# Nachbarschaften im semantischen Raum

Inauguraldissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
des Fachbereichs Sprach- und Literaturwissenschaften
der Universität Trier

vorgelegt von
Armin Wegner, M. A.

begutachtet von
Prof. Dr. Burghard Rieger und
Prof. Dr. Reinhard Köhler

überarbeitete Fassung

Trier, den 25. Juli 2006

# Danksagung

An dieser Stelle möchte ich all jenen Personen danken, die mich bei dieser Arbeit unterstützt haben. Zuallererst möchte ich meinem Doktorvater Herrn Prof. Dr. Burghard Rieger danken, der mir viel Geduld und Vertrauen entgegengebracht und mich in schwierigen Situationen mit seiner Diskussionsbereitschaft und seinen wertvollen Ratschlägen zum Weitermachen motiviert hat. Herrn Prof. Dr. Reinhard Köhler gilt mein Dank für die Übernahme des Zweitgutachtens und für seine unkomplizierte Unterstützung meines Promotionsverfahrens. Herrn Dr. Sven Naumann danke ich insbesondere für das Lemmatisieren und Taggen der Texte. Vor allem danke ich meinen Eltern, die es mir überhaupt ermöglicht haben, diese Arbeit zu verwirklichen und mich auf meinem Weg immer unterstützt haben.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Kapitel 1

# Einleitung

## 1.1 Fragestellung

Diese Arbeit untersucht die mathematischen Methoden und algorithmischen Verfahren der Clusteranalyse im Hinblick auf Bedeutungsrepräsentationen. Dazu wird der RIEGERsche Ansatz zur Abbildung semantischer Ähnlichkeiten von Wörtern in einem semiotischen Modell vorausgesetzt (RIEGER, 1989, S. 194 ff.). Sie basieren auf der Grundhypothese der empirischen Semantik, die besagt, daß der Gebrauch natürlichsprachlicher Zeichen in Texten zur Konstitution ihrer Bedeutung beiträgt und diese festigt. RIEGER modelliert Bedeutung sprachlicher Zeichen als einen zweistufigen Prozeß zunehmender kombinatorischer Einschränkung jeweils noch vorhandener Wahlmöglichkeiten in einem Vektorraummodell zweiter Ordnung. Diesen Einschränkungen entsprechen die linguistischen Grundrelationen der syntagmatischen und paradigmatischen Beziehungen von sprachlichen Ausdrücken. In der ersten Stufe werden die syntagmatischen Assoziationen von Wörtern aus den Regularitäten ihrer Kookkurrenzen in Texten eines Korpus bestimmt. Auf den syntagmatischen Assoziationen aufbauend werden in der zweiten Stufe die paradigmatischen Assoziationen aufgrund des Vergleichs der Kotexte der Vorkommen der Wörter in den Texten ermittelt. Der Prozeß als zeitlicher Ablauf dieser Prozedur berechnet zu jedem Wort einen Punkt in einem hochdimensionalen Vektorraum. Die Punkte heißen *Bedeutungspunkte*, weil deren räumliche Lage zueinander topologische Nachbarschaften bilden und deren Nähe Aufschluß über die Bedeutungsähnlichkeiten der durch sie repräsentierten Wörter geben. Die Menge der Bedeutungspunkte heißt *semantischer Raum*. Dabei ist die Bedeutung eines

Wortes nicht isoliert beschreibbar, sondern wird – wie durch die Koordinaten eines Punktes – durch die Beziehungen zu allen anderen Wörtern bestimmt.

Ziel dieser Arbeit ist es, im Rahmen der deskriptiven und explorativen Datenanalyse Verfahren zu erproben und Bedingungen zu ermitteln, durch die aus einem Korpus ein semantischer Raum berechnet werden kann, dessen Bedeutungspunkte sich in Clustern sammeln, deren Elemente auch intuitiv ähnliche Bedeutungen repräsentieren. Als erstes ist zu untersuchen, welche Wörter oder Texte eines Korpus sich dazu überhaupt eignen und welchen Bedingungen sie genügen müssen. Die nächste wesentliche Frage ist, mit welchen Maßen die syntagmatischen und paradigmatischen Assoziationen der zweistufigen Prozedur berechnet werden können und ob die Maße auch im Sinne der Untersuchung interpretierbar sind. Zuletzt werden die Beziehungen zwischen den Bedeutungspunkten clusteranalytisch untersucht. Dabei wird davon ausgegangen, daß Gruppen von Bedeutungspunkten sich nicht scharf voneinander abgrenzen, sondern fließend ineinander übergehen. Gegenstand der Clusteranalyse sind semantische Räume. Eine bekannte Schwäche der Clusteranalyse im allgemeinen ist die große Anzahl frei wählbarer Parameter und der Einfluß, den jede Wahl eines der bekannten clusteranalytischen Verfahren in Bezug auf die vorauszusetzenden Vorkenntnisse von der Struktur der zu untersuchenden Daten auf die Güte der erwartbaren Ergebnisse hat. Diese generelle Problematik belastet die Abschätzbarkeit von Erfolg und Adäquatheit unüberwachter Klassifikationsverfahren weit über die quantitativ-linguistischen Untersuchungen in der Gebrauchssemantik hinaus. Deshalb wird ein neues Verfahren entwickelt, welches den analysierten Daten in geringerem Maße als bisher Strukturen aufprägt und in höherem Maße als bisher von den analysierten Daten und ihren Strukturen gesteuert wird. Ein weiteres, allgemein leider oft vernachlässigtes Ziel ist es, konkrete, inhaltlich sinnvolle und formal theoretisch abgegrenzte Wertebereiche für die Variablen und Parameter der diskutierten und erprobten Modelle zu bestimmen.

