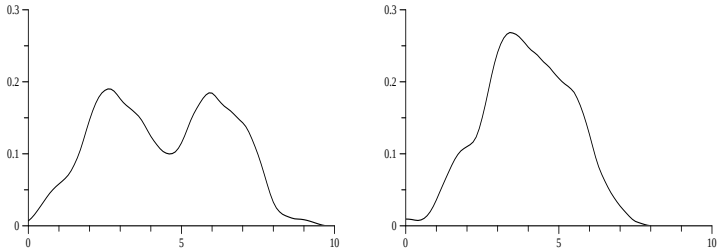


Abb. 7.1-3 Häufigkeitsverteilungen für die X- und Y-Koordinaten der 100 Punkte in Abbildung 7.1-2.

b) Partitionierende Verfahren, bei denen die Anzahl der Cluster vorgegeben werden muss und dann versucht wird, optimale Zuordnungen der Objekte zu Clustern zu finden; dabei werden optimale Zuordnungen durch unterschiedliche Kriterien definiert.

c) Hierarchische Verfahren, die nicht unmittelbar Cluster erzeugen, sondern eine hierarchische Repräsentation der Struktur einer Abstandsmatrix liefern. Sie werden in Kapitel 8 behandelt.

d) Verfahren, die durch Dichotomisierungen einer Abstandsmatrix erzeugte Graphen verwenden. Solche Verfahren werden in Abschnitt 7.3 besprochen.

e) Verfahren, die sich explizit an Häufigkeiten orientieren. Solche Verfah-

ren werden in diesem Text nicht besprochen.[3]

f) Schließlich kann hier auch noch auf eine weitere Vorgehensweise hingewiesen werden, die darin besteht, zunächst räumliche Bilder (einer Abstandsmatrix) zu erzeugen (z.B. mit Verfahren der multidimensionalen Skalierung oder Korrespondenzanalyse) und dann Cluster durch visuelle Anschauung zu bestimmen.[4]

[3] Vgl. Everitt (1993: Kap. 6); Ankerst et al. (1999).

[4] Diese Vorgehensweise wird oft vorgeschlagen, man vgl. beispielsweise Kruskal und Wish (1978: 43ff.), Lorr (1983: 45). Es gibt jedoch auch Kritik, vgl. die Hinweise bei Bailey (1994: 73).

7.2 Klassifikation ordinaler Merkmale

7.3 Verwendung von Graphen

7.4 Modelle für Ähnlichkeiten

Kapitel 8

Hierarchien und Bäume

8.1 Hierarchische Klassifikation
1. Ein allgemeiner Rahmen.
2. SAHN-Algorithmen.
3. Illustration mit Berufsstrukturdaten.
4. Dendrogramme.
5. Bindungen.
6. Vergleiche der SAHN-Verfahren.
7. Monotone Abstandstransformationen.

8.2 Ultrametrische Baummodelle
1. Hierarchien und Bäume.
2. Einfache Baummodelle.
3. Hierarchische Klassifikationsschemas.
4. Darstellung durch Dendrogramme.
5. Alternative Abstandsberechnung.
6. Optimale ultrametrische Modelle.

In diesem Kapitel besprechen wir hierarchische Klassifikationsverfahren, die, anders als partitionierende Verfahren, nicht unmittelbar bestimmte Cluster erzeugen, sondern deren Zweck zunächst darin gesehen werden kann, Einsichten in die Struktur einer Abstandsmatrix zu gewinnen.

Notationen. $\Omega = \{\omega_1, \ldots, \omega_n\}$ ist die Menge der Objekte, die von beliebiger Art sein können. Es wird angenommen, dass eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist, die für jeweils zwei Elemente $\omega_i, \omega_j \in \mathcal{N}$ einen Abstand d_{ij} angibt. Für Teilmengen von Ω, die als Cluster betrachtet werden können, wird meistens der Buchstabe C, für Partitionen wird der Buchstabe P verwendet.

8.1 Hierarchische Klassifikation

Man unterscheidet agglomerative und divisive (zerlegende) Verfahren.[1] Bei agglomerativen Verfahren wird von den einzelnen Objekten ausgegangen, die dann sukzessive zu Clustern zusammengefasst werden; bei divisiven Verfahren wird die Gesamtmenge der Objekte sukzessive in Teilmengen zerlegt. In diesem Kapitel werden nur agglomerativen Verfahren besprochen. Wir beginnen im ersten Abschnitt mit Erläuterungen und Illustrationen einiger hierarchischer Klassifikationsverfahren. Im zweiten Abschnitt wird besprochen, wie diese Verfahren auch als Methoden zur Konstruktion von repräsentierenden Modellen verstanden werden können, eine Betrachtungsweise, die dann auch andere Algorithmen zur Modellberechnung motiviert.

Zu beachten ist, dass hierarchische Klassifikationsverfahren nicht unmittelbar bestimmmte Partitionen einer Objektmenge liefern. Wie das erreicht werden kann, wird erst in Abschnitt 9.1 besprochen. Dort erfolgen auch einige Hinweise auf divisive Verfahren.

1. Ein allgemeiner Rahmen. Man kann sich vorstellen, dass durch ein agglomeratives Klassifikationsverfahren eine Folge von Partitionen $(P_0, P_1, \ldots, P_{n-1})$ der vorausgesetzten Objektmenge Ω entsteht. Begonnen wird mit der Partition $P_0 = \{\{\omega_1\}, \ldots, \{\omega_n\}\}$, deren Cluster jeweils nur ein Objekt enthalten. Bei jedem neuen Level werden dann zwei Cluster der vorangehenden Partition zu einem neuen Cluster zusammengefasst. Diese Betrachtungsweise führt zu einem allgemeinen Schema für agglomerative Verfahren:

(1) Beginne mit der Partition $P_0 = \{\{\omega_1\}, \ldots, \{\omega_n\}\}$.

(2) Wenn eine Partition $P_i = \{C_{i1}, \ldots, C_{in_i}\}$ gegeben ist, die $n_i = n - i$ Cluster enthält, bestimme zwei Cluster C_{ij} and C_{ik}, die zu einem neuen Cluster vereinigt werden sollen.

(3) Erzeuge eine neue Partition P_{i+1}, indem in der Partition P_i die Cluster C_{ij} und C_{ik} durch ein Cluster $C_{ij} \cup C_{ik}$ ersetzt werden.

(4) Solange die neue Partition mehr als ein Cluster enthält, wiederhole den Prozess bei Schritt (2).

2. SAHN-Algorithmen. Zu überlegen ist, wie im zweiten Schritt vorgegangen werden soll. Oft werden sog. *SAHN-Algorithmen* verwendet, eine Abkürzung für *sequential, agglomerative, hierarchical, non-overlapping*.[2]

[1] Übersichten zu den verschiedenen hierarchischen Methoden findet man u.a. bei Gordon (1987); Blashfield und Aldenderfer (1988); Jain und Dubes (1988: 58ff.). Eine neueren Überblick, der auch Erweiterungen des hierarchischen Ansatzes einbezieht, geben Barthélemy, Brucker und Osswald (2007).

[2] Vgl. Jain und Dubes (1988: 79).

Box 8.1-1 Abstandsdefinitionen für einige SAHN-Methoden.

Single Link Methode

$$\rho(C_i, C_j \cup C_k) = \min\{\rho(C_i, C_j), \rho(C_i, C_k)\}$$

Complete Link Methode

$$\rho(C_i, C_j \cup C_k) = \max\{\rho(C_i, C_j), \rho(C_i, C_k)\}$$

WPGMA (Weighted Average) Methode

$$\rho(C_i, C_j \cup C_k) = \frac{1}{2}(\rho(C_i, C_j) + \rho(C_i, C_k))$$

WPGMC (Weighted Centroid) Methode

$$\rho(C_i, C_j \cup C_k) = \frac{1}{2}(\rho(C_i, C_j) + \rho(C_i, C_k)) - \frac{1}{4}\rho(C_j, C_k)$$

UPGMA (Group Average) Methode

$$\rho(C_i, C_j \cup C_k) = \frac{n_j\,\rho(C_i, C_j) + n_k\,\rho(C_i, C_k)}{n_j + n_k}$$

UPGMC (Unweighted Centroid) Methode

$$\rho(C_i, C_j \cup C_k) = \frac{n_j\,\rho(C_i, C_j) + n_k\,\rho(C_i, C_k)}{n_j + n_k} - \frac{n_j\,n_k}{(n_j + n_k)^2}\rho(C_j, C_k)$$

Ward's (Minimum Variance) Methode

$$\rho(C_i, C_j \cup C_k) = \frac{(n_i + n_j)\,\rho(C_i, C_j) + (n_i + n_k)\,\rho(C_i, C_k) - n_i\,\rho(C_j, C_k)}{n_i + n_j + n_k}$$

Die Idee besteht darin, eine Abstandsfunktion für Cluster zu definieren und dann jeweils diejenigen Cluster zusammenzufassen, die den geringsten Abstand aufweisen. Wenn also $P = \{C_1, \ldots, C_m\}$ eine Partition von Ω ist und $\rho(C_j, C_k)$ eine Abstandsfunktion für die Cluster, muss man zwei Cluster $C_{j'}$ und $C_{k'}$ finden, so dass

$$\rho(C_{j'}, C_{k'}) = \min_{j,k}\{\rho(C_j, C_k)\}$$

ist. Die SAHN-Verfahren verwenden eine rekursive Definition der Abstandsfunktion, wobei stets mit $\rho(\{j\}, \{k\}) := d_{jk}$ begonnen wird. In jedem Schritt der agglomerativen Prozedur hat man dann bereits eine Abstandsfunktion ρ für die aktuelle Partition, um die beiden ähnlichsten

Tabelle 8.1-1 Einige SAHN-Methoden, die mit der Formel von Lance und Williams definiert werden können.

Methode	α_1	α_2	β	γ
Single Link	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete Link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
WPGMA (Weighted Average)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
WPGMC (Weighted Centroid)	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
UPGMA (Group Average)	$\frac{n_j}{n_j+n_k}$	$\frac{n_k}{n_j+n_k}$	0	0
UPGMC (Unweighted Centroid)	$\frac{n_j}{n_j+n_k}$	$\frac{n_k}{n_j+n_k}$	$\frac{-n_j\,n_k}{(n_j+n_k)^2}$	0
Ward's (Minimum Variance)	$\frac{n_i+n_j}{n_i+n_j+n_k}$	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$\frac{-n_i}{n_i+n_j+n_k}$	0

Cluster, etwa C_j und C_k, zu finden. Somit genügt es, eine Abstandsfunktion ρ für die neue Partition zu definieren, die aus der Zusammenfassung von C_j und C_k entsteht, also Abstände

$$\rho(C_i, C_j \cup C_k)$$

zu definieren. Die in Box 8.1-1 angegebenen Varianten werden oft verwendet (dabei bedeuten: $n_i := |C_i|$, $n_j := |C_j|$ und $n_k := |C_k|$):[3] In den Abkürzungen steht der erste Buchstabe für *weighted* oder *unweighted*, der letzte für *average* oder *centroid*; die mittleren Buchstaben PGM stehen für *pair group method*.

Lance und Williams (1966) haben folgende Formel entwickelt, mit sich eine ganze Familie von SAHN-Algorithmen definiern lässt, die die oben angegebenen Varianten als Spezialfälle enthält:

$$\rho(C_i, C_j \cup C_k) = \alpha_1\rho(C_i, C_j) + \alpha_2\rho(C_i, C_k) + \beta\rho(C_j, C_k) + \gamma|\rho(C_i, C_j) - \rho(C_i, C_k)| \tag{8.1}$$

3. Illustration mit Berufsstrukturdaten. Zur Illustration der SAHN-Algorithmen verwenden wir die Berufsstrukturdaten aus Abschnitt 2.3 (§ 5). Es gibt $n = 8$ Objekte (Länder); verwendet wird die Abstandsmatrix in Tabelle 2.3-3. Die praktische Durchführung erfolgt mit der `hcls`-Prozedur des TDA-Programms; die folgende Illustration verwendet zunächst die Single-Link-Methode.[4]

Das in Box 8.1-2 dokumentierte Ausgabefile `hca1.lev` zeigt den Ablauf des agglomerativen Verfahrens. In einem ersten Schritt werden die Objekte

[3] In den Bezeichnungen folgen wir Jain und Dubes (1988: 80); man vgl. auch Anderberg (1973: Kap. 6) und Späth (1975: 170-2).

[4] Verwendet wird das Skript `hca1.cf`.

Box 8.1-2 Durch `hca1.cf` erzeugtes Ausgabefile `hca1.lev`.

```
Level  Niveau   Ci   Cj
------------------------
 1     0.0520    3    5
 2     0.1002    4    7
 3     0.1027    3    4
 4     0.1341    3    8
 5     0.1551    1    2
 6     0.1652    1    3
 7     0.1664    1    6
```

Box 8.1-3 Durch `hca1.cf` erzeugtes Ausgabefile `hca1.den`, das das Dendrogramm in Form einer Kantenliste zeigt.

```
  I    J     Niveau
--------------------
  3    9     0.0520
  5    9     0.0520
  4   10     0.1002
  7   10     0.1002
  9   11     0.0507
 10   11     0.0025
  8   12     0.1341
 11   12     0.0314
  1   13     0.1551
  2   13     0.1551
 12   14     0.0311
 13   14     0.0101
  6   15     0.1664
 14   15     0.0012
```

Nr. 3 und 5 bei einem Niveau 0.052 (das ist der Abstand der beiden Objekte in der Abstandsmatrix) zu einem Cluster zusammengefasst:

$$\rho(C_3, C_5) = 0.0520$$

Dann werden die Objekte Nr. 4 und 7 bei einem Niveau 0.1002 zusammengefasst:

$$\rho(C_4, C_7) = 0.1002$$

Dann werden diese beiden Cluster zusammengefasst:

$$\rho(C_3 \cup C_5, C_4 \cup C_7) = \min\{\rho(C_3 \cup C_5, C_4), \rho(C_3 \cup C_5, C_7) = 0.1027$$

Und so weiter, bis alle Objekte in einem Cluster enthalten sind. (Cluster, die mehr als ein Objekt enthalten, werden durch die jeweils kleinste Nummer eines Elements identifiziert.)

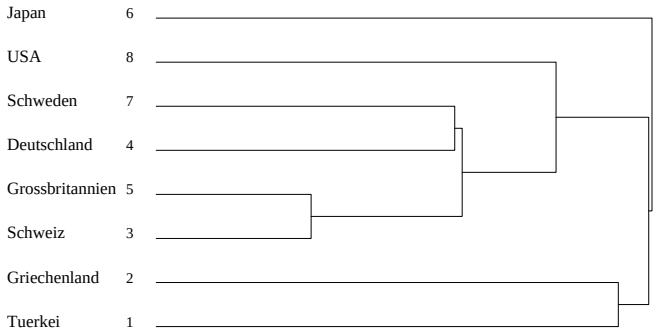


Abb. 8.1-1 Das mit dem Ausgabefile `hca1.pcf` erzeugte Dendrogramm für die hierarchische Klassifikation.

4. Dendrogramme. Ein praktisches Hilfsmittel, um das Ergebnis einer agglomerativen Klassifikation zu veranschaulichen, sind Dendrogramme. Es handelt sich um baumartige Graphen, die zeigen, wie die Objekte bzw. Cluster sukzessive zusammengefasst werden.

Abbildung 8.1-1 zeigt das Dendrogramm für unser gegenwärtiges Beispiel. Man erkennt, wie der agglomerative Prozess verläuft. Man könnte auch eine X-Achse hinzufügen, die die Niveaus anzeigt, bei denen die Zusammenfassung der Objekte bzw. Cluster erfolgt.[5] Als ein Graph betrachtet, kann ein Dendrogramm auch als eine Kantenliste dargestellt (und dann ggf. für weitere Analysezwecke verwendet) werden. Für unser Beispiel erhält man diese Darstellung durch das Ausgabefile `hca1.den` (Box 8.1-3).

5. Bindungen. Von *Bindungen* spricht man, wenn es in einer Abstandsmatrix zwei oder mehr unterschiedliche Paare von Objekten gibt, die den gleichen Abstand haben. Wenn es Bindungen gibt, hängen die Ergebnisse der SAHN-Algorithmen (mit Ausnahme der Single Link-Methode) auch von der zufälligen Ordnung der Daten ab.[6]

6. Vergleiche der SAHN-Verfahren. Auch wenn eine Abstandsmatrix keine Bindungen aufweist, führen die SAHN-Algorithmen meistens zu mehr oder weniger unterschiedlichen Ergebnissen.[7] Abbildung 8.1-2 illustriert das anhand der Berufsstrukturdaten. Ein theoretischer Vergleich setzt allerdings voraus, die Dendrogramme zunächst als unterschiedliche Modelle

[5] Das Wort 'Niveau' bezeichnet hier also den Abstand $\rho(C, C')$ für die Zusammenfassung der Cluster C und C' zu einem neuen Cluster $C \cup C'$.