Ausgangspunkt der Untersuchung sind Texte, die wirkliche Sprecher/Schreiber in konkreten Situationen mit kommunikativer Intention produziert haben, und die als solche von wirklichen Hörern/Lesern erkannt und verstanden werden können. Wenn solche Texte in ähnliche Kotexte eingebettet sind, heißen sie *pragmatisch homogen.* Es werden Korpora aus Zeitungsartikeln eines Ressorts verwendet, weil sie die für die Untersuchungen notwendige pragmatische Homogenität aufweisen. Diese enthalten orthographische und

grammatische Fehler, Setzfehler und für den Inhalt nicht notwendige zusätzliche Angaben wie etwa Agentur, Ort und Autor. Um die Wortarten zu bestimmen, werden die Texte maschinell getaggt, phrasenweise geparst und lemmatisiert. Das ist bei performativen Sprachdaten, die durch wirklichen Sprachgebrauch entstanden sind und nicht etwa zur Absicherung des Erzielens eines gewünschten Untersuchungsergebnisses ausschließlich künstlich und frei von problematischen Strukturen erzeugt wurden, zumindest automatisch noch nicht fehlerfrei möglich. Die Analyse muß auch bezüglich solcher Fehler in den Daten stabil sein.

## 1.2 Übersicht

Zunächst werden in einem mathematischen Repetitorium (Kapitel 2) wichtige Grundbegriffe vorgestellt und einige höhere Konzepte eingeführt. Dabei bleibt es nicht bei vagen Ideen, sondern es werden auch die zugehörigen Formeln vollständig und nachvollziehbar beschrieben. Wenn auch die eine oder andere Definition etwas unvermittelt erscheinen mag, wird es sich in den nachfolgenden Kapiteln doch als nützlich erweisen, vorher eine abstrakte und allgemeingültige Definition gegeben zu haben, auf die aus unterschiedlichsten Zusammenhängen heraus referiert werden kann. Das Hauptaugenmerk wurde dabei auf die Ähnlichkeits-, Distanz- und Assoziationsmaße gerichtet.

In Kapitel 3 werden nach einem Überblick über die unüberwachte Klassifikation durch Clusteranalyseverfahren und eine für die Anwendung von Clusterverfahren zwingend notwendige Formalisierung und grundlegende mathematische Modellierung die wichtigsten hierarchisch-agglomerativen Algorithmen vorgestellt. Mit dem am weitesten verbreiteten Partitionierungsverfahren, dem $c$-Means, wird zu den klassischen unscharfen Methoden übergeleitet. Das Mountain-Clustering bestimmt geeignete Prototypen aufgrund unterschiedlicher Punktdichten der Datenmenge für die Initialisierung von Verfahren zur Verbesserung einer Anfangspartition. Darauf aufbauend wird ein neu entwickeltes, sich allein auf $k$-Nachbarschaften beschränkendes, agglomeratives Verfahren vorgestellt. Seine interne Stoppbedingung ermittelt als lokales Optimalitätskriterium nicht nur eine Partition, sondern damit gleichzeitig auch die Clusteranzahl und die Clusterprototypen automatisch.

In Kapitel 4 werden Möglichkeiten zur Gewichtung der Terme anhand ihrer Vorkommenshäufigkeiten in den Texten des Korpus vorgestellt und an die Untersuchung von Termähnlichkeiten angepaßt. Principal-Component-

Analysis und darauf aufbauendes Latent-Semantic-Indexing werden ausführlich beschrieben und formalisiert. Im Vektorraummodell zweiter Ordnung werden die so veränderten Termvektoren auf Bedeutungspunkte eines semantischen Raumes abgebildet. Dazu werden inhaltlich interpretierbare und plausible Assoziationsmaße speziell für die Berechnung von syntagmatischen und paradigmatischen Ähnlichkeiten von Termen erklärt und entwickelt.

Zuletzt werden in Kapitel 5 die Bedeutungen ausgewählter Wörter aufgrund ihrer Kookkurrenzen in pragmatisch homogenen Korpora aus Zeitungstexten praktisch ermittelt und mittels clusteranalytischer Verfahren diejenigen Wörter zusammengefaßt, deren Verwendung ähnlich ist. Die berechneten Werte werden exemplarisch dargelegt, erläutert und erklärt. Dazu sind zunächst Wörter als Untersuchungseinheiten und die aus ihnen gebildeten Texte als Merkmale auszuwählen und die Merkmalsausprägungen entsprechend ihrer Verteilungen im Korpus zu gewichten. Die Assoziationen der Wörter werden aufgrund dieser Gewichtsvektoren im zwei Stufen umfassenden Vektorraummodell zunächst auf ihre syntagmatischen und dann auf ihre paradigmatischen Ähnlichkeiten untersucht. Diejenigen Kombinationen von Ähnlichkeitsmaßen erster und zweiter Stufe werden gesucht, die nicht nur theoretisch interpretierbar und plausibel sind, sondern mit denen aus den konkreten Daten auch semantische Räume praktisch berechnet werden können, welche die Bedeutungsähnlichkeiten der Wörter geeignet darstellen. Für diese wiederum werden geeignete clusteranalytische Verfahren vorgestellt, die die Wörter aufgrund dieser Ähnlichkeiten in Gruppen zusammengehöriger Wörter einteilen.

Um die in den beiden vorangegangenen Absätzen genannten Aufgaben überhaupt angehen zu können, war vom Autor erheblicher Aufwand beim Programmieren und Testen der erforderlichen Software selbst zu leisten. Dies war notwendig, da bereits verfügbare Software die umfangreichen Datenmengen nicht oder nur unvollständig und manchmal sogar fehlerhaft verarbeitet hat. Zudem kann für wissenschaftliche Zwecke keine Software eingesetzt werden, von der aufgrund mangelhafter Beschreibungen nicht bekannt ist, welche Algorithmen oder Formeln sie genau benutzt, da dadurch die Möglichkeit einer nachvollziehbaren Interpretation der Ergebnisse von vornherein ausgeschlossen ist. Bei der sich daraus ergebenden Notwendigkeit von Neuimplementationen[1]

[1] Es ist beabsichtigt, die Quelltexte der entwickelten Programme zu veröffentlichen, so daß sie auch anderen bei ihrer wissenschaftlichen Arbeit von Nutzen sein mögen. Immerhin war die Entwicklung neuer Software notwendig, weil benötigte Programme gar nicht vorhanden waren und bereits verfügbare Software, die Daten nicht wie gefordert zu verarbeiten vermochte. Die

hat sich darüber hinaus vielfach herausgestellt, daß die Literatur oft nur die grundlegenden Ideen und Konzepte vorstellt, die Modelle nur unzureichend formalisiert, Algorithmen unvollständig angibt und Extrem- und Sonderfälle nicht behandelt. Die erstellten Programme arbeiten mit einem sehr einfachen Datenformat, welches erlaubt, die Daten problemlos einzusehen und zu verändern. Es erleichtert auch den Austausch von Daten zwischen den neu entwickelten und den bereits vorhandenen Programmen.

---

Weitergabe der Quelltexte ermöglicht es, die genaue Arbeitsweise der Algorithmen einzusehen, die Programme an ähnliche Bedürfnisse anzupassen und weiterzuentwickeln.

# Kapitel 2

# Mathematische Vorbemerkungen

## 2.1 Einige nützliche Funktionen und Notationen

### 2.1.1 Indikatorfunktion

Es sei $X$ eine beliebige Menge von Elementen einer Grundgesamtheit $\Omega$. Die auf $\Omega$ erklärte Funktion

$$\mathbf{1}_X \colon \Omega \to \{0,1\}, \quad \omega \mapsto \mathbf{1}_X(\omega) = \begin{cases} 0, & \text{falls } \omega \notin X \\ 1, & \text{falls } \omega \in X \end{cases}$$

heißt Indikatorfunktion von $X$. Sie zeigt mit 1 an, wenn ein Element $\omega$ in einer Teilmenge $X$ von $\Omega$ liegt und mit 0, wenn es darin nicht enthalten ist. Die Indikatorfunktion wird auch charakteristische Funktion einer Menge genannt und dann meist mit $\chi$ bezeichnet.

### 2.1.2 Vorzeichenfunktion

Die Vorzeichenfunktion (Signumfunktion) bildet aus der Menge der reellen Zahlen in die Menge $\{-1,0,1\}$ ab und ist wie folgt definiert:

$$\operatorname{sgn}(x) \colon \mathbb{R} \to \{-1,0,1\}, \quad x \mapsto \operatorname{sgn}(x) = \begin{cases} -1, & \text{falls } x < 0 \\ 0, & \text{falls } x = 0 \\ +1, & \text{falls } x > 0 \end{cases}.$$

Ihr Wert ist $-1$ für eine negative Zahl, $+1$ für eine positive Zahl und 0 für 0 im Argument.

### 2.1.3 Gauss-Klammer

Die GAUSS-Klammer (Ganzzahlfunktion, Abrundungsfunktion, *floor function*)

$$[\,\cdot\,] : \mathbb{R} \to \mathbb{Z}, \quad x \mapsto [x] = \max\{z \in \mathbb{Z} : z \leq x\}$$

rundet eine reelle Zahl auf die nächste ganze Zahl ab. Die GAUSS-Klammer $[x]$ einer reellen Zahl $x$ ist diejenige ganze Zahl, für welche $x - 1 < [x] \leq x$ gilt. Damit ist $[x]$ die kleinste ganze Zahl echt größer als $x - 1$ und die größte ganze Zahl kleiner oder gleich $x$.

### 2.1.4 Kronecker-Symbol

Die für alle Paare $(i, j)$ natürlicher Zahlen durch

$$\delta_{ij} := \begin{cases} 0, & \text{falls } i \neq j \\ 1, & \text{falls } i = j \end{cases}$$

erklärte Funktion heißt KRONECKER-Symbol. Sie ist 1, wenn beide Indizes gleich sind und 0 andernfalls.

## 2.2 Vektorräume

### 2.2.1 Gruppe

Eine Menge $G$ versehen mit einer Abbildung (Verknüpfung, Komposition)

$$*\colon G \times G \to G, \quad (x, y) \mapsto x * y$$

heißt eine Gruppe, wenn die folgenden Axiome erfüllt sind:

(G1) Für alle $x, y, z \in G$ gilt $(x * y) * z = x * (y * z)$ (Assoziativität).

(G2) Es gibt ein neutrales Element $e \in G$ mit $e * x = x$ für alle $x \in G$.

(G3) Zu jedem Element $x \in G$ gibt es ein inverses Element $x^{-1}$ mit $x^{-1} * x = e$.

Gilt zusätzlich

(G4) Für alle $x, y \in G$ gilt $x * y = y * x$ (Kommutativität).

heißt die Gruppe kommutativ (abelsch). Um deutlich zu machen, welche so definierte Abbildung aus einer Menge $G$ eine Gruppe macht, wird die Gruppe als Paar $(G, *)$ geschrieben. Vgl. FISCHER (1989, S. 31 ff.).

**Beispiele**

**1.** Die Menge $\mathbb{R}$ der reellen Zahlen mit der gewöhnlichen Addition $+$ als Verknüpfung ist eine additiv geschriebene kommutative Gruppe $(\mathbb{R}, +)$. Bei einer additiv geschriebenen kommutativen Gruppe heißt das neutrale Element auch Nullelement und die inversen Elemente auch negative Elemente. Das Nullelement ist $0 \in \mathbb{R}$. Das negative Element zu $x \in \mathbb{R}$ ist $-x \in \mathbb{R}$.

**2.** Die Menge $\mathbb{R} \setminus \{0\}$ der reellen Zahlen vermindert um die Null und versehen mit der üblichen Multiplikation $\cdot$ ist eine multiplikativ geschriebene kommutative Gruppe $(\mathbb{R}, \cdot)$. Das neutrale Element einer multiplikativ geschriebenen kommutativen Gruppe nennt heißt auch Einselement und die inversen Elemente auch reziproke Elemente. Das Einselement ist $1 \in \mathbb{R}$. Das reziproke Element zu $x \in \mathbb{R}$ ist $1/x \in \mathbb{R}$.

### 2.2.2 Körper

Eine Menge $K$ versehen mit einer Addition

$$+ \colon K \times K \to K, \quad (x, y) \mapsto x + y$$

und einer Multiplikation

$$\cdot \colon K \times K \to K, \quad (x, y) \mapsto x \cdot y$$

als innere Verknüpfungen heißt Körper, wenn die folgenden Axiome erfüllt sind:

(K1) $(K, +)$ ist eine kommutative Gruppe.

Ihr neutrales Element wird mit 0 und das zu $x \in K$ inverse Element mit $-x$ bezeichnet.