[6] Man vgl. die Diskussion bei Jain und Dubes (1988: Kap. 4); Klemm (1995).

[7] Ausführliche Hinweise auf Eigenschaften der unterschiedlichen SAHN-Methoden findet man bei Sneath und Sokal (1973); Klemm (1995).

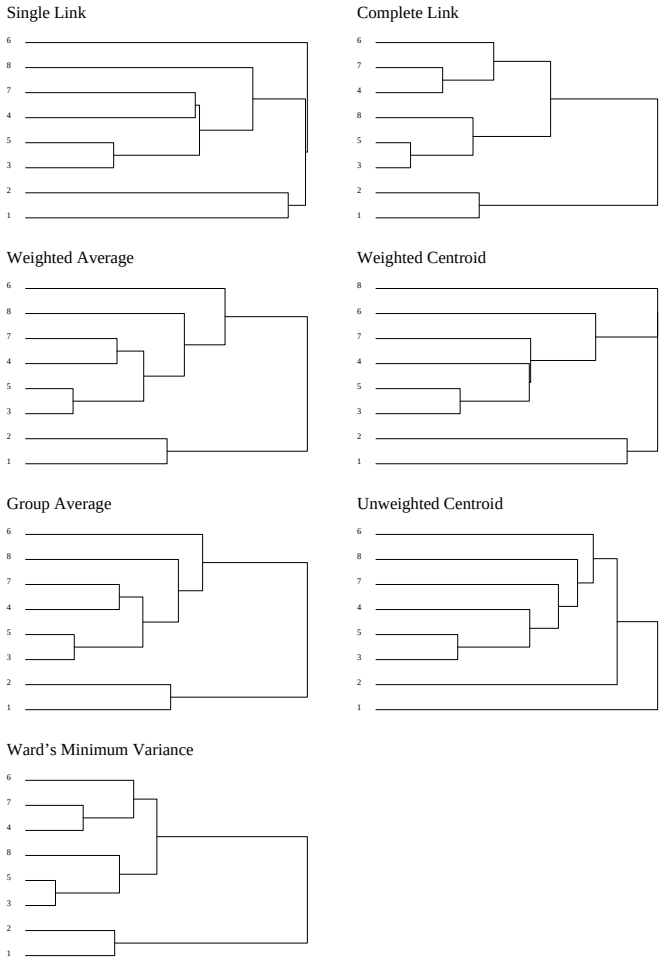


Abb. 8.1-2 Unterschiedliche Ergebnisse der SAHN-Algorithmen bei den Berufsstrukturdaten.

der gegebenen Abstandsmatrix zu verstehen.

Zu beachten ist auch, dass durch die SAHN-Algorithmen nicht immer ein (korrektes) Dendrogramm entsteht. Dies hängt sowohl von der Beschaffenheit der jeweils verwendeten Abstandsmatrix als auch vom Klassifikationsverfahren ab. Milligan (1979) hat gezeigt, dass bei den durch die Formel (8.1) von Lance und Williams definierten Verfahren die Monotoniebedingung dann erfüllt ist, wenn folgende Bedingungen gelten:[8]

$$\gamma \geq 0,\ \alpha_1 + \alpha_2 + \beta \geq 1,\ \alpha_1 \geq 0,\ \alpha_2 \geq 0 \quad \text{oder} \tag{8.2}$$
$$\gamma < 0,\ \alpha_1 + \alpha_2 + \beta \geq 1,\ \alpha_1 \geq |\gamma|,\ \alpha_2 \geq |\gamma|$$

In fünf der sieben in Tabelle 8.1-1 angegebenen Verfahren ist die Bedingung erfüllt. Bei den Centroid-Verfahren (WPGMC und UPGMC) ist sie nicht erfüllt. Tatsächlich tritt auch bei den Berufsstrukturdaten bei der WPGMC-Methode im letzten Schritt eine Verlezung der Monotoniebedingung auf.

7. Monotone Abstandstransformationen.

[8] Vgl. auch Jain und Dubes (1988: 83ff.).

8.2 Ultrametrische Baummodelle

In diesem Abschnitt betrachten wir hierarchische Klassifikationsverfahren als Methoden zur Konstruktion von Modellen für Abstandsfunktionen. Entsprechend der in Abschnitt **??** (§ ??) besprochenen Unterscheidung handelt es sich um repräsentierende, nicht um erklärende Modelle.

1. Hierarchien und Bäume. Sei $\Omega = \{\omega_1, \ldots, \omega_n\}$ eine beliebige Objektmenge. Unter einer *Hierarchie* (für Ω) verstehen wir eine Menge $\mathcal{H}$, deren Elemente Teilmengen von Ω sind und die folgenden Bedingungen genügt:

a) $\Omega \in \mathcal{H}$, $\emptyset \notin \mathcal{H}$

b) für alle $\omega \in \Omega$: $\{\omega\} \in \mathcal{H}$

c) wenn $h_1, h_2 \in \mathcal{H}$, dann $h_1 \cap h_2 \in \{h_1, h_2, \emptyset\}$

Ist beispielssweise $\Omega = \{1, 2, 3, 4, 5\}$, wäre

$$\mathcal{H} := \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{2, 3\}, \{1, 2, 3, 4, 5\}\}$$

eine Hierarchie. Offenbar können auch noch anderen Hierarchien gebildet werden.

Wenn $\mathcal{H}$ eine Hierarchie ist, kann sie auch durch einen Graphen repräsentiert werden. Man kann $\mathcal{H}$ mit der Knotenmenge des Graphen identifizieren und festlegen, dass zwei Knoten h_1 und h_2 durch eine Kante verbunden sind, wenn h_1 eine Teilmenge von h_2 ist und es keinen anderen Knoten h mit der Eigenschaft $h_1 \subset h \subset h_2$ gibt.

Graphen, durch die Hierarchien repräsentiert werden, sind Beispiele für eine besondere Art von Graphen, die man *Bäume* nennt. Wir verwenden folgende allgemeine Definition: Ein Graph $\mathcal{G} = (\mathcal{K}, \mathcal{L})$, wobei $\mathcal{K}$ die Knotenmenge und $\mathcal{L}$ die Kantenmenge bezeichnet, ist ein Baum, wenn $|\mathcal{K}| = |\mathcal{L}| + 1$ ist und je zwei Knoten durch genau einen Pfad verbunden sind.[9] Knoten, die den Grad 1 haben, werden als *Blätter* bezeichnet; wenn der Grad größer als 1 ist, spricht man von *internen Knoten*.

Man kann für einen Baum eine Orientierung einführen, indem man einen internen Knoten als *Wurzel des Baums* bestimmt. Dann gibt es von jedem Blatt des Baums genau einen Pfad zu seiner Wurzel. Zu beachten ist, dass es meistens unterschiedliche Möglichkeiten gibt, Wurzeln zu definieren. So könnte man in dem oben angeführten Beispiel sowohl den Knoten, der der Menge Ω entspricht, als auch den Knoten, der der Menge $\{2, 3\}$ entspricht, als eine Wurzel des Baums bestimmen.

2. Einfache Baummodelle. Sei $\mathcal{G} = (\mathcal{K}, \mathcal{L}, v)$ ein Baum mit einer Bewertungsfunktion $v : \mathcal{L} \longrightarrow \mathbf{R}$. Die Bewertungsfunktion ordnet jeder Kante

[9] Damit äquivalent: Ein zusammenhängender Graph ohne Zyklen.

Box 8.2-1 Kantenliste eines minimalen aufspannenden Baums für die Abstandsmatrix in Tabelle 2.3-3.

i	j	v(i,j)
3	5	0.0520
4	7	0.1002
4	5	0.1027
5	8	0.1341
1	2	0.1551
2	8	0.1652
4	6	0.1664

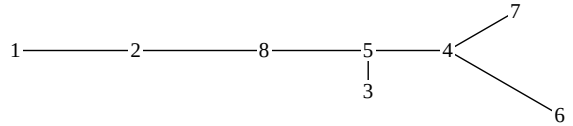


Abb. 8.2-1 Der Baum mit der Kantenliste in Box 8.2-1.

$l \in \mathcal{L}$ eine positive Zahl $v(l)$ zu, die als Länge oder Wert der Kante interpretiert werden kann. Dann kann man für jeweils zwei Knoten $i, j \in \mathcal{K}$ einen Abstand

$$d_{ij}^v := \text{ Summe der Kantenbewertungen des Pfades von } i \text{ nach } j$$

definieren und erhält auf diese Weise eine Abstandsfunktion für die Knotenmenge des Baums.

Diese Möglichkeit, Abstandsfunktionen durch Bäume zu definieren, führt nun zu der Idee, Bäume als Modelle für vorgegebene Abstandsmatrizen zu verwenden. Wir beginnen mit einer einfachen Variante, die später modifiziert und erweitert wird. Als Ausgangspunkt wird angenommen, dass für eine Objektmenge $\Omega = \{\omega_1, \ldots, \omega_n\}$ Abstände gegeben sind; d_{ij} ist der Abstand zwischen ω_i und ω_j. Gesucht ist nun ein Baum $\mathcal{G}$ mit der Knotenmenge Ω und einer Kantenbewertung v, so dass die daraus resultierenden Abstände d_{ij}^v möglichst gut zu den vorausgesetzten Abständen d_{ij} passen.

Wenn man die Optimalitätsforderung („möglichst gut“) zunächst ignoriert, kann man einen sogenannten *minimalen aufspannenden Baum* (auch kurz *Minimalbaum* genannt[10]) verwenden. Um den Begriff und die Konstruktion zu erklären, verwenden wir als Beispiel die Berufsstrukturdaten aus Abschnitt 2.3, und zwar die Abstandsmatrix in Tabelle 2.3-3 für die acht Länder; sie wird im Folgenden $\mathbf{D} = (d_{ij})$ genannt. Gesucht ist also

[10] Im Englischen: *minimal spanning tree*.

Tabelle 8.2-1 Aus dem Baum mit der Kantenliste in Box 8.2-1 gebildete Abstandsmatrix $\mathbf{D}^v$.

0.0000	0.1551	0.5064	0.5571	0.4544	0.7235	0.6573	0.3203
0.1551	0.0000	0.3513	0.4020	0.2993	0.5684	0.5022	0.1652
0.5064	0.3513	0.0000	0.1547	0.0520	0.3211	0.2549	0.1861
0.5571	0.4020	0.1547	0.0000	0.1027	0.1664	0.1002	0.2368
0.4544	0.2993	0.0520	0.1027	0.0000	0.2691	0.2029	0.1341
0.7235	0.5684	0.3211	0.1664	0.2691	0.0000	0.2666	0.4032
0.6573	0.5022	0.2549	0.1002	0.2029	0.2666	0.0000	0.3370
0.3203	0.1652	0.1861	0.2368	0.1341	0.4032	0.3370	0.0000

ein Baum mit der Knotenmenge $\mathcal{K} = \{1, \ldots, 8\}$, so dass die Knoten den Ländern entsprechen. Man benötigt genau sieben Kanten, damit ein Baum entsteht. Welche soll man nehmen?

Bei der Konstruktion eines Minimalbaums wird angenommen, dass die Kantenbewertungen durch die Abstände d_{ij} vorgegeben sind. Wenn also für den zu konstruierenden Baum eine Kante verwendet wird, die die Knoten i und j verbindet, erhält sie die Bewertung d_{ij}. Also kann man sich an folgender Idee orientieren: Wähle die Kanten für den zu konstruierenden Baum so aus, dass die Summe ihre Kantenbewertungen minimal wird. Der auf diese Weise entstehende Baum wird ein Minimalbaum genannt.

Box 8.2-1 zeigt einen Minimalbaum für das Beispiel in Form einer Kantenliste.[11] Die angegebenen Kantenbewertungen entsprechen offenbar den Abständen in Tabelle 2.3-3. Abbildung 8.2-1 zeigt den Baum in einer graphischen Darstellung, wobei die Länge der Kanten näherungsweise proportional zu ihren Bewertungen sein sollte.[12]

Allerdings ist ohne weiteres nicht klar, ob der Baum ein gutes Modell für die Abstandsmatrix $\mathbf{D}$ (in Tabelle 2.3-3) ist. Denn der Baum repräsentiert zwar korrekt sieben Abstände; er impliziert aber (insgesamt 28) Abstände zwischen allen acht Knoten. Wenn man sie ausrechnet, erhält man eine neue Abstandsmatrix $\mathbf{D}^v = (d^v_{ij})$, die in Tabelle 8.2-1 gezeigt wird.[13] Vergleicht man $\mathbf{D}^v$ mit $\mathbf{D}$, findet man, dass es erhebliche Unterschiede gibt, was auch durch folgende Abstandsberechnungen sichtbar wird:[14]

$$\|\mathbf{D}^v - \mathbf{D}\| = 1.04 \quad \|\mathbf{D}^v\| = 2.70 \quad \|\mathbf{D}\| = 1.73$$

3. Hierarchische Klassifikationsschemas. Bei einfachen Baummodellen werden nur die Elemente einer vorgegebenen Objektmenge Ω als Knoten

[11] Erzeugt mit dem Skript `gm1.cf`.

[12] Erstellt mit dem Skript `gmplot1.cf`.

[13] Berechnet mit dem Skript `gm1a.cf`.

[14] Berechnet mit dem Skript `gm1b.cf`.

verwendet. Allgemeinere Baummodelle entstehen, wenn man Bäume verwendet, die darüber hinaus noch andere Knoten enthalten. Bei einem oft verwendeten Ansatz wird ein Baum verwendet, dessen Blätter den vorgegebenen Objekten entsprechen; dann werden interne Knoten hinzugefügt, so dass man eine möglichst gute Repräsentation der für die Objekte gegebenen Abstandsmatrix $\mathbf{D}$ durch die durch den Baum induzierte Abstandsmatrix $\mathbf{D}^v$ erreicht.[15] Einen Spezialfall bilden hierarchische Klassifikationsschemas (HKS), mit denen wir uns im Folgenden beschäftigen.[16]

Wie bisher nehmen wir an, dass eine Objektmenge $\Omega = \{\omega_1, \ldots, \omega_n\}$ mit einer Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist. Unter einem *hierarchischen Klassifikationsschema* (HKS) für Ω verstehen wir eine Hierarchie, für deren Elemente eine Indexfunktion definiert ist. Für die Hierarchie verwenden wir die Notation

$$\mathcal{H} := \{C_1, \ldots, C_n, \ldots, C_q\}$$

Die Elemente sind Teilmengen von Ω, und wir nehmen an, dass $C_i = \{\omega_i\}$ für $i = 1, \ldots, n$ und $C_q = \{\Omega\}$ ist. Außerdem gibt es für jedes Element $C \in \mathcal{H}$ einen Wert $\sigma(C) \geq 0$, wobei gilt, dass $\sigma(C_i) = 0$ für $i = 1, \ldots, n$. Die Funktion σ wird als *Indexfunktion* (der Hierarchie) bezeichnet.[17]

Wie diese Indexwerte entstehen, hängt vom Konstruktionsverfahren ab. Davon hängt auch ab, ob eine Indexfunktion entsteht, die folgende Monotoniebedingung erfüllt:

$$C \subset C' \implies \sigma(C) < \sigma(C')$$

Sie ist zum Beispiel erfüllt, wenn ein HKS mit der Single-Link-Methode konstruiert wird, bei einigen anderen Verfahren können auch nichtmonotone Indexfunktionen entstehen. Wir nehmen für die weiteren Überlegungen an, dass die Monotoniebedingung erfüllt ist.

Ist nun ein HKS $\mathcal{H}$ mit einer Indexfunktion σ gegeben, kann ein Baum mit der Knotenmenge $\mathcal{H}$ gebildet werden, dessen Blätter den Elementen $C_1, \ldots, C_n$ entsprechen und dessen Wurzel dem Element C_q entspricht (vgl. § 1). Mit der Indexfunktion kann auch eine Abstandsfunktion für den Baum definiert werden. Zunächst wird eine Bewertungsfunktion v für die Kanten des Baums festgelegt. Wenn eine Kante l die Knoten C_i und C_j verbindet, kann man annehmen, dass C_j auf einem Pfad liegt, der von C_i zur Wurzel C_q führt; also liefert die Definition

$$v(l) := \sigma(C_j) - \sigma(C_i)$$

[15] Eine gute Darstellung dieses Ansatzes geben Barthélemy und Guénoche (1991).

[16] Die erste ausführliche Untersuchung stammt von Johnson (1967). Eine gute neuere Darstellung findet man bei Jain und Dubes (1988: 58ff.).

[17] Im Englischen: *indexed hierarchy*.