(K2) Die auf $K \setminus \{0\}$ beschränkte Multiplikation ist wohldefiniert, weil $x \cdot y \neq 0$ für alle $x, y \in K \setminus \{0\}$ gilt. $(K \setminus \{0\}, \cdot)$ ist eine kommutative Gruppe.

Ihr neutrales Element wird mit 1 und das zu $x \in K$ inverse Element mit $x^{-1}$ bezeichnet.

(K3) Für alle $x, y, z \in K$ gelten

$$x \cdot (y + z) = (x \cdot y) + (x \cdot z) \quad \text{und}$$
$$(x + y) \cdot z = (x \cdot z) + (y \cdot z)$$

(Distributivität).

Für $x \cdot y$ schreibt man meist auch nur $xy$. Ein Menge $K$ und die auf ihr so definierten Operationen der Addition $+$ und der Multiplikation $\cdot$ werden als Tripel $(K, +, \cdot)$ geschrieben. Vgl. Fischer (1989, S. 33 ff.).

**Beispiel**

Die Menge $\mathbb{R}$ der reellen Zahlen mit der üblichen Addition $+$ und Multiplikation $\cdot$ ist ein Körper $(\mathbb{R}, +, \cdot)$.

### 2.2.3 Vektorraum

Sei $K$ ein Körper. Eine Menge $V$ zusammen mit einer Addition

$$+ \colon V \times V \to V, \quad (x, y) \mapsto x + y$$

und einer äußeren Multiplikation (Skalarmultiplikation)

$$\cdot \colon K \times V \to V, \quad (\lambda, x) \mapsto \lambda \cdot x$$

heißt $K$-Vektorraum (oder Vektorraum über $K$), wenn die folgenden Axiome gelten:

(V1) $(V, +)$ ist eine kommutative Gruppe.

(V2) Für alle $x, y \in V$ und alle $\lambda, \mu \in K$ gelten die Distributivgesetze

$$(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x,$$
$$\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$$

und das Assoziativgesetz

$$(\lambda\mu) \cdot x = \lambda \cdot (\mu \cdot x)$$

und die 1 auf $K$ ist neutral,

$$1 \cdot x = x.$$

Die Elemente eines Vektorraums heißen Vektoren. Die Elemente des zugehörigen Körpers heißen Skalare und der Körper selbst Skalarenkörper. $0 \in V$ heißt Nullvektor. Statt $\lambda \cdot x$ wird meist kurz $\lambda x$ geschrieben. Die Addition im Vektorraum und die im Körper binden weniger stark als die Multiplikation mit Skalaren.

Eine nichtleere Teilmenge $W$ eines $K$-Vektorraums $V$ wird Untervektorraum von $V$ genannt, wenn sie abgeschlossen bezüglich der in $V$ definierten Addition und Multiplikation mit Skalaren ist. Sie ist dann selbst ein $K$-Vektorraum zusammen mit den aus $V$ induzierten Operationen. Vgl. FISCHER (1989, S. 36 ff.).

**Beispiele**

**1.** Die Menge $\mathbb{R}$ der reellen Zahlen ist ein Vektorraum über sich selbst bezüglich der üblichen Addition und Multiplikation.

**2.** Das $n$-fache Kreuzprodukt $\mathbb{R} \times \cdots \times \mathbb{R} = \mathbb{R}^n$ der Menge der reellen Zahlen ist ein Vektorraum über $\mathbb{R}$ als Skalarenkörper bezüglich der komponentenweisen Operationen im Körper $K$.

**3.** Es sei $K$ ein Körper und $m, n \in \mathbb{N}$. Die Menge $K^{m \times n} := K^{\{1,2,\ldots,m\} \times \{1,2,\ldots,n\}}$ aller $m \times n$-Matrizen mit Komponenten aus $K$ ist ein $K$-Vektorraum bezüglich der komponentenweisen Addition und Skalarmultiplikation.

### 2.2.4 Linearkombination

Es sei $V$ in $K$-Vektorraum, $n \in \mathbb{N}$, $\lambda_1, \lambda_2, \ldots, \lambda_n \in K$ und $x_1, x_2, \ldots, x_n \in V$. Dann heißt die Summe der skalaren Vielfachen dieser Vektoren,

$$\sum_{i=1}^{n} \lambda_i x_i \in V,$$

eine Linearkombination der Vektoren $x_1, x_2, \ldots, x_n$.

### 2.2.5 Lineare Abhängigkeit

Eine endliche Teilmenge $E$ eines Vektorraums $V$ heißt linear unabhängig, wenn sich die Elemente von $E$ nur trivial linear zu 0 kombinieren lassen,

$$\underset{(\lambda_x)_{x\in E} \in K^E}{\forall} \left( \sum_{x\in E} \lambda_x x = 0 \quad \Rightarrow \quad \underset{x\in E}{\forall} \lambda_x = 0 \right).$$

Sie heißt linear abhängig, wenn sie nicht linear unabhängig ist. Dann gibt es von 0 verschiedene Skalare mit denen die Elemente von $E$ linear zum Nullvektor kombiniert werden können,

$$\underset{(\lambda_x)_{x\in E} \in K^E}{\exists} \left( (\lambda_x)_{x\in E} \neq (0)_{x\in E} \text{ und } \sum_{x\in E} \lambda_x x = 0 \right).$$

Eine beliebige Teilmenge $M \subset V$ heißt linear unabhängig, wenn jede endliche Teilmenge von $M$ linear unabhängig ist.

### 2.2.6 Lineare Hülle

Es sei $V$ ein $K$-Vektorraum und $M$ eine nichtleere Teilmenge von $V$. Die Menge aller Vektoren aus $V$, die sich durch Linearkombination von Elementen aus $M$ erzeugen lassen,

$$\mathcal{L}(M) := \{x \in V : \underset{n\in\mathbb{N}}{\exists} \underset{\lambda_1,\lambda_2,\ldots,\lambda_n \in K}{\exists} \underset{x_1,x_2,\ldots,x_n \in M}{\exists} x = \sum_{i=1}^{n} \lambda_i x_i\},$$

heißt lineare Hülle (linearer Spann) von $M$. Sie ist der kleinste Untervektorraum von $V$, der $M$ enthält.