Tabelle 8.2-2 Aus dem mit der Single Link-Methode erzeugten Dendrogramm gewonnene Abstandsmatrix D^h für die acht Länder.

0.0000	0.1551	0.1652	0.1652	0.1652	0.1664	0.1652	0.1652
0.1551	0.0000	0.1652	0.1652	0.1652	0.1664	0.1652	0.1652
0.1652	0.1652	0.0000	0.1027	0.0520	0.1664	0.1027	0.1341
0.1652	0.1652	0.1027	0.0000	0.1027	0.1664	0.1002	0.1341
0.1652	0.1652	0.0520	0.1027	0.0000	0.1664	0.1027	0.1341
0.1664	0.1664	0.1664	0.1664	0.1664	0.0000	0.1664	0.1664
0.1652	0.1652	0.1027	0.1002	0.1027	0.1664	0.0000	0.1341
0.1652	0.1652	0.1341	0.1341	0.1341	0.1664	0.1341	0.0000

eine Bewertungsfunktion, die jeder Kante eine positive Bewertung zuordnet. Wie in § 2 besprochen wurde, erhält man dann auch eine Abstandsfunktion, die je zwei Knoten C_i und C_j einen Abstand $d^v(C_i, C_j)$ zuordnet, der die Summe der Bewertungen der Kanten auf dem Pfad von C_i nach C_j angibt.

4. Darstellung durch Dendrogramme. Der Baum, der durch ein HKS entsteht, hat eine besondere Eigenschaft: Alle Blätter haben von der Wurzel den gleichen Abstand. Ein solcher Baum wird auch als ein *Dendrogramm* bezeichnet. Ein Dendrogramm kann auch dadurch charakterisiert werden, dass die durch seine Levelfunktion definierte Abstandsfunktion d^v folgende Bedingung erfüllt:

$$\text{Für alle } i,j,k\text{: } d^v_{ij} \leq \max\{d^v_{ik}, d^v_{jk}\} \tag{8.3}$$

Eine Abstandsfunktion, die diese Bedingung erfüllt, wird *ultrametrisch* genannt (sie impliziert die Dreiecksungleichung).

Als Beispiel kann man an die Konstruktion eines HKS für die Berufsstrukturdaten in Abschnitt 8.1 (§ 3) denken. Box 8.1-3 zeigt die Indexfunktion, Abbildung 8.1-1 zeigt das Dendrogramm. In diesem Beispiel hat der Baum $q = 15$ Knoten; die Knoten $C_1, \ldots, C_8$ entsprechen den Ländern, C_{15} ist die Gesamtmenge der Länder.

5. Alternative Abstandsberechnung. Wenn das Baummodell durch ein Dendrogramm gegeben ist, kann alternativ zur Abstandsfunktion d^v_{ij} (Summe der Kantenbewertungen des Pfads von i nach j) auch eine Abstandsfunktion

$$d^h_{ij} := \text{ Index des kleinsten Clusters, das } i \text{ und } j \text{ enthält}$$

gebildet werden.[18] Schränkt man die Abstandsfunktion auf die ursprünglichen Objekte (die Blätter des Baums) ein, gilt offenbar: $d^v_{ij} = 2\, d^h_{ij}$. Tabelle

[18] Diese Abstandsdefinition wurde bereits von Johnson (1967: 244) vorgeschlagen, und sie wird auch meistens von Computerprogrammen für hierarchische Clusteranalysen verwendet; dies gilt auch für die hier verwendete TDA-Prozedur `hcls`. Vgl. auch die Hinweise bei Corter (1996: 15f.).

Tabelle 8.2-3 Vergleich der unterschiedlichen SAHN-Verfahren für die Berufsstrukturdaten.

Methode	$\|\mathbf{D} - \mathbf{D}^h\|$
Single Link	0.7580
Complete Link	0.8901
Weighted Average	0.4579
Weighted Centroid	0.7260
Group Average	0.4380
Unweighted Centroid	0.7059
Ward's Minimum Variance	1.4246

8.2-2 zeigt die so eingeschränkte Abstandsmatrix $\mathbf{D}^h = (d^h_{ij})$ für das Beispiel. Man erkennt, dass es wiederum nur sieben $(n-1)$ unterschiedliche Abstandswerte gibt.

Analog zur Vorgehensweise in § 2 können auch diese Abstandsmatrizen mit der vorgegebenen Abstandsmatrix $\mathbf{D}$ verglichen werden, um den Grad ihrer Repräsentation durch das hierarchische Klassifikationsschema zu erfassen. Tabelle 8.2-3 vergleicht die unterschiedlichen SAHN-Verfahren für die Berufsstrukturdaten.

6. Optimale ultrametrische Modelle. Hierarchische Klassifikationsverfahren liefern eine Abstandsmatrix $\mathbf{D}^h$, die dann durch $\|\mathbf{D} - \mathbf{D}^h\|$ mit der vorgegebenen Abstandsmatrix $\mathbf{D}$ verglichen werden kann. Offenbar gelangt man zu einem im Sinne dieses Kriteriums optimalen Modell, wenn man folgende Aufgabe löst: Finde eine ultrametrische Abstandsmatrix $\mathbf{D}^u$, die den Wert des Kriteriums $\|\mathbf{D} - \mathbf{D}^u\|$ minimal macht.

Für die Lösung dieser Aufgabe sind verschiedene Algorithmen vorgeschlagen worden.[19] Wir verwenden zur Illustration wiederum die Berufsstrukturdaten. Tabelle 8.2-4 zeigt für dieses Beispiel die im Sinne des LS-Kriterums optimale ultrametrische Abstandsmatrix $\mathbf{D}^u$. Das Kriterium hat in diesem Fall den Wert $\|\mathbf{D} - \mathbf{D}^u\| = 0.39$, offenbar niedriger als die in Tabelle 8.2-3 angegebenen Werte, die mit den SAHN-Algorithmen erzielt werden konnten.

Da die Abstandsmatrix $\mathbf{D}^u$ ultrametrisch ist, kann sie auch durch ein Dendrogramm dargestellt werden. Das Dendrogramm kann beispielsweise dadurch erzeugt werden, dass man $\mathbf{D}^u$ als Ausgangsmatrix für ein hierarchisches Klassifikationsverfahren mit der Single-Link-Methode verwendet. Abbildung 8.2-2 zeigt das Ergebnis für das Beispiel.

[19] Vgl. De Soete (1984, 1988); De Soete, Carroll und De Sarbo, W. S. (1987); Sriram und Lewis (1993); Corter (1996: 26ff.).

Tabelle 8.2-4 Optimale ultrametrische Abstandsmatrix $\mathbf{D}^u$ für die Berufsstrukturdaten.

0.0000	0.3282	0.3282	0.3282	0.3282	0.3282	0.3282	0.3282
0.3282	0.0000	0.2184	0.2184	0.2184	0.2184	0.2184	0.2184
0.3282	0.2184	0.0000	0.1371	0.0517	0.1975	0.1371	0.1815
0.3282	0.2184	0.1371	0.0000	0.1371	0.1975	0.1004	0.1815
0.3282	0.2184	0.0517	0.1371	0.0000	0.1975	0.1371	0.1815
0.3282	0.2184	0.1975	0.1975	0.1975	0.0000	0.1975	0.1975
0.3282	0.2184	0.1371	0.1004	0.1371	0.1975	0.0000	0.1815
0.3282	0.2184	0.1815	0.1815	0.1815	0.1975	0.1815	0.0000

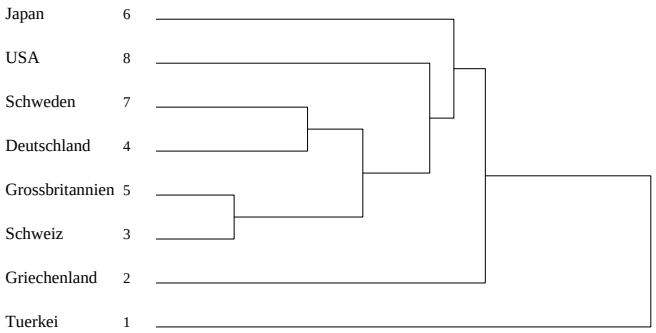


Abb. 8.2-2 Das der ultrametrischen Abstandsmatrix $\mathbf{D}^u$ in Tabelle 8.2-4 entsprechende Dendrogramm.

Kapitel 9

Bildung von Partitionen

9.1 Partitionen aus Hierarchien
1. Unterschiedliche Ansätze.
2. Rechentechnische Probleme.

9.2 Verwendung von Clusterzentren
1. Unterschiedliche Ansätze.
2. Rechentechnische Probleme.

Die im vorangegangenen Kapitel besprochenen hiearchischen Klassifikationsverfahren führen nicht unmittelbar zu Partitionen einer Objektmenge, sondern liefern Repräsentationen der für die Objekte gegebenen Abstandsfunktion. Bevor wir uns im nächsten Kapitel mit weiteren Möglichkeiten einer Konstruktion repräsentierender Modelle beschäftigen, soll in diesem Kapitel besprochen werden, wie Partitionen gebildet werden können.

Eine Möglichkeit besteht darin, Partitionen aus einem hierarchischen Klassifikationsschema zu bilden, indem man mehr oder weniger willkürlich einen bestimmten Indexwert festlegt, der das Dendrogramm in Cluster zerlegt. Das wird in Abschnitt 9.1 illustriert. In den restlichen Abschnitten werden einige partitionierende Klassifikationsverfahren besprochen. Bei diesen Verfahren muss die Anzahl der zu bildenden Cluster, im Folgenden k genannt, vorgegeben werden. Gesucht ist dann eine Partitionierung der Objektmenge in k disjunkte Teilmengen $C_1, \ldots, C_k$, so dass es sich um gut definierte Cluster handelt. Dafür können unterschiedliche Kriterien verwendet werden. Man kann in erster Linie zwei Ansätze unterscheiden. Einerseits Ansätze, die für jedes Cluster ein Clusterzentrum suchen und dann alle Objekte ihrem nächstgelegenen Clusterzentrum zuordnen; sie werden in Abschnitt 9.2 besprochen. Andererseits Ansätze, die ohne Clusterzentren auskommen; sie werden in Abschnitt **??** besprochen. Schließlich werden in Abschnitt **??** einige Möglichkeiten dargestellt, um Partitionen zu vergleichen und zu beurteilen.

9.1 Partitionen aus Hierarchien

9.2 Verwendung von Clusterzentren

1. Ansätze mit Clusterzentren.

a) Ein erstes Kriterium setzt voraus, dass die Objekte durch Zeilen (oder Spalten) einer Datenmatrix gegeben sind:

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{im})' \qquad (i = 1, \ldots, n)$$

und dass euklidische Abstände verwendet werden. Dann kann für jedes Cluster C_l ein Mittelwert

$$\bar{\mathbf{x}}_l := \frac{1}{n_l} \sum\nolimits_{i \in C_l} \mathbf{x}_i$$

definiert werden (n_l bezeichnet die Anzahl der Elemente in C_l), und es wird möglich, folgendes Kriterium zu verwenden:

$$\sum_{l=1}^{k} \sum_{i \in C_l} \| \mathbf{x}_i - \bar{\mathbf{x}}_l \|^2 \longrightarrow \min \tag{9.1}$$

Es wird auch als *Abstandsquadratsummenkriterium* bezeichnet. Jeder Vektor $\mathbf{x}_i$ wird dem nächstgelegenen Clusterzentrum $\bar{\mathbf{x}}_l$ zugeordnet, und die Clusterzentren sollen so bestimmt werden, dass die gesamten quadrierten Abstände zu den Clusterzentren möglichst klein werden.

b) Wenn zunächst eine beliebige Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist, kann das Abstandsquadratsummenkriterium nicht verwendet werden. Objekte als Zentren.

2. Ansätze ohne Clusterzentren. Die beiden in § 1 besprochenen Kriterien verwenden Clusterzentren. Entweder oder. Man kann auch versuchen, ohne Clusterzentren auszukommen.

a) Man kann beispielsweise folgendes Kriterium betrachten:

$$\sum_{l=1}^{k} \frac{1}{n_l} \sum_{i,j \in C_l : j<i} d_{ij} \longrightarrow \min \tag{9.2}$$

Das Ziel besteht in diesem Fall darin, ...

Unterschiedliche Varianten des Kriteriums entstehen, wenn man nicht durch n_j, sondern beispielsweise durch $n_j(n_j - 1)$ dividiert; H. Späth, der mit unterschiedlichen Varianten ausführlich experimentiert hat, hält die in (9.2) angegebene Variante für die praktisch brauchbarste.[1]

[1] Vgl. Späth (1983: 92ff.; 1988).

b) Anstatt sich auf eine Art von durchschnittlicher Abstandsgröße in den Clustern zu beziehen, wie bei den Kriterien der Art (9.2), kann man auch den durch $\max\{d_{ij} \mid i, j \in C_l\}$ definierten Clusterdurchmesser verwenden. Dann entsteht folgendes Kriterium:

$$\sum_{l=1}^{k} \max\{d_{ij} \mid i, j \in C_l\} \longrightarrow \min \tag{9.3}$$

In diesem Fall sollen also die Cluster so gebildet werden, dass die Summe ihrer Durchmesser minimal wird.

3. Rechentechnische Probleme. Die Hauptschwierigkeit resultiert daraus, dass die Anzahl der Möglichkeiten zur Einteilung einer Menge von N Objekten in k Cluster schnell außerordentlich groß wird. Zum Beispiel kann man 10 Objekte auf 34105 verschiedene Weisen in vier Cluster einteilen; aber bei 19 Objekten gibt es bereits 11,259,666,000 Möglichkeiten.[2] Es ist deshalb in den meisten Fällen praktisch nicht möglich, Cluster zu finden, die den globalen Minima der oben angegebenen Kriterien entsprechen.

Die normalerweise verwendeten Verfahren können nur lokale Minima der Kriterien finden. Oft handelt es sich um sog. Austauschverfahren. D.h. ausgehend von einer (irgendwie, zufällig) gebildeten Anfangspartition werden solange Objekte zwischen den Partitionen ausgetauscht, bis sich das Kriterium nicht weiter verkleinern lässt. Für das Kriterium (9.1) wird eine besonders oft verwendete Variante dieses Austauschverfahrens als *k-means Algorithmus* bezeichnet.[3] Für das Kriterium (9.2) wurde ein Austauschverfahren von H. Späth entwickelt.[4] Teilweise andere Verfahren wurden für das Kriterium (9.3) vorgeschlagen.[5]

Bei der Verwendung partitonierender Verfahren sollte man also daran denken, dass die normalerweise verfügbaren Programme nur lokale Minima der Kriterien finden können. Es ist deshalb sinnvoll, die Berechnungen mit unterschiedlichen Anfangspartitionen zu wiederholen, um einen gewissen Einblick in das Auftreten unterschiedlicher lokaler Minima zu gewinnen.

4. Illustration mit artifiziellen Daten. Zur Illustration partitionierender Verfahren verwenden wir das Kriterium (9.2). Wir beginnen mit den in Abschnitt 7.1 (§ 5) beschriebenen artifiziellen Daten: 100 Realisierungen einer zweidimensionalen Normalverteilung (Abbildung 7.1-2). Daraus wird eine $(100, 100)$-Matrix mit euklidischen Abständen gebildet und als Eingabe für

[2] Vgl. Jain und Dubes (1988: 91).

[3] Vgl. Hartigan (1975: Kap. 4); Bacher (1994: 308ff.).

[4] Vgl. Späth (1983: 143ff.). Dieses Verfahren liegt auch der TDA-Prozedur `clp` zugrunde, die für die späteren Illustrationen verwendet wird.

[5] Vgl. Hansen und Jaumard (1987); Charikar und Panigrahy (2001).

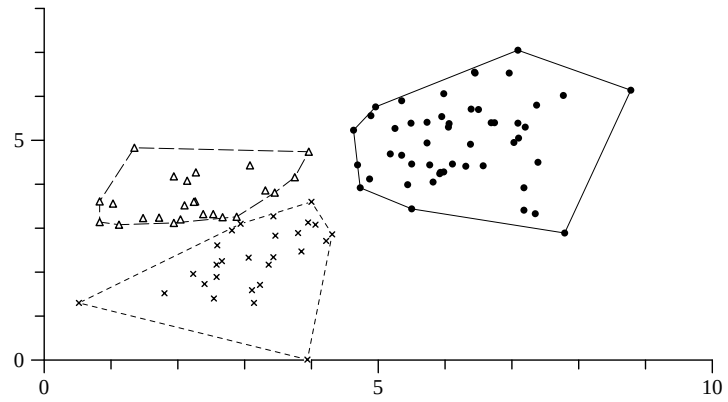


Abb. 9.2-1 Einteilung der 100 Punkte aus Abbildung 7.1-2 in drei Cluster unter Verwendung des Kriteriums (9.2). Die Kreise deuten die Clustermittelpunkte an.

ein partitionierendes Verfahren verwendet.[6]

Versucht man, durch Minimierung des Kriteriums (9.2) zwei Cluster zu bilden, erhält man bei 100 Wiederholungen mit zufällig gebildeten Anfangspartitionen als beste Lösung folgende Einteilung: Die Punkte 1–51 (ohne Nr. 32) gehören zum ersten, die Punkte 32 und 52–100 zum zweiten Cluster.[7] Bis auf den Punkte Nr. 32 (der mit den Koordinaten $(3.96, 4.74)$ einen Grenzfall bildet) entspricht dies Ergebnis dem datenerzeugenden Prozess.

Aber warum zwei Cluster? Bildet man drei Cluster, entsteht sogleich ein vollständig anderes Bild, s. Abbildung 9.2-1.[8] In diesem Fall wird auch bei 100 Wiederholungen ein optimales Ergebnis nur in acht Fällen erreicht.

5. Beispiele mit Berufsstrukturdaten. Jetzt verwenden wir die Berufsstrukturdaten aus Abschnitt 2.3. Wir beginnen mit der Abstandsmatrix in Tabelle 2.3-3 für die acht Länder. Folgende Tabelle zeigt das Ergebnis, wenn man zwei, drei bzw. vier Cluster bildet:[9]

Land	1	2	3	4	5	6	7	8
2 Cluster	1	1	2	2	2	2	2	2
3 Cluster	1	1	3	2	3	2	2	3
4 Cluster	1	1	2	4	2	3	4	2

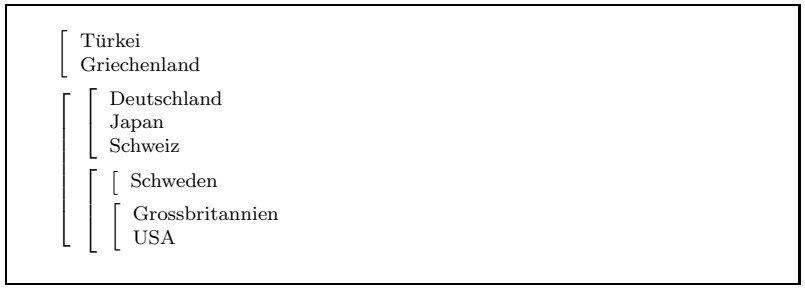
[6] Das Datenfile mit der Abstandsmatrix wurde mit dem Skript `cl1a.cf` erzeugt und `cl1a.dat` genannt.

[7] Verwendet wurde das Skript `clp1.cf`.

[8] Verwendet wurde das Skript `clp1a.cf`.

[9] Verwendet wurde das Skript `clp2.cf`.

Box 9.2-1 Verwendung der Berufsstrukturdaten aus Tabelle 2.3-3 für eine Einteilung der Länder in zwei, drei bzw. vier Cluster.



In diesem Beispiel entsteht auch eine hierarchische Struktur, wie sie in Box 9.2-1 noch einmal verdeutlicht wird. (Das ist bei partitionierenden Verfahren im Allgemeinen nicht der Fall.)

Verwendet man die Abstandsmatrix für die geschlechtsspezifischen Berufsverteilungen aus Abschnitt 2.3 (§ 6), findet man wiederum eine hierarchische Struktur:[10]

$$\Big(\mathrm{M1},\ldots,\mathrm{M8}\Big)\ \Big(\big(\mathrm{F1},\mathrm{F2},\mathrm{F8}\big)\big((\mathrm{F3},\mathrm{F4},\mathrm{F5},\mathrm{F7})(\mathrm{F6})\big)\Big)$$

[10] Erzeugt mit dem Skript `clp3.cf`.

Kapitel 10

Unscharfe Klassifikation

10.1 Pyramidale Klassifikation

1. Pyramidale Klassifikationsschemas.
2. Ein agglomerativer Algorithmus.
3. Illustration der Berechnung.
4. Indexfunktionen und Abstände.
5. Berufsstrukturdaten.

10.1 Pyramidale Klassifikation

1. Pyramidale Klassifikationsschemas. Pyramidale Klassifikation wurde von E. Diday vorgeschlagen und dann von zahlreichen Autoren weiter verfolgt.[1] Es handelt sich um eine Verallgemeinerung der hierarchischen Klassifikation. Eine formale Definition kann folgendermaßen gegeben werden. Ausgangspunkt ist eine Objektmenge $\Omega = \{\omega_1, \ldots, \omega_n\}$. Eine Menge von Teilmengen von Ω, im Folgenden $\mathcal{H}$ genannt, ist ein *pyramidales Klassifikationsschema* (kurz: eine *Pyramide*) für Ω, wenn folgende Bedingungen erfüllt sind:

a) $\Omega \in \mathcal{H}$ und $\{\omega_i\} \in \mathcal{H}$ für $i = 1, \ldots, n$.

b) Für alle $h, h' \in \mathcal{H}$: $h \cap h' = \emptyset$ oder $h \cap h' \in \mathcal{H}$.

c) Es gibt eine vollständige lineare Ordnung für Ω, so dass alle Elemente von $\mathcal{H}$ bzgl. dieser Ordnung zusammenhängend sind.

Offenbar ist jedes hierarchische Klassifikationsschema auch ein pyramidales Klassifikationsschema. Ein Unterschied liegt darin, dass jedes Element einer Pyramide zwei unmittelbare Nachfolger haben kann.[2] Zum Beispiel: $\Omega = \{\omega_1, \omega_2, \omega_3\}$ und

$$\mathcal{H} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_1, \omega_2\}, \{\omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_3\}\}$$

In diesem Beispiel hat $\{\omega_2\}$ zwei unmittelbare Nachfolger. Mehr als zwei unmittelbare Nachfolger sind jedoch nicht möglich.[3] Daraus folgt, dass ein pyramidales Klassifikationsschema für n Objekte maximal $n\,(n+1)/2$ Elemente hat.

2. Ein agglomerativer Algorithmus. Zur praktischen Berechnung von pyramidalen Klassifikationsschemas verwenden wir eine Variante eines agglomerativen Algorithmus, der von Diday (1986: 215) vorgeschlagen wurde. Ausgangspunkt ist eine Objektmenge $\Omega = \{\omega_1, \ldots, \omega_n\}$ und eine Distanzmatrix $\mathbf{D} = (d_{ij})$.

1) Setze $\mathcal{H} := \{\{\omega_1\}, \ldots, \{\omega_n\}\}$.
2) Bilde eine Menge $\mathcal{M}$, die aus allen Paaren (h_1, h_2) von Elementen von $\mathcal{H}$ besteht, die folgenden Bedingungen genügen:
   a) Wenn $h_1 = \{\omega\}$ ist, kann ω bzgl. der (während des Verfahrens zu bildenden) linearen Ordnung unmittelbar vor dem kleinsten Element oder unmittelbar nach dem größten Element von h_2 plaziert werden.
   b) Wenn $h_2 = \{\omega\}$ ist, kann ω bzgl. der (während des Verfahrens zu bildenden) linearen Ordnung unmittelbar vor dem kleinsten Element oder unmittelbar nach dem größten Element von h_1 plaziert werden.
   c) $h_1 \cup h_2$ ist bzgl. der bisher gebildeten linearen Ordnung zusammenhängend.
   d) Es gibt kein $h \in \mathcal{H}$, so dass $h_1 \subseteq h$ und $h_2 \subseteq h$ ist.
   e) Es gibt kein $h \in \mathcal{H}$, so dass $h \subset h_1 \cup h_2$ und $h \not\subseteq h_1$ und $h \not\subseteq h_2$.

   Wenn kein Paar gefunden werden kann, dass diesen Bedingungen genügt, wird abgebrochen.
3) Wähle das Paar, bei dem h_1 und h_2 den kleinsten Abstand aufweisen und füge $h_1 \cup h_2$ als neues Element zur bisher gebildeten Menge $\mathcal{H}$ hinzu. Soweit erforderlich, ergänze die lineare Ordnung für die Elemente von Ω. Dann Fortsetzung bei (2).

Ergänzend ist festzulegen, wie Abstände zwischen den Elementen von $\mathcal{H}$ gebildet werden sollen. Das kann auf analoge Weise geschehen, wie in Abschnitt 8.1 (§ 2) für hierarchische Klassifikationsverfahren besprochen wurde. Es wird also durch den Algorithmus zugleich eine neue Abstandsfunktion $d_{\mathcal{H}}$ für $\mathcal{H}$ gebildet. Anfangs wird mit der Definition

$$d_{\mathcal{H}}(\{\omega_i\}, \{\omega_j\}) := d_{ij}$$

begonnen. Dann kann für die Fortsetzung beispielsweise die Complete-Link-Methode verwendet werden, also die Definition

$$d_{\mathcal{H}}(h, h_1 \cup h_2)) := \max\{d_{\mathcal{H}}(h, h_1), d_{\mathcal{H}}(h, h_2)\}$$

Diese Methode hat auch den Vorteil, dass eine Pyramide ohne Inversionen entsteht.[4]

3. Illustration der Berechnung. Um die Berechnung eines pyramidalen Klassifikationsschemas zu verdeutlichen, verwenden wir vier Objekte mit folgender Abstandsmatrix:

$$\mathbf{D} := \begin{pmatrix} 0 & 1 & 3 & 2 \\ 1 & 0 & 6 & 5 \\ 3 & 6 & 0 & 4 \\ 2 & 5 & 4 & 0 \end{pmatrix} \tag{10.1}$$

Box 10.1-1 zeigt das Ergebnis des in § 2 beschriebenen Algorithmus mit der Complete-Link-Methode.[5] Das pyramidale Klassifikationsschema $\mathcal{H}$ hat in diesem Beispiel 10 Elemente $h_1, \ldots, h_{10}$, deren Nummern in der ersten Spalte angegeben sind. Im unteren Teil der Box wird die durch den Algorithmus gebildete Abstandsfunktion $d_{\mathcal{H}}$ dargestellt.

[1] Vgl. Diday (1986); Gaul und Schader (1994); Lasch (1996); Gil, Capdevila und Arcas (????); Aude, Diaz-Lazcoz, Codani und Risler (1999).

[2] h' ist ein unmittelbarer Nachfolger von h, wenn $h \subset h'$ ist und es kein h'' mit der Eigenschaft $h \subset h'' \subset h'$ gibt.

[3] Vgl. Diday (1986: 207).

[4] Vgl. die Diskussion bei Lasch (1996).

[5] Erzeugt mit dem Skript `clpyr1.cf`.

Box 10.1-1 Pyramidales Klassifikationsschema für die Abstandsmatrix (10.1).

```
 1. cluster number
 2. first element
 3. second element
 4. number of objects
 5. minimal element
 6. maximal element
 7. next on left side
 8. next on right side
 9. index

 1    2    3    4    5    6    7    8      9
-------------------------------------------------
 1    0    0    1    1    1    2    4     0.0000
 2    0    0    1    2    2    0    1     0.0000
 3    0    0    1    3    3    4    0     0.0000
 4    0    0    1    4    4    1    3     0.0000
 5    2    1    2    2    1    0    0     1.0000
 6    4    1    2    1    4    0    0     2.0000
 7    4    3    2    4    3    0    0     4.0000
 8    7    6    3    1    3    0    0     4.0000
 9    6    5    3    2    4    0    0     5.0000
10    9    8    4    2    3    0    0     6.0000


Abstandsmatrix fuer die Cluster

 1 :  0  1  3  2  1  2  3  3  2  3
 2 :  1  0  6  5  1  5  6  6  5  6
 3 :  3  6  0  4  6  4  4  4  6  6
 4 :  2  5  4  0  5  2  4  4  5  5
 5 :  1  1  6  5  0  5  6  6  5  6
 6 :  2  5  4  2  5  0  4  4  5  5
 7 :  3  6  4  4  6  4  0  4  6  6
 8 :  3  6  4  4  6  4  4  0  6  6
 9 :  2  5  6  5  5  5  6  6  0  6
10 :  3  6  6  5  6  5  6  6  6  0
```

Zur bildlichen Veranschaulichung können, ähnlich zu Dendrogrammen, pyramidenartige Graphen vewendet werden. Abbildung 10.1-1 zeigt die Darstellung für das gegenwärtige Beispiel.

4. Indexfunktionen und Abstände. Analog zur Vorgehensweise bei der hierarchischen Klassifikation können Indexfunktionen verwendet werden. Ist ein pyramidales Klassifikationsschema $\mathcal{H}$ gegeben, ist eine Indexfunktion eine Funktion $l : \mathcal{H} \longrightarrow \mathbf{R}$, die jedem Cluster $h \in \mathcal{H}$ eine nicht-negative Zahl $l(h)$ zuordnet und folgende Bedingung erfüllt: Wenn $h \subset h'$, dann $l(h) < l(h')$.

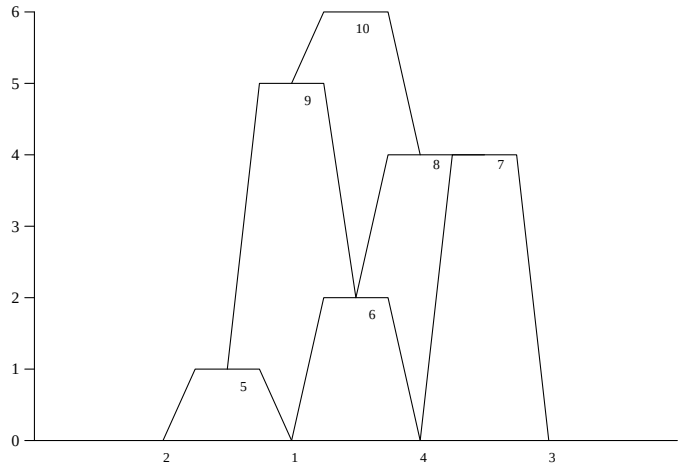


Abb. 10.1-1 Darstellung des pyramidalen Klassifikationsschemas mit den Daten aus Box 10.1-1 (erzeugt mit `clplot4.cf`).

Um eine bestimmte Indexfunktion zu definieren, kann die für die Cluster gebildete Abstandsfunktion $d_{\mathcal{H}}$ verwendet werden.[6] Für den in § 2 beschriebenen Algorithmus eignet sich folgende rekursive Definition:

$$l(\{\omega_i\}) := 0\,, \quad l(h_1 \cup h_2) := d_{\mathcal{H}}(h_1, h_2) \tag{10.2}$$

Die letzte Spalte in Box 10.1-1 zeigt die Werte dieser Indexfunktion für das Beispiel.

Eine Indexfunktion kann dann verwendet werden, um eine neue, das Klassifikationsschema charakterisierende Abstandsmatrix $\mathbf{D}^* = (d^*_{ij})$ für die Ausgangsobjekte zu bilden:

$$d^*_{ij} := \min\{l(h) \,|\, \omega_i, \omega_j \in h\} \tag{10.3}$$

Für das Beispiel sieht diese Abstandsmatrix folgendermaßen aus:

$$\mathbf{D}^* := \begin{pmatrix} 0 & 1 & 4 & 2 \\ 1 & 0 & 6 & 5 \\ 4 & 6 & 0 & 4 \\ 2 & 5 & 4 & 0 \end{pmatrix} \qquad \mathbf{D}^{**} := \begin{pmatrix} 0 & 1 & 6 & 6 \\ 1 & 0 & 6 & 6 \\ 6 & 6 & 0 & 4 \\ 6 & 6 & 4 & 0 \end{pmatrix} \tag{10.4}$$

Mit einer Ausnahme entspricht sie der Ausgangsmatrix in (10.1). $\mathbf{D}^{**}$ ist zum Vergleich die durch eine hierarchische Klassifikation mit der Complete-Link-Methode erzeugte Abstandsmatrix. Offenbar erlaubt die pyramdiale Klassifikation eine bessere Anpassung an die vorgegebenen Abstände.

[6] Vgl. Lasch (1996: 239).

Kapitel 11

Asymmetrische Beziehungen

11.1 Mobilitätstabellen

1. Definition von Mobilitätstabellen.
2. Beispiele zur beruflichen Mobilität.
3. Beispiele mit Eltern und Kindern.

11.2 Input-Output-Tabellen

1. Eine allgemeine Definition.
2. Verflechtung von Wirtschaftssektoren.
3. Konstruktion einer Abstandsfunktion.

11.3 Kapitalverflechtungen

1. Eine allgemeine Definition.

Bei vielen sozialwissenschaftlichen Anwendungen gibt es (zunächst) keine Abstände, sondern asymmetrische Beziehungen. In diesem Kapitel betrachten wir drei Varianten: Dominanzbeziehungen, Mobilitätstabellen und Input-Output-Tabellen.