### 2.2.7 Erzeugendensystem

Eine Teilmenge $M$ eines Vektorraums $V$ heißt Erzeugendensystem von $V$, wenn ihre Lineare Hülle $\mathcal{L}(M)$ den Vektorraum aufspannt, $\mathcal{L}(M) = V$.

### 2.2.8 Basis

Eine Teilmenge $B$ eines $K$-Vektorraums $V$ heißt Basis, wenn sie linear unabhängig ist und den Vektorraum aufspannt. Jeder Vektor ist eindeutig durch eine Linearkombination endlich vieler Basisvektoren beschrieben,

$$\underset{x \in V}{\forall} \quad \underset{(\alpha_b)_{b \in B} \in K^B}{\exists_1} \quad E := \{b \in B : \alpha_b \neq 0\} \text{ endlich} \quad \text{und} \quad x = \sum_{b \in E} \alpha_b \, b.$$

Ein Vektorraum kann viele verschiedene Basen haben.

**Beispiele**

**1.** Für jedes $1 \leq i \leq n$, $n \in \mathbb{N}$, sei $e_i := (\delta_{ij})_{1 \leq j \leq n}$ der $n$-elementige Vektor in $\mathbb{R}^n$, dessen $i$-te Komponente 1 und alle anderen 0 sind. Dann ist $\{e_i : 1 \leq i \leq n\}$ die sogenannte Standardbasis (kanonische Basis) des $\mathbb{R}$-Vektorraums $\mathbb{R}^n$. Ihre Elemente heißen Einheitsvektoren.

**2.** Es sei $K$ ein Körper und $m, n \in \mathbb{N}$. Die Matrizen $E_{s,t} := (\delta_{is}\delta_{kt})_{1 \leq i \leq m,\, 1 \leq k \leq n}$ für alle $1 \leq s \leq m$ und $1 \leq t \leq n$ bilden eine $mn$-elementige Basis des Vektorraums $K^{m \times n}$. Die Komponente mit dem Index $(s, t)$ von $E_{s,t}$ ist 1. Alle anderen Komponenten sind 0.

### 2.2.9 Dimension

Die Dimension eines Vektorraumes $V$, geschrieben $\dim(V)$, ist die Anzahl der Elemente einer Basis. Alle Basen eines Vektorraums sind gleich groß.

**Beispiele**

**1.** Die Dimension des $n$-dimensionalen reellen Vektorraums $\mathbb{R}^n$ ist $\dim(\mathbb{R}^n) = n$, $n \in \mathbb{N}$.

**2.** Die einzige nichtleere Teilmenge des Nullvektorraums $\{0\}$ ist der Vektorraum selbst. Die Teilmenge $\{0\}$ ist linear abhängig und kann folglich keine Basis sein. Deshalb definiert man die Dimension des Vektorraums, der nur aus dem Nullvektor besteht, als 0, was äquivalent dazu ist, die leere Menge als Basis zu definieren.

### 2.2.10 Koordinaten

Bilden die Vektoren $b_1, b_2, \ldots, b_n$ eine Basis $B$ des Vektorraumes $V$ über dem Körper $K$, so heißen die eindeutig bestimmten Skalare $\lambda_1, \lambda_2, \ldots, \lambda_n \in K$ in der Linearkombination

$$v := \sum_{i=1}^{n} \lambda_i b_i$$

Koordinaten und der aus ihnen gebildete Vektor

$$[v]_B := (\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Koordinatenvektor von $v \in V$ bezüglich der Basis $B$. Der lineare Isomorphismus $[\cdot]_B : V \to K^n$ ordnet jedem Vektor $v \in V$ seinen Koordinatenvektor bezüglich der Basis $B$ zu. Er heißt Koordinatenabbildung von $V$ bezüglich $B$. Der Wahl einer Basis entspricht die Wahl eines Isomorphismusses $V \to K^n$. Die Umkehrabbildung $[\cdot]_B^{-1}$ heißt Koordinatensystem in $V$. Vgl. Fischer (1989, S. 76 f.), Lay (1997, S. 240 f.).

**Beispiele**

**1.** Die Koordinaten eines Vektors $x$ in $\mathbb{R}^n$ bezüglich der Standardbasis $S$ sind gerade dessen Komponenten, $[x]_S = x$.

**2.** Der Koordinatenvektor des $i$-ten Basisvektors $b_i$ einer $n$-elementigen Basis $B$ ist $[b_i]_B = (\delta_{ij})_{1 \le j \le n}$.

## 2.3 Matrizen

Ein rechteckiges Schema

$$A := (a_{ij})_{1 \le i \le m,\, 1 \le j \le n} := \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

von Elementen aus einem Körper $K$ heißt eine Matrix mit $m$ Zeilen und $n$ Spalten. Die Elemente $a_{ij}$ heißen Komponenten der Matrix (Fischer, 1989, S. 50). Die Matrix, die nur aus den Zeilen $k$ bis $l$ und den Spalten $s$ bis $t$ von $A$ besteht, wird

$A(k{:}l, s{:}t) := (a_{ij})_{k \le i \le l,\, s \le j \le t}$ geschrieben. Wird ein Bereich nicht eingeschränkt, läßt man die Grenzen weg und notiert nur den Doppelpunkt.

### 2.3.1 Spur

Die Spur einer quadratischen Matrix $A := (a_{ij})_{1 \le i,j \le n}$ ist die Summe ihrer Diagonalelemente,

$$\operatorname{Spur}(A) := \sum_{i=1}^{n} a_{ii}.$$

## 2.4 Lineare Abbildungen

### 2.4.1 Lineare Abbildung

Es seien $V$, $W$ Vektorräume über demselben Körper $K$. Eine Abbildung $f\colon V \to W$ heißt linear, wenn sie homogen,

$$\underset{\lambda \in K}{\forall} \quad \underset{x \in V}{\forall} \quad f(\lambda x) = \lambda f(x), \tag{2.1}$$

und additiv,

$$\underset{x, y \in V}{\forall} \quad f(x + y) = f(x) + f(y), \tag{2.2}$$

ist. Bei linearen Abbildungen ist es unerheblich, ob zwei Vektoren erst addiert und dann ihre Summe abgebildet oder erst die Vektoren abgebildet und dann die Bilder addiert werden (2.2). Gleiches gilt für die Multiplikation eines Vektors mit einem Skalar (2.1).