11.1 Mobilitätstabellen

11.2 Input-Output-Tabellen

1. Eine allgemeine Definition. Als Ausgangspunkt dient die Vorstellung, dass man sich auf Objekte irgendeiner Art beziehen kann, die sich wechselseitig beeinflussen. Was mit „Beeinflussung“ gemeint ist, muss je nach Anwendung festgelegt werden. Exemplarisch kann man an folgende Möglichkeiten denken: (a) Unternehmen, zwischen denen Güter ausgetauscht werden. (b) Objekte irgendeiner Art, zwischen denen Informationen ausgetauscht werden. (c) Aktiengesellschaften, die wechselseitig Kapitalanteile besitzen. (d) Länder, die durch Migrationsströme verbunden sind.

In jedem Fall muss man sich auf eine Menge von Objekten beziehen, die wir wie bisher durch $\Omega = \{\omega_1, \ldots, \omega_n\}$ repräsentieren. Dann kann eine *Input-Output-Tabelle* formal als eine (n, n)-Matrix

$$\mathbf{A} = (a_{ij})$$

definiert werden, wobei a_{ij} die Menge (beispielsweise Gütern oder Informationen) angibt, die von ω_i nach ω_j fließt.[1] Bei den Koeffizienten der Matrix handelt es sich um reelle Zahlen; und wenn es nicht ausdrücklich anders gesagt wird, nehmen wir an, dass alle Koeffizienten nichtnegativ sind: $a_{ij} \geq 0$.

2. Verflechtung von Wirtschaftssektoren. Input-Output-Tabellen werden insbesondere verwendet, um den Austausch von Gütern zwischen Wirtschaftssektoren zu erfassen. Zur Illustration betrachten wir eine aggregierte I-O-Tabelle für Deutschland für das Jahr 1995.[2]

Die vollständige Tabelle basiert auf 56 Sektoren; Box 11.2-1 beschreibt die 12 Sektoren, die für die Aggregation verwendet werden. Erfasst werden die im Inland produzierten Vorleistungsströme (in Mrd. DM). Box 11.2-2 zeigt die entsprechende I-O-Tabelle.

3. Konstruktion einer Abstandsfunktion. I-O-Tabelle können als asymmetrische Tabellen beschrieben und analysiert werden; man kann aus ihnen aber auch symmetrische Abstandsmatrizen konstruieren, die dann mit Verfahren der Clusteranalyse oder der multidimensionalen Skalierung untersucht werden können.

Zur Illustration konstruieren wir aus der I-O-Tabelle in Box 11.2-2 eine Abstandsmatrix. Wird die Tabelle $\mathbf{A} = (a_{ij})$ genannt, kann in einem ersten Schritt eine symmetrische Matrix $\mathbf{B} = (b_{ij})$ erzeugt werden, wobei $b_{ij} := (a_{ij} + a_{ji})/2$ die im Durchschnitt zwischen zwei Sektoren fließenden Vorleistungen erfasst. Dann erhält man durch $d_{ij} := 1/b_{ij}$ Größen, die sich als Abstände interpretieren lassen: Je kleiner d_{ij} ist, desto stärker sind die Sektoren i und j durch Güterströme verflochten.

[1] Die an dieser Stelle zu verwendenden Sprechweisen hängen natürlich von der jeweiligen Anwendung ab.

[2] Fachserie 10, R. 2 Input-Output-Rechnung, 1995.

Tabelle 11.2-1 Aus der I-O-Tabelle in Box 11.2-2 konstruierte Abstandsmatrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.4545	0.8000	0.4762	3.3333	1.1765	0.7407	0.0373	2.5000	0.2353	0.2439	0.7407	1.0526
2	0.8000	0.0476	0.0935	0.1724	0.1389	0.2778	0.4545	0.2273	0.1042	0.1005	0.3226	0.1460
3	0.4762	0.0935	0.0153	0.1299	0.0575	0.1149	0.2532	0.0300	0.0619	0.0451	0.2198	0.1190
4	3.3333	0.1724	0.1299	0.0180	0.0303	0.3279	0.7407	0.0847	0.0806	0.1266	0.6250	0.6897
5	1.1765	0.1389	0.0575	0.0303	0.0067	0.1538	1.1765	0.0862	0.0309	0.0317	0.1481	0.2326
6	0.7407	0.2778	0.1149	0.3279	0.1538	0.0182	0.3636	0.1183	0.0480	0.0485	0.2500	0.1538
7	0.0373	0.4545	0.2532	0.7407	1.1765	0.3636	0.0304	1.8182	0.0461	0.0885	0.2667	0.4255
8	2.5000	0.2273	0.0300	0.0847	0.0862	0.1183	1.8182	0.1282	0.0532	0.0190	0.1695	0.1818
9	0.2353	0.1042	0.0619	0.0806	0.0309	0.0480	0.0461	0.0532	0.0071	0.0130	0.0749	0.0353
10	0.2439	0.1005	0.0451	0.1266	0.0317	0.0485	0.0885	0.0190	0.0130	0.0026	0.0459	0.0322
11	0.7407	0.3226	0.2198	0.6250	0.1481	0.2500	0.2667	0.1695	0.0749	0.0459	0.1099	0.0095
12	1.0526	0.1460	0.1190	0.6897	0.2326	0.1538	0.4255	0.1818	0.0353	0.0322	0.0095	0.0334

Box 11.2-1 Sektoren der aggregierten I-O-Tabelle für 1995.

1	Land- und Forstwirtschaft, Fischerei
2	Bergbau, Gewinnung von Steinen und Erden, Energie- und Wasserversorgung
3	Mineralölverarbeitung, chem. Industrie, Glasgewinnung, Verarbeitung von Steinen u. Erden
4	Metallerzeugung und -bearbeitung
5	Maschinen-, Fahrzeugbau, Datenverarbeitungsgeräte, Elektrotechnik
6	Textil- und Bekleidungsgewerbe, Leder-, Holz-, Papiergewerbe, Recycling u.ä.
7	Ernährungsgewerbe und Tabakverarbeitung
8	Baugewerbe
9	Handel, Verkehr, Nachrichtenübermittlung, Gastgewerbe
10	Finanzierung, Vermietung und Unternehmensdienstleister
11	Gesundheits-, Veterinär- und Sozialwesen, Erziehung und Unterricht, Entsorgung
12	Öffentliche Verwaltung, Verteidigung, Sozialversicherung, sonstige öff. und private Dienstleistungen, häusliche Dienste

Box 11.2-2 Aggregierte I-O-Tabelle für 1995 (inländische Produktion in Mrd. DM). Fachserie 10, R. 2, 1995: 50-51.

	1	2	3	4	5	6	7	8	9	10	11	12
1	2.2	0.1	0.1	0.0	0.0	2.4	47.2	0.0	1.1	1.4	0.6	1.3
2	2.4	21.0	19.4	8.8	7.9	6.5	4.4	5.0	12.6	4.0	5.5	5.5
3	4.1	2.0	65.3	8.9	29.0	11.7	4.9	64.2	10.2	2.1	6.2	14.0
4	0.6	2.8	6.5	55.6	58.4	3.4	2.7	21.4	4.8	0.9	2.2	1.5
5	1.7	6.5	5.8	7.5	150.0	2.5	1.6	20.6	16.8	4.1	12.4	5.9
6	0.3	0.7	5.7	2.7	10.5	54.9	5.4	15.4	17.5	12.9	6.4	8.8
7	6.4	0.0	3.0	0.0	0.1	0.1	32.9	0.0	21.6	0.1	6.0	2.9
8	0.8	3.8	2.5	2.2	2.6	1.5	1.1	7.8	10.3	46.4	10.7	8.8
9	7.4	6.6	22.1	20.0	48.0	24.2	21.8	27.3	141.1	21.3	18.9	42.5
10	6.8	15.9	42.2	14.9	59.0	28.3	22.5	58.6	132.5	388.6	36.8	38.9
11	2.1	0.7	2.9	1.0	1.1	1.6	1.5	1.1	7.8	6.8	9.1	203.9
12	0.6	8.2	2.8	1.4	2.7	4.2	1.8	2.2	14.2	23.3	6.7	29.9

Tabelle 11.2-1 zeigt die auf diese Weise konstruierte Abstandsmatrix; das entsprechende Datenfile wird `iot1.dat` genannt.[3]

[3]Das Datenfile wurde aus der I-O-Tabelle in Box 11.2-2 (`iota95.dat`) mit folgendem Skript (`iot1.cf`) erzeugt:


```
mfmt = 8.4;
mdeff(A) = iota95.dat;
mtransp(A,AT);
mexpr((A+AT)/2,B);
mexpr(1/B,B);
mpr(B)=iot1.dat;
```

11.3 Kapitalverflechtungen

Kapitel 12

Sequenzdaten

12.1 Bestsellersequenzen

1. Der formale Rahmen. Unter *Sequenzdaten* vestehen wir Daten für eine Folge statistischer Variablen, die den gleichen Merkmalsraum haben:

$$(Y_1, \ldots, Y_q) : \Omega \longrightarrow \mathcal{Y}^q$$

Die Referenzmenge Ω kann sich auf Objekte oder Situationen irgendeiner Art beziehen. In vielen Anwendungen entspricht die Folge $t = 1, \ldots, q$ einer Folge von Zeitstellen (zum Beispiel Tage, Monate oder Jahre); es kann sich aber auch um Folgen in einer nichtzeitlichen Dimension handeln. Es wird angenommen, dass alle Variablen Y_t den gleichen Merkmalsraum $\mathcal{Y}$ haben; der Merkmalsraum kann qualitativ oder quantitativ sein. Es wird jedoch nicht vorausgesetzt, dass es für jedes Objekt und jede Stelle einen gültigen Wert gibt. Wir verwenden folgende Konvention: Nichtnegative Werte entsprechen definierten Merkmalsausprägungen; negative Werte geben an, dass es keinen gültigen Wert gibt.

Eine Datenmatrix für Sequenzdaten sieht formal folgendermaßen aus:

$$\mathbf{Y} := \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix} \tag{12.1}$$

Jede Zeile entspricht einem Objekt. Die $i$te Zeile liefert die Sequenz für das $i$te Objekt; wenn es sich um eine zeitliche Sequenz handelt, kann man auch von einer *Zeitreiher* sprechen. Die interpretierbare Sequenz beginnt bei der ersten Stelle und endet bei der letzten Stelle, in der y_{it} nicht negativ ist. Allerdings kann es auch innerhalb dieser Grenzen negative Werte geben; die Sequenz weist dann Lücken auf.

Ein Datensatz mit Sequenzdaten kann natürlich noch weitere Variablen enthalten, die zusätzliche Informationen über die Objekte oder Situationen liefern. Bemerkenswert ist auch, dass die Zustandsvariablen Y_t mehrdimensional sein können. Es gibt dann für jedes Objekt oder jede Situation zwei oder mehr parallele Sequenzen.

2. Daten aus Bestsellerlisten. Zur Illustration verwenden wir Daten aus Beststellerlisten.[1] Es handelt sich um 313 wöchentliche Bestsellerlisten für belletristische Literatur, die im *New York Times Book Review* im Zeitraum vom 5. Januar 1994 bis zum 26. Dezember 1999 veröffentlicht wurden. In jeder Liste gibt es 17 Rangplätze (1 bis 17). Erfasst wurden 546 Titel, die während des genannten Zeitraums in mindestens einer Liste aufgetreten sind.

Orientiert man sich an der Datenmatrix $\mathbf{Y}$ in (12.1), ist in diesem

[1] Wir danken Hugo Verdaasdonk, der uns die Daten freundlicherweise zur Verfügung gestelt hat. Zu einer interessanten kultursoziologischen Verwendung der Daten vgl. man Verdaasdonk (2003).

Beispiel $n = 546$ und $q = 313$. Jede Spalte der Matrix entspricht einer Liste und enthält die Zahlen $1, \ldots, 17$ für die Rangplätze der Bestseller dieser Woche. Jede Zeile der Matrix entspricht einem Titel und zeigt, in welchen Wochen der Titel einen Rangplatz auf einer Bestsellerliste hatte.

3. Auswertung von Sequenzdaten. Bei der Verwendung und Auswertung von Sequenzdaten gibt es insbesondere folgende Möglichkeiten:

a) Man kann die Sequenzdaten verwenden, um die Objekte zu charakterisieren. In unserem Beispiel kann man etwa für jeden Titel ermitteln, wieviele Wochen er auf einer Bestsellerliste aufgetreten ist und welchen durchschnittlichen Rangplatz er eingenommen hat. Man konstruiert also neue Variablen, deren Verteilungen dann beschrieben werden können.

b) Man kann versuchen, unterschiedliche Verlaufsmuster der Sequenzen zu konstruieren. Dann können Häufigkeiten für das Auftreten der verschiedenen Verlaufsmuster ermittelt werden. (Damit beschäftigen wir uns in Kapitel 13.)

c) Man kann die Sequenzen im Hinblick auf ihre Ähnlichkeit untersuchen, also Abstandsfunktionen für die Sequenzen konstruieren. Dafür gibt es unterschiedliche Möglichkeiten. Eine einfache Möglichkeit, bei der die Zeitstruktur ignoriert wird, besprechen wir in § 4; eine komplexere Variante, die unter dem Namen „optimal matching“ bekanntgeworden ist, wird in Abschnitt 12.3 besprochen.

4. Bildung einer Abstandsmatrix. Sequenzdaten können auf einfache Weise ausgewertet werden, wenn bzw. insoweit es sinnvoll möglich ist, von ihrem sequentiellen (zeitlichen) Charakter zu abstrahieren. In unserem Beispiel liegt es nahe, 17 Variablen $X_1, \ldots, X_{17}$ zu definieren, wobei $X_j(\omega)$ angibt, in wievielen Wochen sich der Titel ω auf Platz j befand. Die neue Datenmatrix hat dann die Form

$$\mathbf{X} := \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

wobei jetzt $m = 17$ ist. In dieser Form kann sie offenbar unmitelbar verwendet werden, um Abstände zwischen ihren Zeilen zu definieren.

Allerdings muss beachtet werden, dass es links und rechts zensierte Sequenzen gibt. Wir sprechen von einer *links zensierten Sequenz*, wenn nicht ausgeschlossen werden kann, dass es bereits vor der ersten Zeitstelle gültige Werte gegeben hat; in unserem Beispiel: wenn ein Titel bereits auf einer früheren, nicht erfassten Liste aufgetreten sein könnte. Wir nehmen an, dass dies dann der Fall ist, wenn bei einem Titel bereits die Variable Y_1 einen gültigen Wert hat. Ganz analog sprechen wir von einer *rechts*

zensierten Sequenz, wenn nicht ausgeschlossen werden kann, dass nach dem Ende des Beobachtungsfensters noch gültige Werte auftreten können. Für unser Beispiel nehmen wir an, dass das dann der Fall ist, wenn bei einem Titel die letzte Variable, also Y_{313}, einen gültigen Wert hat.

Mit diesen Definitionen findet man in unserem Beispiel 15 links und 15 rechts zensierte Sequenzen. Schließt man sie aus, verbleiben für die Abstandskonstruktion 516 Titel.

Als Abstände können absolute Differenzen (die sog. City-Block-Metrik) verwendet werden, wobei Unterschiede in den Rangplätzen durch Gewichte berücksichtigt werden. Wir bilden also eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ mit

$$d_{ij} := \sum\nolimits_{k=1,17} w_k \, |x_{ik} - x_{jk}| \tag{12.2}$$

und bestimmten die Gewichte durch $w_k := k$. So entsteht eine Abstandsmatrix mit 516 Zeilen und Spalten.

12.2 Berufliche Mobilität

12.3 Abstandsfunktionen für Sequenzen

Kapitel 13

Konstruktion von Mustern

13.1 Zum Reden von Mustern

13.2 Muster in zeitlichen Verläufen

Anhang A

Mathematische Ergänzungen

Dieser Anhang enthält zu einigen der in den vorangegangenen Kapiteln besprochenen Methoden und Rechenverfahren ergänzende Erläuterungen zu mathematischen Aspekten.

A.1 Nichtmetrische Seriation

In diesem Abschnitt wird besprochen, wie das nicht-metrische Seriationsproblem (Abschnitt 5.1) mit kombinatorischen Methoden (näherungsweise) gelöst werden kann.[1] Wir beziehen uns auf n Objekte, für die eine Abstandsmatrix $\mathbf{D} = (d_{ij})$ gegeben ist. Zu bestimmen sind Zahlen $x_1, \ldots, x_n$, für deren Abstände $d^x_{ij} := |x_i - x_j|$ folgende Bedingung (möglichst gut) erfüllt sein soll:

$$d_{ij} < d_{kl} \implies d^x_{ij} \leq d^x_{kl} \quad \text{und} \quad d_{ij} > d_{kl} \implies d^x_{ij} \geq d^x_{kl} \tag{A.1}$$

Den Leitfaden liefert folgende Überlegung. Da es nur auf die Ordnung der Abstände ankommt, kann man anstelle beliebiger reeller Zahlen die ersten n natürlichen Zahlen verwenden. Die Aufgabe besteht dann darin, eine *Permutation*

$$\pi : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$$

zu finden, so dass eine durch die Zahlen $x_i := \pi(i)$ definierte Abstandsmatrix die Bedingung (A.1) möglichst gut erfüllt.

Ein einfaches Beispiel kann die Überlegung verdeutlichen. Es sei $n = 4$, und die vorgegebene Abstandsmatrix sei

$$\mathbf{D} := \begin{pmatrix} 0 & 5 & 2 & 1 \\ 5 & 0 & 2 & 3 \\ 2 & 2 & 0 & 1 \\ 1 & 3 & 1 & 0 \end{pmatrix}$$

[1] Dazu Literatur: Szczotka (1972); Hubert (1974b, 1987).

Verwendet man die Permutation $\pi(1) = 1, \pi(2) = 4, \pi(3) = 3, \pi(4) = 2$, erhält man die Zahlen $x_1 = 1, x_2 = 4, x_3 = 3, x_4 = 2$ und daraus die Abstandsmatrix

$$\mathbf{D}^x = \begin{pmatrix} 0 & 3 & 2 & 1 \\ 3 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$$

Durch einen Vergleich von $\mathbf{D}$ und $\mathbf{D}^x$ kann man sich davon überzeugen, dass die Bedingung (A.1) erfüllt ist.

Es bleibt die Frage, wie eine geeignete Permutation gefunden werden kann. Bildet man aus den Zahlen $1, \ldots, n$ eine Abstandsmatrix, sieht sie folgendermaßen aus:

$$\mathbf{D}^s = \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots & n-1 \\ 1 & 0 & 1 & 2 & \cdots & n-2 \\ 2 & 1 & 0 & 1 & \cdots & n-3 \\ 3 & 2 & 1 & 0 & \cdots & n-4 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ n-1 & n-2 & n-3 & n-4 & \cdots & 0 \end{pmatrix}$$

Anstatt die Zeilen und Spalten dieser Matrix zu permutieren, so dass durch einen Vergleich mit der Matrix $\mathbf{D}$ die Bedingung (A.1) erfüllbar wird, kann man auch umgekehrt vorgehen: Man kann nach einer Permutation der Zeilen und Spalten von $\mathbf{D}$ suchen, so dass ihre Koeffizienten die gleiche Ordnung aufweisen wie die in $\mathbf{D}^s$. Man sucht dann nach einer Permutation π, so dass die Elemente der Matrix

$$\mathbf{D}^\pi = \begin{pmatrix} d_{\pi(1)\pi(1)} & \cdots & d_{\pi(1)\pi(n)} \\ d_{\pi(2)\pi(1)} & \cdots & d_{\pi(2)\pi(n)} \\ \vdots & & \vdots \\ d_{\pi(n)\pi(1)} & \cdots & d_{\pi(n)\pi(n)} \end{pmatrix}$$

möglichst die gleiche Ordnung aufweisen wie die Elemente in $\mathbf{D}^s$.

Wenn sich eine Permutation π finden lässt, so dass sich die Ordnungen der Elemente in $\mathbf{D}^\pi$ und $\mathbf{D}^s$ genau entsprechen, lässt sich $\mathbf{D}$ perfekt nichtmetrisch eindimensional skalieren. Oft kann man jedoch nur erwarten, eine „möglichst ähnliche“ Ordnung der Matrixelemente zu finden. Was man erreichen kann, hängt von der Matrix $\mathbf{D}$ ab und kann nur festgestellt werden, indem man es ausprobiert. Wünschenswert ist also ein Verfahren, mit dem man eine Permutation π finden kann, so dass die Ordnung der Elemente in $\mathbf{D}^\pi$ der Ordnung der Elemente in $\mathbf{D}^s$ möglichst nahe kommt. Dazu muss zunächst dieses Ziel präzisiert werden. Ein geeignetes Kriterium ist

$$S(\pi) := \sum\nolimits_{i,j} |i-j|\, d_{\pi(i)\pi(j)} \longrightarrow \max \tag{A.2}$$

Dass es sich um eine geeignetes Kriterium handelt, sieht man, wenn man sie so schreibt:

$$S(\pi) = \sum\nolimits_{i,j} (\mathbf{D}^s)_{ij}\, (\mathbf{D}^\pi)_{ij}$$

Durch ihre Maximierung wird also erreicht, dass kleine bzw. große Elemente in $\mathbf{D}^s$ mit kleinen bzw. großen Elementen in $\mathbf{D}^\pi$ in möglichst weitgehende Übereinstimmung gebracht werden. Wie das praktisch gemacht werden kann, soll hier nicht besprochen werden. Es sei nur erwähnt, dass die Maximierung des Kriteriums $S(\pi)$ ein Spezialfall eines allgemeineren Problems ist, das in der Literatur unter dem Namen „quadratic assignment problem“ diskutiert wird. Verfahren, die ein globales Maximum finden, sind sehr rechenaufwendig und nur für verhältnismäßig kleine Abstandsmatrizen ($n \leq 18$) praktikabel. Es gibt jedoch effiziente Verfahren, die auch bei größeren Matrizen näherungsweise gute Lösungen finden.

A.2 Metrische eindimensionale Skalierung

1. Die folgenden Ausführungen betreffen das Problem der metrischen eindimensionalen Skalierung, dass in Abschnitt 5.1 besprochen wurde. Ausgangspunkt ist eine (n,n)-Abstandsmatrix $\mathbf{D} = (d_{ij})$. Die Aufgabe besteht darin, reelle Zahlen $x_1, \ldots, x_n$ zu finden, die das Kriterium

$$f(x_1, \ldots, x_n) := \sum_{i=2}^{n} \sum_{j=1}^{i-1} (d_{ij} - |x_i - x_j|)^2 \tag{A.3}$$

minimal machen.[2] Man kann sich auch vorstellen, dass durch dieses Kriterium zwei Abstandsmatrizen verglichen werden. Einerseits die gegebene Matrix $\mathbf{D}$, und andererseits eine durch die Zahlen $\mathbf{x} = (x_1, \ldots, x_n)$ induzierte Abstandsmatrix

$$\mathbf{D^x} = \begin{pmatrix} |x_1 - x_1| & |x_1 - x_2| & \cdots & |x_1 - x_n| \\ |x_2 - x_1| & |x_2 - x_2| & \cdots & |x_2 - x_n| \\ \vdots & \vdots & & \vdots \\ |x_n - x_1| & |x_n - x_2| & \cdots & |x_n - x_n| \end{pmatrix}$$

Offenbar gilt

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \frac{1}{2} \| \mathbf{D} - \mathbf{D^x} \|^2$$

D.h. die Zahlen, die das Kriterium f minimal machen, liefern zugleiche eine neue Abstandsmatrix $\mathbf{D^x}$, deren euklidischer Abstand zur vorgegebenen Abstandsmatrix $\mathbf{D}$ minimal ist. Aus dieser Darstellung erkennt man auch, dass der Lösungsvektor $\mathbf{x}$ nicht eindeutig bestimmt ist. Wenn man $\mathbf{x}$ mit -1 multipliziert oder eine beliebige Konstante hinzufügt, verändert sich die induzierte Abstandsmatrix $\mathbf{D^x}$ und also auch der Wert des Kriteriums nicht. – Zur Illustration verwenden wir die Abstandsmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 3 & 4 \\ 1 & 0 & 2 & 4 \\ 3 & 2 & 0 & 6 \\ 4 & 4 & 6 & 0 \end{pmatrix} \tag{A.4}$$

Ein Minimum für das Kriterium erhält man in diesem Beispiel durch $\hat{\mathbf{x}} = (0, 0.75, 2.75, -3.5)$ mit der induzierten Abstandsmatrix

$$\mathbf{D^{\hat{x}}} = (|\hat{x}_i - \hat{x}_j|) = \begin{pmatrix} 0 & 0.75 & 2.75 & 3.5 \\ 0.75 & 0 & 2 & 4.25 \\ 2.75 & 2 & 0 & 6.25 \\ 3.5 & 4.25 & 6.25 & 0 \end{pmatrix}$$

[2] Da $\mathbf{D}$ symmetrisch ist und alle Hauptdiagonalelemente Null sind, kann man sich zur Definition des Kriteriums auf die Elemente unterhalb der Hauptdiagonalen beschränken.

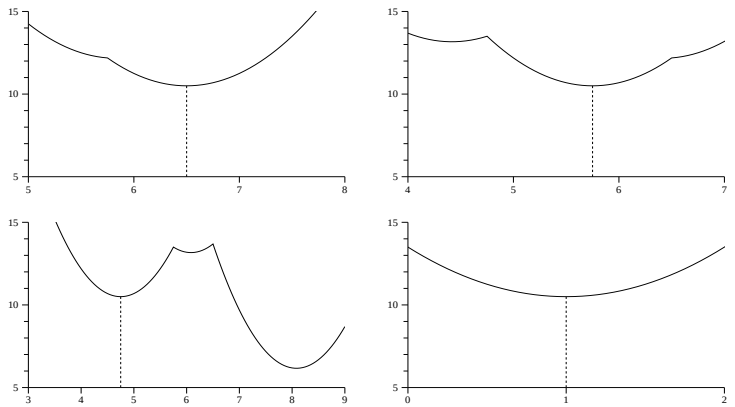


Abb. A.2-1 Projektion der Zielfunktion $f(x_1, x_2, x_3, x_4)$ auf die vier Koordinaten in Umgebungen des lokalen Minimums $(6.5, 5.75, 4.75, 1)$.

und man findet

$$f(\hat{\mathbf{x}}) = f(0, -0.75, -2.75, 3.5) = \frac{1}{2} \| \mathbf{D} - \mathbf{D^{\hat{x}}} \|^2 = 0.5$$

Wie aber findet man einen Vektor $\hat{\mathbf{x}}$, der das Kriterium f minimal macht? Die Schwierigkeiten resultieren daraus, dass diese Funktion nicht überall stetig differenzierbar und auch nicht global konvex ist, so dass die üblichen Minimierungsmethoden nicht (ohne weiteres) verwendet werden können. Zur Illustration bleiben wir bei unserem Beispiel. Die dem Kriterium entsprechende Zielfunktion hat zahlreiche lokale Minima, z.B. an der Stelle $\mathbf{x} = (6.5, 5.75, 4.75, 1)$, an der die Funktion den Wert 10.5 annimmt. Abbildung A.2-1 zeigt, wie die Funktion in einer Umgebung von $\mathbf{x}$ aussieht. Man erkennt nicht nur, dass es sich um ein lokales Minimum handelt, sondern sieht auch einige der Stellen, an denen die Funktion nicht stetig differenzierbar ist.

2. Eine Möglichkeit, mit diesen Schwierigkeiten umzugehen, besteht darin, der Reihe nach alle Permutationen der Abstandsmatrix $\mathbf{D}$ zu betrachten.[3] Permutationen beziehen sich in diesem Zusammenhang auf die n Objekte, deren paarweise Abstände durch $\mathbf{D}$ gegeben sind. Jede Permutation ist eine ein-eindeutige Abbildung

$$\pi : \{1, \ldots, n\} \longrightarrow \{1, \ldots, n\}$$

so dass $\pi(i)$ die neue Nummer des $i$ten Objekts ist. Zum Beispiel könnte man die Permutation $\pi(1) = 4,\ \pi(2) = 1,\ \pi(3) = 2,\ \pi(4) = 3$ verwenden, um aus der in (A.4) angegebenen Abstandsmatrix die permutierte

[3] Die folgenden Überlegungen orientieren sich an Defays (1978).

Abstandsmatrix

$$\mathbf{D}^\pi \;=\; (d_{ij}^\pi) \;=\; \begin{pmatrix} 0 & 4 & 4 & 6 \\ 4 & 0 & 1 & 3 \\ 4 & 1 & 0 & 2 \\ 6 & 3 & 2 & 0 \end{pmatrix} \tag{A.5}$$

zu bilden. Die Menge aller möglichen Permutationen bezeichnen wir mit Π_n. Sie enthält $n!$ Elemente.

Indem man sich jeweils gesondert auf eine permutierte Abstandsmatrix $\mathbf{D}^\pi$ bezieht, kann man den Parameterraum, in dem man nach einer Lösung sucht, sinnvoll einschränken. Anstatt alle möglichen Vektoren $(x_1, \ldots, x_n)$ zu betrachten, kann man die Suche auf den Parameterraum

$$P_n \;:=\; \{\, (x_1, \ldots, x_n) \,|\, x_1 \le x_2 \le \cdots \le x_n, x_1 < x_n, \sum_{j=1}^{n} x_j = 0 \}$$

beschränken. Ist $\pi \in \Pi_n$ irgendeine Permutation, wird also nach einem Vektor $(x_1, \ldots, x_n) \in P_n$ gesucht, der das Kriterium

$$f^\pi(x_1, \ldots, x_n) \;:=\; \sum_{i=2}^{n} \sum_{j=1}^{i-1} (d_{ij}^\pi - \,|\, x_i - x_j \,|\,)^2 \;=\; \sum_{i=2}^{n} \sum_{j=1}^{i-1} (d_{ij}^\pi - x_i + x_j \,)^2$$

minimal macht. Der Vorteil liegt darin, dass diese neue Zielfunktion innerhalb des eingeschränkten Parameterraums stetig differenzierbar ist.

Infolgedessen kann man dann auch versuchen, die ersten Ableitungen der Zielfunktion zu bilden und daraus notwendige Bedingungen für ein (lokales) Minimum zu finden. Als erste Ableitungen findet man

$$\frac{\partial f^\pi(x_1, \ldots, x_n)}{\partial x_k} \;=\; -2 \sum_{j=1}^{k-1} (d_{kj}^\pi - \, x_k + x_j \,) \;+\; 2 \sum_{j=k+1}^{n} (d_{kj}^\pi - \, x_j + x_k \,)$$

für $k = 1, \ldots, n$, und daraus gewinnt man als notwendige Bedingungen für ein Minimum die Gleichungen

$$\sum_{j=1}^{k-1} (d_{kj}^\pi - \, x_k + x_j \,) \;=\; \sum_{j=k+1}^{n} (d_{kj}^\pi - \, x_j + x_k \,) \tag{A.6}$$

für $k = 1, \ldots, n$. Eine einfache Umformung ergibt

$$\sum_{j=1}^{k-1} d_{kj}^\pi \;-\; \sum_{j=k+1}^{n} d_{kj}^\pi \;=\; \sum_{j=1}^{k-1} (x_k - x_j \,) \;+\; \sum_{j=k+1}^{n} (x_k - x_j)$$

Definiert man zur Abkürzung

$$d_{\cdot k}^\pi \;:=\; \sum_{j=1}^{k-1} d_{kj}^\pi \quad \text{und} \quad d_{k\cdot}^\pi \;:=\; \sum_{j=k+1}^{n} d_{kj}^\pi$$

erhält man schließlich die Darstellung

$$x_k \;=\; \frac{1}{n} (d_{\cdot k}^\pi - d_{k\cdot}^\pi) + \frac{1}{n} \sum_{j=1}^{n} x_j \quad (\text{für } k = 1, \ldots, n) \tag{A.7}$$

Aus dieser Darstellung erkennt man, dass die Gleichungen (A.6) keine eindeutige Lösung haben. Man kann die Zahlen x_k mit einer beliebigen Konstanten addieren, ohne dass sich an den Gleichungen (A.7) etwas ändert. Insbesondere kann man also erreichen, dass die in der Definition von P_n geforderte Bedingung $\Sigma_k x_k = 0$ erfüllt wird, indem man den Vektor

$$\mathbf{x}^\pi \;=\; (x_1^\pi, \ldots, x_n^\pi) \quad \text{mit} \quad x_k^\pi \;:=\; \frac{1}{n} (d_{\cdot k}^\pi - d_{k\cdot}^\pi)$$

betrachtet.[4] Allerdings gilt nicht unbedingt, dass der so definierte Vektor $\mathbf{x}^\pi$ ein Element des Parameterraums P_n ist, denn es gilt nicht notwendigerweise $x_1^\pi \le x_2^\pi \le \cdots \le x_n^\pi$. Wir werden jedoch später zeigen, dass man diejenigen Permutationen, bei denen $\mathbf{x}^\pi$ kein Element von P_n ist, bei der Suche nach einem globalen Minimum der Zielfunktion ignorieren kann.