### 2.4.2 Matrix einer linearen Abbildung

Jede lineare Abbildung aus einem endlich-dimensionalen Vektorraum in einen weiteren kann durch eine Matrix dargestellt werden. Eine lineare Abbildung ist durch die Bilder einer Basis eindeutig festgelegt. Es sei $f$ eine lineare Abbildung aus einem $n$-dimensionalen $K$-Vektorraum $V$ in einen $m$-dimensionalen $K$-Vektorraum $W$, $\mathcal{V} := \{v_1, v_2, \ldots, v_n\}$ eine Basis von $V$ und

$\mathcal{W} := \{w_1, w_n, \ldots, w_m\}$ eine Basis von $W$. Dann kann jeder Vektor $f(v_j) \in W$ eindeutig als Linearkombination der Basisvektoren aus $\mathcal{W}$ ausgedrückt werden,

$$\underset{1 \leq j \leq n}{\forall} \quad \underset{a_{1j},\, a_{2j},\, \ldots,\, a_{mj} \in K}{\exists_1} \quad f(v_j) = \sum_{i=1}^{m} a_{ij}\, w_i.$$

Die $m \times n$-Matrix $A := (a_{ij})_{1 \leq i \leq m,\, 1 \leq j \leq n}$ stellt die Abbildung $f$ bezüglich der Basis $\mathcal{V}$ in der Quelle und der Basis $\mathcal{W}$ im Ziel dar. Die Zeilenanzahl von $A$ ist die Dimension des Ziels und die Spaltenanzahl die der Quelle. Die $j$-te Spalte von $A$ ist der Koordinatenvektor von $f(v_j)$ bezüglich der Basis $\mathcal{W}$, also $A = ([f(v_1)]_{\mathcal{W}}, [f(v_2)]_{\mathcal{W}}, \ldots, [f(v_n)]_{\mathcal{W}})$. Es gilt für alle $v \in V$

$$[f(v)]_{\mathcal{W}} = A[v]_{\mathcal{V}}.$$

(Fischer, 1989, S. 78 f.)

**Beispiel**

Es sei $f : \mathbb{R}^n \to \mathbb{R}^m$ eine lineare Abbildung, $S_n$ die Standardbasis mit den Einheitsvektoren $e_1, e_2, \ldots, e_n$ in $\mathbb{R}^n$ und $S_m$ die Standardbasis in $\mathbb{R}^m$. Die Spalten der $m \times n$-Matrix $A$ seien die Bilder der Einheitsvektoren, $A = (f(e_1), f(e_2), \ldots, f(e_n))$. Dann gilt $f(x) = [f(x)]_{S_m} = A[x]_{S_n} = Ax$ für alle $x \in \mathbb{R}^n$.

### 2.4.3 Inverse Matrix

Eine quadratische Matrix $A$ wird invertierbar oder regulär genannt, wenn es eine Matrix $B$ gibt, so daß $AB = BA = E$ ist. Eine solche Matrix ist eindeutig bestimmt. Sie wird mit $A^{-1}$ bezeichnet und inverse Matrix zu $A$ genannt. Eine nicht invertierbare Matrix heißt auch singulär. Die Inverse der Inversen einer regulären Matrix $A$ ist die Matrix selbst, $(A^{-1})^{-1} = A$. Die Inverse des Produktes zweier invertierbarer Matrizen $A$ und $B$ ist das Produkt ihrer Inversen in umgekehrter Reihenfolge, $(AB)^{-1} = B^{-1}A^{-1}$ (Fischer, 1989, S. 87).

$$\begin{tikzcd} V \arrow[r, "\mathrm{id}_V"] \arrow[d, "{[\cdot]_{\mathcal{A}}}"'] & V \arrow[d, "{[\cdot]_{\mathcal{B}}}"] \\ K^n \arrow[r, "W"'] & K^n \end{tikzcd}$$

**Abbildung 2.1:** Koordinatentransformation des Basiswechsels von $\mathcal{A}$ nach $\mathcal{B}$

### 2.4.4 Basiswechsel und Koordinatentransformation

Es sei $V$ ein endlich-dimensionaler $K$-Vektorraum mit den Basen $\mathcal{A} := \{a_1, a_2, \ldots, a_n\}$ und $\mathcal{B} := \{b_1, b_2, \ldots, b_n\}$. Jeder Basisvektor aus $\mathcal{A}$ läßt sich eindeutig durch eine Linearkombination der Basisvektoren aus $\mathcal{B}$ darstellen,

$$\underset{1 \le j \le n}{\forall} \quad \underset{w_{1j},\, w_{2j},\, \ldots,\, w_{nj} \in K}{\exists_1} \quad a_j = \sum_{i=1}^{n} w_{ij}\, b_i.$$

Die Matrix $W := (w_{ij})_{1 \le i,j \le n}$ heißt Koordinatentransformationsmatrix des Basiswechsels von $\mathcal{A}$ nach $\mathcal{B}$. Die $j$-te Spalte von $W$ ist der Koordinatenvektor von $a_j$ bezüglich der Basis $\mathcal{B}$, $W = ([a_1]_{\mathcal{B}}, [a_2]_{\mathcal{B}}, \ldots, [a_n]_{\mathcal{B}})$. Die Koordinatentransformationsmatrix des Übergangs von $\mathcal{A}$ nach $\mathcal{B}$ ist die Matrix, die die Identitätsabbildung $\mathrm{id}_V$ in $V$ bezüglich der Basen $\mathcal{A}$ in der Quelle und $\mathcal{B}$ im Ziel darstellt (Abschnitt 2.4.2). Der lineare Operator $W$ bildet den Koordinatenvektor $[v]_{\mathcal{A}}$ eines Vektors $v \in V$ bezüglich $\mathcal{A}$ auf den Koordinatenvektor $[v]_{\mathcal{B}}$ von $v$ bezüglich $\mathcal{B}$ ab,

$$[v]_{\mathcal{B}} = [\mathrm{id}_V(v)]_{\mathcal{B}} = W\,[v]_{\mathcal{A}}$$

(Abbildung 2.1). Jede Koordinatentransformationsmatrix ist regulär. Die Inverse der Koordinatentransformationsmatrix heißt Basiswechselmatrix. Sie bildet die Basisvektoren von $\mathcal{A}$ auf die Basisvektoren von $\mathcal{B}$ ab,

$$\underset{1 \le j \le n}{\forall} \quad b_j = W^{-1} a_j.$$

Es sei nun $A := [a_1, a_2, \ldots, a_n]$ die Matrix, deren Spalten die Basisvektoren von $\mathcal{A}$ sind, und $B := [b_1, b_2, \ldots, b_n]$ die Matrix, deren Spalten die Basisvektoren von $\mathcal{B}$ sind. Somit ist $B = W^{-1}A$. Die Basiswechselmatrix ist $W^{-1} = BA^{-1}$ und die Koordinatenwechselmatrix ist $W = AB^{-1}$. Die Transformationsmatrix des umgekehrten Basiswechsels ist deren Inverse. Umgekehrt ist jede reguläre