Nehmen wir also zunächst an, dass $\mathbf{x}^\pi \in P_n$ ist. Dann bleibt immer noch zu überlegen, ob $\mathbf{x}^\pi$ ein (lokales) Minimum der Zielfunktion f^π liefert. Das kann mithilfe der zweiten Ableitungen entschieden werden, die bei unserer Zielfunktion auch eine sehr einfache Gestalt haben, nämlich

$$\frac{\partial f^\pi(x_1, \ldots, x_n)}{\partial x_k \partial x_l} \;=\; \begin{cases} -2 & \text{wenn } k \neq l \\ -2 + 2\,n & \text{wenn } k = l \end{cases}$$

Also sieht die sog. Hesse-Matrix folgendermaßen aus:

$$\mathbf{H} \;=\; \left(\frac{\partial f^\pi(x_1, \ldots, x_n)}{\partial x_k \partial x_l} \right) \;=\; \begin{pmatrix} -2 & \cdots & -2 \\ \vdots & & \vdots \\ -2 & \cdots & -2 \end{pmatrix} + \begin{pmatrix} 2n & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 2n \end{pmatrix}$$

Diese Matrix ist jedoch innerhalb des Parameterraums P_n positiv definit. Denn sei $\mathbf{x}$ irgendein Vektor in P_n. Dann ist

$$\mathbf{x}'\mathbf{H}\mathbf{x} \;=\; 2\,n\,\mathbf{x}'\mathbf{x} \;>\; 0$$

Also ist die Zielfunktion f^π innerhalb des Parameterraums P_n streng konvex, so dass man weiß: Wenn $\mathbf{x}^\pi \in P_n$ ist, dann hat man an dieser Stelle ein Minimum der Funktion f^π gefunden. – Zur Illustration verwenden wir

[4] Diese Möglichkeit resultiert aus der Symmetrie der Abstandsmatrix $\mathbf{D}$. Denn infolgedessen gilt stets: $\sum_{k=1}^{n} d_{k\cdot} = \sum_{k=1}^{n} d_{\cdot k}$.

die in (A.5) angegebene Abstandsmatrix $\mathbf{D}^\pi$. Man findet:

k	$d^\pi_{\cdot k}$	$d^\pi_{k\cdot}$	$d^\pi_{\cdot k} - d^\pi_{k\cdot}$	x^π_k
1	0	14	−14	−3.50
2	4	4	0	0.00
3	5	2	3	0.75
4	11	0	11	2.75

so dass in diesem Fall offenbar $\mathbf{x}^\pi \in P_n$ ist. Somit hat man auch ein (lokales) Minimum von f^π gefunden, das den Wert der Zielfunktion

$$f^\pi(\hat{\mathbf{x}^\pi}) \;=\; f^\pi(-3.5, 0, 0.75, 2.75) \;=\; \frac{1}{2}\;\|\,\mathbf{D}^\pi - \mathbf{D}^{\mathbf{x}^\pi}\,\|^2 = \; 0.5$$

liefert. Und indem man die Permutation rückgängig macht, also π^{-1} bildet,[5] findet man den Vektor

$$\mathbf{x}^{\pi^{-1}\pi} \;=\; (0, 0.75, 2.75, -3.5)$$

bei dem die Komponenten so angeordnet sind, wie es der ursprünglichen Reihenfolge der Objekte für die Bildung der nicht-permutierten Abstandsmatrix $\mathbf{D}$ entspricht. Dieser Vektor ist ersichtlich mit dem Vektor $\hat{\mathbf{x}}$ identisch, der zu Beginn dieses Abschnitts verwendet worden ist.

3. Zunächst hat man allerdings nur ein Minimum der Zielfunktion innerhalb des Parameterraums P_n gefunden, also bei einer bestimmten vorausgesetzten Reihenfolge (Permutation) der Objekte. Somit bleibt die Frage, ob bzw. wie man auch ein globales Minimum finden kann. Zuvor beschäftigen wir uns jedoch mit der Frage, welchen Wert die Zielfunktion f^π an der Stelle $\mathbf{x}^\pi$ annimmt, wenn vorausgesetzt werden kann, dass $\mathbf{x}^\pi \in P_n$ ist. Zunächst gilt sicherlich:

$$\begin{aligned} f^\pi(\mathbf{x}^\pi) \;&=\; \sum_{i=2}^{n}\sum_{j=1}^{i-1}(d_{ij} - (x^\pi_i - x^\pi_j))^2 \\ &=\; \sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}^2 \;+\; \sum_{i=2}^{n}\sum_{j=1}^{i-1}(x^\pi_i - x^\pi_j)^2 \;-\; 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}(x^\pi_i - x^\pi_j) \end{aligned}$$

[5] Mit π^{-1} bezeichnen wir die zu π inverse Permutation, die dadurch definiert ist, dass für $i = 1, \ldots, n$ gilt: $\pi^{-1}(\pi(i)) = i$.

Nun gilt jedoch für den mittleren Term auf der rechten Seite:[6]

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x^\pi_i - x^\pi_j)^2 \;=\; n\sum_{i=1}^{n} x^\pi_i x^\pi_i$$

und für den rechten Term auf der rechten Seite findet man ebenfalls einen einfachen Ausdruck:[7]

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}\,(x^\pi_i - x^\pi_j) \;=\; n\sum_{i=1}^{n} x^\pi_i x^\pi_i$$

Also erhält man insgesamt:

$$f^\pi(\mathbf{x}^\pi) \;=\; \sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}^2 \;-\; n\sum_{i=1}^{n} x^\pi_i x^\pi_i \tag{A.8}$$

Zur Illustration kann wieder unser Beispiel dienen. In diesem Fall ist

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}^2 \;=\; 82 \quad \text{und} \quad n\sum_{i=1}^{4} x^\pi_i x^\pi_i \;=\; 81.5$$

so dass die Zielfunktion den Wert $f^\pi(\mathbf{x}^\pi) = 0.5$ annimmt.

[6] Man sieht dies folgendermaßen:

$$\begin{aligned} \sum_{i=2}^{n}\sum_{j=1}^{i-1}(x^\pi_i - x^\pi_j)^2 &= \sum_{i=2}^{n}\sum_{j=1}^{i-1} x^\pi_i x^\pi_i + x^\pi_j x^\pi_j - 2\,x^\pi_i x^\pi_j \\ &\sum_{i=2}^{n}(i-1)\,x^\pi_i x^\pi_i + \sum_{j=1}^{n-1}(n-i)\,x^\pi_j x^\pi_j + \sum_{i=1}^{n} x^\pi_i x^\pi_i - \sum_{i=1}^{n}\sum_{j=1}^{n} x^\pi_i x^\pi_j \;= \\ &\sum_{i=1}^{n}(i-1)\,x^\pi_i x^\pi_i + \sum_{i=1}^{n}(n-i)\,x^\pi_i x^\pi_i + \sum_{i=1}^{n} x^\pi_i x^\pi_i \;= \\ &(n-1)\sum_{i=1}^{n} x^\pi_i x^\pi_i + \sum_{i=1}^{n} x^\pi_i x^\pi_i \;=\; n\sum_{i=1}^{n} x^\pi_i x^\pi_i \end{aligned}$$

denn da $\mathbf{x}^\pi \in P_n$ ist, ist $\Sigma_i x^\pi_i = 0$ und infolgedessen auch die Doppelsumme in der zweiten Zeile auf der rechten Seite.

[7] Man sieht dies folgendermaßen:

$$\begin{aligned} \sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}\,(x^\pi_i - x^\pi_j) &= \sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}x^\pi_i - \sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}x^\pi_j \;= \\ &\sum_{i=2}^{n} d_{\cdot i}x^\pi_i - \sum_{j=1}^{n-1}\sum_{i=j+1}^{n} d_{ji}x^\pi_j \;=\; \sum_{i=2}^{n} d_{\cdot i}x^\pi_i - \sum_{j=1}^{n-1} d_{j\cdot}x^\pi_j \;= \\ &\sum_{i=1}^{n} d_{\cdot i}x^\pi_i - \sum_{i=1}^{n} d_{i\cdot}x^\pi_i \;=\; \sum_{i=1}^{n} x^\pi_i\,(d_{\cdot i} - d_{i\cdot}) \;=\; n\sum_{i=1}^{n} x^\pi_i x^\pi_i \end{aligned}$$

denn $d_{\cdot 1} = d_{n\cdot} = 0$.

A.3 Skalierung mit Eigenvektoren

In diesem Abschnitt wird besprochen, wie das Kriterium $q_1(\mathbf{s})/q(\mathbf{s})$ für die Skalierung mit Eigenvektoren maximiert werden kann (vgl. Abschnitt 6.1, dessen Notationen hier übernommen werden). Um die als Kriterium verwendete Funktion mit Matrizen schreiben zu können, werden Diagonalmatrizen $\mathbf{C} := \mathrm{diag}(f_{.1}, \ldots, f_{.m})'$ und $\mathbf{R} := \mathrm{diag}(f_{1.}, \ldots, f_{n.})'$ verwendet. Wegen der Normierungsbedingung (6.3) ist $\bar{s} = 0$, und man kann schreiben:

$$q(\mathbf{s}) = \sum_i \sum_j f_{ij}\, s_j^2 = \mathbf{s}'\mathbf{C}\mathbf{s}$$

und

$$q_1(\mathbf{s}) = \sum_i \frac{1}{f_{i.}} \left(\sum_j f_{ij}\, s_j \right)^2 = \mathbf{s}'\mathbf{F}'\mathbf{R}^{-1}\mathbf{F}\mathbf{s}$$

Also kann die Zielfunktion insgesamt so geschrieben werden:

$$\frac{q_1(\mathbf{s})}{q(\mathbf{s})} = \frac{\mathbf{s}'\mathbf{F}'\mathbf{R}^{-1}\mathbf{F}\mathbf{s}}{\mathbf{s}'\mathbf{C}\mathbf{s}} \longrightarrow \max$$

Nun wird definiert:

$$\mathbf{w} := \mathbf{C}^{1/2} \quad \text{mit} \quad \mathbf{C}^{1/2} = \mathrm{diag}(\sqrt{f_{.1}}, \ldots, \sqrt{f_{.m}})$$

und somit ist $\mathbf{s}'\mathbf{C}\mathbf{s} = \mathbf{w}'\mathbf{w}$. Definiert man außerdem[8]

$$\mathbf{A} := \mathbf{R}^{-1}\mathbf{F}\,\mathbf{C}^{-1}$$

kann man die Zielfunktion folgendermaßen als eine Funktion von $\mathbf{w}$ schreiben:

$$\lambda(\mathbf{w}) = \frac{\mathbf{w}'\mathbf{A}'\mathbf{A}\mathbf{w}}{\mathbf{w}'\mathbf{w}} \longrightarrow \max \tag{A.9}$$

Differenziert man $\lambda(\mathbf{w})$ nach $\mathbf{w}$, findet man folgende Bedingungen für ein Maximum:

$$\mathbf{A}'\mathbf{A}\mathbf{w} = \lambda(\mathbf{w})\,\mathbf{w}$$

Das heißt, dass die Funktion $\lambda(\mathbf{w})$ dann Extremstellen hat, wenn man für $\mathbf{w}$ einen Eigenvektor von $\mathbf{A}'\mathbf{A}$ verwendet, und die Funktion hat dann als Wert den zugehörigen Eigenwert.

Der größte Eigenwert ist allerdings stets 1 und hat als zugehörigen Eigenvektor $\mathbf{w} = \mathbf{C}^{-1}\mathbf{1}_m$.[9] Das sieht man folgendermaßen:[10]

$$\begin{aligned}\mathbf{A}'\mathbf{A}\mathbf{w} &= \mathbf{C}^{-1/2}\mathbf{F}'\mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\mathbf{1}_n = \mathbf{C}^{-1/2}\mathbf{F}'\mathbf{R}^{-1/2}\mathbf{F}\mathbf{1}_n \\ &= \mathbf{C}^{-1/2}\mathbf{F}'\mathbf{R}^{-1/2}\mathbf{F}\mathbf{1}_n = \mathbf{C}^{-1/2}\mathbf{F}'\mathbf{1}_n = \mathbf{C}^{1/2}\mathbf{1}_n = \mathbf{w}\end{aligned}$$

[8] Dies entspricht der in Abschnitt 6.1 verwendeten Matrix $\mathbf{A}$.

[9] $\mathbf{1}_m$ dient hier zur Bezeichnung eines Vektors, der aus m Einsen besteht.

[10] Man beachte: $\mathbf{F}\mathbf{1}_n = \mathbf{R}\mathbf{1}_n$ und $\mathbf{F}'\mathbf{1}_m = \mathbf{C}\mathbf{1}_m$.

A.4 Regression mit Scores

In diesem Abschnitt wird besprochen, wie Werte für die Parameter der in Abschnitt 6.3 definierten Regression mit Scores berechnet werden können. Ausgangspunkt ist die Funktion (6.6), die minimiert werden soll. Wegen

$$\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{e}'\mathbf{e}$$

kann man stattdessen auch die Funktion

$$\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha}} \longrightarrow \max \tag{A.10}$$

verwenden, wobei $\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha} = c$ mit einer beliebigen Konstanten c als Nebenbedingung verwendet wird. Nun folgt aus dem Regressionsansatz

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\alpha} \tag{A.11}$$

also $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha}$, wobei

$$\mathbf{P} := \mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \tag{A.12}$$

so dass die Zielfunktion auch so geschrieben werden kann:

$$f(\boldsymbol{\alpha}) := \frac{\boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha}} \longrightarrow \max \tag{A.13}$$

Bedingungen für Extremwerte findet man aus den ersten Ableitungen (man beachte, dass $\mathbf{P}$ symmetrisch ist):

$$\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \frac{(2\mathbf{P}\boldsymbol{\alpha})(\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha}) - (\boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha})(2\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha})}{(\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha})^2} = 0$$

Daraus folgt

$$\mathbf{P}\boldsymbol{\alpha} = \frac{\boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha}}\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha}$$

Definiert man $\mathbf{Q} := (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{P}$ und $\lambda := (\boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha})/(\boldsymbol{\alpha}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\alpha})$, kann man die Gleichung auch so schreiben:

$$\mathbf{Q}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{A.14}$$

Die Scores für das Regressionsproblem gewinnt man also als Eigenvektoren dieser Gleichung; und zur Maximierung der Funktion (A.10) sollte der zum maximalen Eigenwert gehörende Eigenvektor verwendet werden. Dieser Eigenvektor kann dann beliebig normiert werden. Und schließlich kann $\boldsymbol{\beta}$ aus (A.11) berechnet werden.

Anhang B

Hinweise auf Programme

Dieser Anhang enthält einige Hinweise auf Programme und deren Prozeduren, die für praktische Berechnungen verwendet werden können. Es handelt sich nicht um eine Einführung in die Verwendung der Programme.

B.1 Verwendete TDA-Prozeduren

In diesem Abschnitt werden einige der TDA-Prozeduren, die in den vorangegangenen Kapiteln verwendet wurden, kurz in alphabetischer Reihenfolge erläutert. Nicht ausdrücklich genannt werden Matrixbefehle, die für einfache und schnelle Berechnungen oft nützlich sind. Um beispielsweise für die Daten in einem File `fn` eine Singularwertzerlegung durchzuführen, genügen die Befehle

```
mdeff(X)=fn; msvd1(X,Q,U,V);
```

Der erste Befehl liest das Datenfile und erzeugt die Matrix $\mathbf{X}$; der zweite Befehl führt die Singularwertzerlegung durch und erzeugt als Ergebnis die Matrizen $\mathbf{Q}$, $\mathbf{U}$ und $\mathbf{V}$, die dann ausgedruckt oder weiterverwendet werden können

`clp` verwendet Varianten des Kriteriums (9.2) in Abschnitt 9.2 für partitionierende Clusteranalysen.

`dmet` kann verwendet werden, um Abstandsmatrizen zu modifizieren, so dass die Dreiecksungleichung erfüllt wird (vgl. Abschnitt 1.1, § 6).

`hcls` kann für hierarchische Clusteranalysen mit SAHN-Algorithmen verwendet werden (vgl. Abschnitt 8.1, § 2).[1]

`mdsc` Diese Prozedur ermöglicht multidimensionale Skalierung mit Hauptkoordinaten (Abschnitt 4.2).

`mdsm` ist eine Prozedur für die metrische multidimensionale Skalierung, wie sie in Abschnitt 4.3 besprochen wurde. Es können mit unterschiedlichen Minimierungsalgorithmen zweidimensionale Konfigurationen erzeugt werden. Die Prozedur erlaubt eine beliebige Anzahl von Wiederholungen, die von zufällig erzeugten Anfangskonfigurationen ausgehen. Auf diese Weise kann das Problem, dass die Minimierungsalgorithmen nur lokale Minima finden können, in gewissem Umfang einschätzbar gemacht werden.

[1] Die Prozedur beruht auf einem von Späth (1975: 172) publizierten Algorithmus.