Matrix die Matrix eines geeigneten Basiswechsels. Vgl. FISCHER (1989, S. 88 f.) und LIPSCHUTZ (1977, S. 153).

### 2.4.5 Kern einer linearen Abbildung

Es seien $V$ und $W$ Vektorräume über demselben Körper $K$ und $f\colon V \to W$ eine lineare Abbildung. Die Menge der Elemente in $V$, die auf die 0 in $W$ abgebildet werden, heißt Kern von $f$,

$$\mathrm{Kern}(f) := \overset{-1}{f}(0) = \{v \in V : f(v) = 0\}.$$

Der Kern von $f$ ist ein Untervektorraum von $V$.

### 2.4.6 Determinante

Es sei $K$ ein Körper, $n \in \mathbb{N}$. Dann ist $K^{n\times n}$ die Menge aller quadratischen $n \times n$-Matrizen mit Komponenten aus $K$. Eine Abbildung

$$\det\colon K^{n\times n} \to K$$

heißt Determinantenabbildung, wenn folgendes gilt:

(D1) det ist linear in jeder Zeile bei festgehaltenen anderen Zeilen.

(D2) det ist alternierend, das heißt, hat eine Matrix zwei gleiche Zeilen, wird ihr 0 zugeordnet.

(D3) det ist normiert, das heißt $\det(E) = 1$.

Für jedes $n \in \mathbb{N}$ gibt es genau eine Determinantenabbildung, die sogenannte Determinante.

## 2.5 Eigenwerte

### 2.5.1 Transponierte Matrix

Die Transponierte $A^T$ einer Matrix $A$ ist die Matrix, die sich ergibt, wenn die Spalten als Zeilen geschrieben werden. Ist $A = (a_{ij})_{1\le i\le m,\, 1\le j\le n}$ eine $m\times n$-Matrix, so ist ihre Transponierte $A^T = (a_{ji})_{1\le j\le n,\, 1\le i\le m}$ eine $n \times m$-Matrix.

### 2.5.2 Rang einer Matrix

Der Spaltenrang einer Matrix ist die Anzahl der linear unabhängigen Spalten und der Zeilenrang die der linear unabhängigen Zeilen. Der Spaltenrang und der Zeilenrang einer Matrix sind immer gleich. Dieser Wert heißt Rang der Matrix. Für eine $m \times n$-Matrix $A$ gilt $\mathrm{Rang}(A) = \mathrm{Rang}(A^T) = \mathrm{Rang}(A^T A) = \mathrm{Rang}(AA^T) \leq \min(m, n)$ (Magnus, 1988, S. 2 f.).

### 2.5.3 Eigenwerte und Eigenvektoren

Es sei $A$ eine quadratische $n \times n$-Matrix mit Komponenten aus einem Körper $K$. Ein Skalar $\lambda \in K$ heißt Eigenwert von $A$, wenn es einen von 0 verschiedenen Vektor $x \in K^n$ gibt, der

$$Ax = \lambda x$$

erfüllt. Der Vektor $x$ heißt dann Eigenvektor von $A$ zum Eigenwert $\lambda$. Der Skalar $0 \in K$ kann ein Eigenwert sein. Der Nullvektor $0 \in K^n$ ist jedoch niemals ein Eigenvektor. Mit $x$ ist auch jedes skalare Vielfache $\mu x$ von $x$ mit einem Skalar $\mu \neq 0$ ein Eigenvektor zum selben Eigenwert $\lambda$, denn $A(\mu x) = \mu Ax = \mu\lambda x = \lambda(\mu x)$. Die Menge

$$\mathrm{Eig}(A, \lambda) := \{x \in K^n : Ax = \lambda x\} = \mathrm{Kern}(A - \lambda E)$$

heißt Eigenraum von $A$ bezüglich $\lambda$. Der um den Nullvektor verminderte Eigenraum $\mathrm{Eig}(A, \lambda)$ ist die Menge aller Eigenvektoren von $A$ zum Eigenwert $\lambda$. Die Eigenräume verschiedener Eigenwerte $\lambda_1 \neq \lambda_2$ haben nur den Nullvektor gemeinsam, $\mathrm{Eig}(A, \lambda_1) \cap \mathrm{Eig}(A, \lambda_2) = \{0\}$. Die Dimension des Eigenraumes $\mathrm{Eig}(A, \lambda)$ heißt geometrische Vielfachheit des Eigenwertes $\lambda$ bezüglich $A$. Ein Skalar $\lambda$ ist ein Eigenwert einer quadratischen Matrix $A$ genau dann, wenn er die charakteristische Gleichung $\det(A - \lambda E) = 0$ erfüllt (Lay, 1997, S. 307). Die Eigenwerte einer quadratischen $n \times n$-Matrix $A$ sind somit die $n$ Nullstellen (Wurzeln) ihres charakteristischen Polynoms $\det(A - \lambda E)$. Eine $n \times n$-Matrix hat genau $n$ Eigenwerte, wenn deren algebraische Vielfachheiten mitgezählt werden (Lütkepohl, 1996, S. 64). Folglich hat sie höchstens $n$ verschiedene Eigenwerte. Der Rang einer quadratischen $n \times n$-Matrix $A$ ist größer oder gleich der Anzahl $r$ der von Null verschiedenen Eigenwerte, $\mathrm{Rang}(A) \geq r$ (Lütkepohl, 1996, S. 66). Die Summe der Eigenwerte einer Matrix ist ihre Spur. Das Produkt der Eigenwerte einer Matrix ergibt ihre Determinante.