`pdatd` Diese Prozedur kann verwendet werden, um Distanzmatrizen zu erzeugen.

`scla` kann zur Ermittlung separierbarer Cluster verwendet werden (vgl. Abschnitt 7.1).

Literatur

Anderberg, M. R. 1973. Cluster Analysis for Applications. New York: Academic Press.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., Sander, J. 1999: OPTICS: Ordering Points to Identify the Clustering Structure. Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, 49–60.

Arabie, P., Boorman, S. A. 1982. Blockmodels: Development and Prospects. In: H. C. Hudson (ed.), Classifying Social Data, 177–198. San Francisco: Jossey-Bass.

Aude, J-C., Diaz-Lazcoz, Y., Codani, J-J., Risler, J-L. 1999. Applications of the Pyramidal Clustering Method to Biological Objects. Computers & Chemistry 23, 303–315.

Augustson, J. G., Minker, J. 1970. An Analysis of Some Graph Theoretical Cluster Techniques. Journal of the ACM 17, 571–588.

Bacher, J. 1994. Clusteranalyse. München: Oldenbourg.

Bailey, K. D. 1983. Sociological Classification and Cluster Analysis. Quality and Quantity 17, 251–268.

Bailey, K. D. 1994. Typologies and Taxonomies. An Introduction to Classification Techniques. London: Sage.

Barthélemy, J.-P., Brucker, F., Osswald, C. 2007. Combinatorial Optimisation and Hierarchical Classifications. Annals of Operations Research 153, 179–214.

Barthélemy, J.-P., Guénoche, A. 1991. Trees and Proximity Representations. New York: Wiley.

Batagelj, V., Bren, M. (1995). Comparing Resemblance Measures. Journal of Classification 12, 73–90.

Bénasséni, J., Dosse, M. B., Joly, S. 2007. On a General Transformation Making a Dissimilarity Matrix Euclidean. Journal of Classification 24, 33–51.

Blashfield, R. K., Aldenderfer, M. S. 1988. The Methods and Problems of Cluster Analysis. In: J. R. Nesselroade, R. B. Cattell (eds.), Handbook of Multivariate Experimental Psychology, 447–473. New York: Plenum Press.

Blasius,J. 2001. Korrespondenzanalyse. München: Oldenbourg.

Bock, H. H. 1974. Automatische Klassifikation. Göttingen: Vandenhoeck & Ruprecht.

Borg, I., Lingoes, J. 1987. Multidimensional Similarity Structure Analysis. New York: Springer.

Bortz, J., Döring, N. 1995. Forschungsmethoden und Evaluation. Berlin: Springer-Verlag.

Brusco, M. J. 2002. Integer Programming Methods for Seriation and Unidimensional Scaling of Proximity Matrices: A Review and Some Extensions. Journal of Classification 19, 45–67.

Charikar, M., Panigrahy, R. 2001. Clustering to Minimize the Sum of Cluster Diameters. Proceedings on 33rd Ann. ACM Symposium on Theory of Computing, 1–10.

Charles, M., Grusky, D. B. 1995. Models for Describing the Underlying Structure of Sex Segregation. American Journal of Sociology 100, 931–971.

Clausen, S.-E. 1998. Applied Correspondence Analysis. An Introduction. London: Sage.

Cliff, N., McCormick, D. J., Zatkin, J. L., Cudeck, R. A., Collins, L. M. 1986. BINCLUS: Nonhierarchical Clustering of Binary Data. Journal of Mathematical Psychology 21, 201–227.

Commandeur, J. F. 1991. Matching Configurations. Leiden University: DSWO Press.

Corter, J. E. 1996. Tree Models of Similarity and Association. Thousand Oakes: Sage.

Cox, T. F., Cox, M. A. 1994. Multidimensional Scaling. London: Chapman & Hall.

Defays, D. 1978. A Short Note on a Method of Seriation. British Journal of Math. and Statist. Psychology 31, 49–53.

De Soete, G. 1984. Ultrametric Tree Representations of Incomplete Dissimilarity Data. Journal of Classification 1, 235–242.

De Soete, G. 1988. Tree Representations of Proximity Data by Least Squares Methods. In: H. H. Bock (ed.), Classification and Related Methods of Data Analysis, 147–156. Amsterdam: Elsevier.

De Soete, G., Carroll, J. D., De Sarbo, W. S. 1987. Least Squares Algorithms for Constructing Constrained Ultrametric and Additive Tree Representations of Symmetric Proximity Data. Journal of Classification 4, 155–173.

Diday, E. 1986. Orders and Overlapping Clusters by Pyramids. In: J. de Leeuw, W. Heiser, J. Meulman, F. Critchley (eds.), Multidimensional Data Analysis, 201–234. Leiden: DSWO Press.

Elisseeff, V. 1968. De L'Application des Propriétés du Scalogramme a L'Étude des Objets. In: Calcul et Formalisation Dans les Sciences de L'Homme, 107–120. Paris: Éditions du Centre National de la Recherche Scientifique.

Everitt, B. S. 1993. Cluster Analysis, 3rd ed. London: Arnold.

Falk, M., Becker, R., Marohn, F. 1995. Angewandte Statistik mit SAS. Berlin: Springer-Verlag.

Faust, K., Wasserman, S. 1993. Correlation and Association Models for Studying Measurements on Ordinal Relations. Sociological Methodology 23, 177–215.

Fox, J. 1982. Selective Aspects of Measuring Resemblance for Taxonomy. In: H. C. Hudson (ed.), Classifying Social Data, 127–151. San Francisco: Jossey-Bass.

Ganti, V., Gehrke, J., Ramakrishnan, R. 1999. CACTUS – Clustering Categorical Data Using Summaries. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 73–83.

Gaul, W., Schader, M. 1994. Pyramidal Classification Based on Incomplete Dissimilarity Data. Journal of Classification 11, 171–193.

Gil, A. J., Capdevila, C., Arcas, A. xxxx. On the Efficiency and Sensitivity of a Pyramidal Classification Algorithm.

Gordon, A. D. 1987. A Review of Hierarchical Classification. Journal of the Royal Statistical Society, A 150, 119–137.

Gower, J. C. 1988. Classification, Geometry and Data Analysis. In: H. H. Bock (ed.), Classification and Related Methods of Data Analysis, 3–14. Amsterdam: Elsevier Science Publ.

Green, P. E., Carmone, F. J., Smith, S. M. 1989. Multidimensional Scaling. Concepts and Applications. Boston: Allyn and Bacon.

Greenacre, M. J. 1993. Correspondence Analysis in Practice. New York: Academic Press.

Greenacre, M. J., Blasius, J. (eds.) 1994. Correspondence Analysis in the Social Sciences. New York: Academic Press.

Groenen, P. J. F. 1993. The Majorization Approach to Multidimensional Scaling. Leiden: DSWO Press.

Groenen, P. J. F., Heiser, W. J., Meulman, J. J. 1998. City-Block Scaling: Smoothing Strategies for Avoiding Local Minima. In: I. Balderjahn, R. Mathar, M Schader (eds.), Classification, Data Analysis, and Data Highways, 46–53. Berlin: Springer-Verlag.

Hansen, P., Jaumard, B. 1987. Minimum Sum of Diameters Clustering. Journal of Classification 4, 215–226.

Hartigan, J. A. 1975. Clustering Algorithms. New York: Wiley.

Heiser, W. J. 1988. Multidimensional Scaling with Least Absolute Residuals. In: H. H. Bock (ed.), Classification and Related Methods of Data Analysis, 455–462. North Holland: Elsevier.

Hempel, C. G., Oppenheim, P. 1936. Der Typusbegriff im Lichte der neuen Logik. Leiden: A. W. Sijthoff's.

Holtmann, D. 1975. Metrische multidimensionale Skalierung und ein „inhaltliches" Verfahren zur Bestimmung der Achsen. Zeitschrift für Soziologie 4, 248–253.

Homans, G. C. 1951. The Human Group. London: Routledge & Kegan Paul.

Höppner, F., Klawonn, F., Kruse, R. 1997. Fuzzy-Clusteranalyse. Braunschweig: Vieweg.

Hout, M. 1983. Mobility Tables. Newbury Park: Sage.

Hubert, L. J. 1974a. Some Applications of Graph Theory to Clustering. Psychometrika 39, 283–309.

Hubert, L. J. 1974b. Some Applications of Graph Theory and Related Non-Metric Techniques to Problems of Approximate Seriation: The Case of Symmetric Proximity Measures. British Journal of Mathematical & Statistical Psychology 27, 133–153.

Hubert, L. J. 1987. Assignment Methods in Combinatorial Data Analysis. New York: Marcel Dekker.

Hubert, L. J., Arabie, P. 1988. Relying on Necessary Conditions for Optimization: Unidimensional Scaling and Some Extensions. In: H. H. Bock (ed.), Classification and Related Methods of Data Analysis, 463–472. Amsterdam: Elsevier.

Hubert, L. J., Arabie, P., Hesson-Mcinnis, M. 1992. Multidimensional Scaling in the City-Block Metric: A Combinatorial Approach. Journal of Classification 9, 211–236.

Hubert, L. J., Arabie, P., Meulman, J. J. 2002. Linear Unidimensional Scaling in the L_2-Norm: Basic Optimization Methods Using MATLAB. Journal of Classification 19, 303–328

Hubert, L. J., Schultz, J. 1976. Quadratic Assignment as a General Data Analysis Strategy. British Journal of Mathematical and Statistical Psychology 29, 190–241.

Ihm, P. 1978. Statistik in der Archaeologie. Bonn: Rheinland-Verlag.

Jain, A. K., Dubes, R. C. 1988. Algorithms for Clustering Data. Englewood Cliffs: Prentice Hall.

Johnson, S. C. 1967. Hierarchical Clustering Schemes. Psychometrika 32, 241–254.

Jones, L. E., Koehly, L. M. 1993. Multidimensional Scaling. In: G. Keren, C. Lewis (eds.), A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues, 95–163. Hillsdale: Lawrence Erlbaum.

Kaufman, L., Rousseeuw, P. J. 1990. Finding Groups in Data. An Introduction to Cluster Analysis. New York: Wiley.

Kendall, D. G. 1971. Seiation from Abundance Matrices. In: F. R. Hodson, D. G. Kendall, P. Tautu (eds.), Mathematics in the Archeological and Historical Sciences, 215–252. Edinburgh: Edinburgh University Press.

Klemm, E. 1995. Das Problem der Distanzbindungen in der hierarchischen Clusteranalyse. Frankfurt: Peter Lang.

Kluge, S. 1999. Empirisch begründete Typenbildung. Zur Konstruktion von Typen und Typologien in der qualitativen Sozialforschung. Opladen: Leske + Budrich.

Kruskal, J. B. 1964a. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29, 1–27.

Kruskal, J. B. 1964b. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika 29, 115–129.

Kruskal, J. B., Wish, M. 1978. Multidimensional Scaling. London: Sage.

Lance, G. N., Williams, W. T. 1966. A Generalized Sorting Strategy for Computer Classifications. Nature 212, 218.

Lasch, R. 1996. Pyramidal Clustering Schemes. Statistical Papers 37, 235–251.

Lau, K., Leung, P. L., Tse, K. 1998. A Nonlinear Programming Approach to Metric Unidimensional Scaling. Journal of Classification 15, 3–14.

Lawick-Goodall, J. van 1971. In the Shadow of Man. London: W. Collins Sons & Co.

Laxton, R. H. 1997. Seriation in Archaeology: Modelling, Methods and Prior Information. In: R. Klar, O. Opitz (eds.), Classification and Knowledge Organization, 617–630. Berlin: Springer-Verlag.

Lazarsfeld, P. F. 1937. Some Remarks on the Typological Procedures in Social Research. Zeitschrift fuer Sozialforschung 6, 119–139.

Leeuw, J. de 1977. Correctness of Kruskal's Algorithms for Monotone Regression with Ties. Psychometrika 42, 141–144.

Lorr, M. 1983. Cluster Analysis for Social Scientists. San Francisco: Jossey-Bass.

Mardia, K. V., Kent, J. T., Bibby, J. M. 1979. Multivariate Analysis. New York: Academic Press.

Michaud, P. 1983. Opinions Aggregation. In: J. Janssen, J.-F. Marcotorchino, J.-M. Proth (eds.), New Trends in Data Analysis and Applications, 5–27. Amsterdam: North-Holland.

Milligan, G. W. 1979. Ultrametric Hierarchical Clustering Algorithms. Psychometrika 44 (1979), 343-346.

Mirkin, B. 1996. Mathematical Classification and Clustering. Dordrecht: Kluwer.

Nishisato, S. 1980. Analysis of Categorical Data: Dual Scaling and Its Applications. Toronto: University of Toronto Press.

Nishisato, S. 1994. Elements of Dual Scaling: An Introduction to Practical Data Analysis. Hillsdale, N.J.: Lawrence Erlbaum.

Pliner, V. M. 1984. A Class of Metric Scaling Models. Automation and Remote Control 45, 789–794.

Pliner, V. M. 1986. The Problem of Multivariate Metric Scaling. Automation and Remote Control 47, 560–567.

Pliner, V. M. 1996. Metric Unidimensional Scaling and Global Optimization. Journal of Classification 13, 3–18.

Rohwer, G., Pötter, U. 2001. Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Rohwer, G., Pötter, U. 2002a. Methoden sozialwissenschaftlicher Datenkonstruktion. Weinheim: Juventa.

Rohwer, G., Pötter, U. 2002b. Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung. Weinheim: Juventa.

Rubinfeld, D. L. 1982. Multiple Regression with a Qualitative Dependent Variable. Journal of Economics and Business 34, 67–78.

Schader, M., Tüshaus, U. 1988. Analysis of Qualitative Data: A Heuristic for Finding a Complete Preorder. In: H. H. Bock (ed.), Classification and Related Methods of Data Analysis, 341–346. Amsterdam: Elsevier.

Schriever, B. F. 1983. Scaling of Order Dependent Categorical Variables with Correspondence Analysis. International Statistical Review 51, 225–238.

Shepard, R. N. 1962. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, II. Psychometrika 27, 219–246.

Sneath, P. H. A., Sokal, R. R. 1973. Numerical Taxonomy. San Francisco: Freeman and Company.

Sodeur, W. 1974. Empirische Verfahren zur Klassifikation. Stuttgart: Teubner.

Späth, H. 1975. Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. München: Oldenbourg.

Späth, H. 1983. Cluster-Formation und -Analyse. München: Oldenbourg.

Späth, H. 1988. Homogeneous and Heterogeneous Clusters for Distance Matrices. In: H. H. Bock (ed.), Classification and Related Methods of Data Analysis, 157–164. Amsterdam: Elsevier.

Sriram, N., Lewis, S. 1993. Constructing Optimal Ultrametrics. Journal of Classification 10, 241–268.

Stinchcombe, A. L. 1968. Constructing Social Theories. New York: Harcourt, Brace & World.

Szczotka, F. A. 1972. On a Method of Ordering and Clustering of Objects. Zastosowania Matematyki (Applicationes Mathematicae) 13, 23–33.

Torgerson, W. S. 1952. Multidimensional Scalin: I. Theory and Method. Psychometrika 17, 401–419.

Torgerson, W. S. 1958. Theory and Methods of Scaling. New York: Wiley.

Trosset, M. W. 1993. Optimization Problems Associated with Multidimensional Scaling. Paper TR93-13, Department of Computational & Applied Mathematics, Houston: Rice University.

Trosset, M. W. 1997. Numerical Algorithms for Multidimensional Scaling. In: R. Klar, O. Opitz (eds.), Classification and Knowledge Organization, 80–92. Berlin: Springer-Verlag.

Tüshaus, U. 1983. Aggregation binärer Relationen in der qualitativen Datenanalyse. Königstein: Athenaeum-Hain-Hanstein.

Verdaasdonk, H. 2003. Valuation as Rational Decision-Making: A Critique of Bourdieu's Analysis of Cultural Value. Poetics 31, 357–374.

West, D. H. 1983. Algorithm 608. Approximative Solution of the Quadratic Assignment Problem. ACM Transactions on Mathematical Software 9, 461–466.

Young, F. W., Hamer, R. M. 1987. Multidimensional Scaling: History, Theory, and Applications, Hillsdale: Lawrence Erlbaum.

Young, G., Householder, A. S. 1938. Discussion of a Set of Points in Terms of Their Mutual Distances. Psychometrika 3, 19–22.

Ziegler, R. 1973. Typologien und Klassifikationen. In: G. Albrecht, H. Daheim, F. Sack (Hg.), Soziologie. Sprache, Bezug zur Praxis, Verhältnis zu anderen Wissenschaften, 11–47, Opladen: Westdeutscher Verlag.

Namenverzeichnis

Stichwortverzeichnis