Eigenvektoren zu verschiedenen Eigenwerten sind linear unabhängig (JOHNSON et al., 1993, S. 256).

### 2.5.4 Ähnliche Matrizen

Zwei quadratische Matrizen $A$ und $B$ heißen ähnlich, wenn es eine invertierbare Matrix $P$ gibt, so daß

$$A = PBP^{-1}.$$

Ähnliche Matrizen haben dieselben Eigenwerte (JOHNSON et al., 1993, S. 269).

### 2.5.5 Diagonalisierbare Matrix

Eine quadratische Matrix heißt diagonalisierbar, wenn sie ähnlich zu einer Diagonalmatrix ist. Die Eigenwerte einer Diagonalmatrix sind die Elemente ihrer Hauptdiagonalen (LAY, 1997, S. 300). Eine $n \times n$-Matrix ist diagonalisierbar genau dann, wenn sie $n$ linear unabhängige Eigenvektoren besitzt (JOHNSON et al., 1993, S. 270).

### 2.5.6 Eigenzerlegung

Es sei $A$ eine quadratische $n \times n$-Matrix. Ihre Eigenwerte seien $\lambda_1, \lambda_2, \ldots, \lambda_n$ und die zugehörigen Eigenvektoren $x_1, x_2, \ldots, x_n$. Die Gleichungen $Ax_i = \lambda_i x_i$ können als Matrixgleichung $AP = P\Lambda$ geschrieben werden, wobei die Spalten von $P$ die Eigenvektoren sind und $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ die Diagonalmatrix der zugehörigen Eigenwerte ist. Dann läßt sich $A$ darstellen als Eigenzerlegung (*eigen decomposition, matrix diagonalization*)

$$A = P\Lambda P^{-1}$$

(LAY, 1997, S. 314).

## 2.6 Bilinearformen und quadratische Formen

### 2.6.1 Bilinearform

Es seien $V$ und $W$ Vektorräume über demselben Körper $K$. Eine skalarwertige Abbildung

$$f\colon V \times W \to K,$$

heißt bilinear oder eine Bilinearform, wenn gilt:

(BF1) Für jedes $w \in W$ ist

$$f(\cdot, w) : V \to K, \quad v \mapsto f(v,w)$$

linear. Für ein beliebiges, aber festes zweite Argument ist $f$ im ersten linear.

(BF2) Für jedes $v \in V$ ist

$$f(v, \cdot) : W \to K, \quad w \mapsto f(v,w)$$

linear. Für ein beliebiges, aber festes erstes Argument ist $f$ im zweiten linear.

### 2.6.2 Matrix einer Bilinearform

Es seien $V$ und $W$ endlich-dimensionale Vektorräume über demselben Körper $K$, $B := \{v_1, v_2, \ldots, v_n\}$ eine Basis in $V$, $C := \{w_1, w_2, \ldots, w_m\}$ eine Basis in $W$ und $f : V \times W \to K$ eine Bilinearform. Dann stellt die $m \times n$-Matrix $A := (a_{Eis})_{1 \le i \le m,\, 1 \le j \le n}$ mit $a_{ij} := f(v_i, w_j)$ die Bilinearform $f$ bezüglich der Basen $B$ und $C$ dar, denn für alle $v \in V$ und $w \in W$ ist

$$f(v, w) = [v]_B^T A [w]_C$$

(Nering, 1970, S. 156 f.). Im allgemeinen wird einer Bilinearform eine Eigenschaft zugeschrieben, wenn ihre darstellende Matrix, diese Eigenschaft hat.

### 2.6.3 Symmetrische Bilinearform

Eine Bilinearform $f\colon V \times V \to K$ heißt symmetrisch, wenn für alle $v, w \in V$

$$f(v, w) = f(w, v)$$

gilt (Fischer, 1989, S. 183). Damit diese Definition sinnvoll ist, muß die Bilinearform auf Paaren von Vektoren desselben Vektorraums definiert sein. Eine Bilinearform ist symmetrisch genau dann, wenn die sie darstellende Matrix symmetrisch ist (Nering, 1970, S. 158).

**Beispiel**

Es seien $x, y \in \mathbb{R}^n$ Spaltenvektoren und $A \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix. Dann ist durch $x^T A y$ eine symmetrische Bilinearform auf $\mathbb{R}^n$ gegeben (Fischer, 1989, S. 186).

### 2.6.4 Schiefsymmetrische Bilinearform

Eine Bilinearform $f : V \times V \to K$ heißt schiefsymmetrisch (*skew-symmetric*), wenn für alle $v \in V$

$$f(v, v) = 0$$

gilt (Nering, 1970, S. 158). Es sei $K$ ein Körper, in dem für das Einselement und das Nullelement die Beziehung $1 + 1 \neq 0$ gilt. Dann ist eine Bilinearform schiefsymmetrisch genau dann, wenn eine sie darstellende Matrix $A$ die Eigenschaft $A^T = -A$ hat (Nering, 1970, S. 158). Ist $1 + 1 \neq 0$, so kann jede Bilinearform $f$ eindeutig als Summe einer symmetrischen Bilinearform $f_s$ und einer schiefsymmetrischen Bilinearform $f_{ss}$ dargestellt werden. Es ist dann $f_s(u, v) = \frac{1}{2}\big(f(u, v) + f(v, u)\big)$ und $f_{ss}(u, v) = \frac{1}{2}\big(f(u, v) - f(v, u)\big)$. Die Voraussetzung $1 + 1 \neq 0$ stellt sicher, daß der Koeffizient $\frac{1}{2}$ wohldefiniert ist (Nering, 1970, S. 159).

### 2.6.5 Positiv definite Bilinearform und Matrix

Eine reellwertige symmetrische Bilinearform $f\colon V \times V \to \mathbb{R}$ heißt positiv definit, wenn für alle $v \in V$ mit $v \neq 0$

$$f(v, v) > 0$$

gilt (Fischer, 1989, S. 184).

Eine symmetrische $n \times n$-Matrix $A$ über $\mathbb{R}$ heißt positiv definit, wenn die aus ihr gebildete Bilinearform $x^T Ay$, $x, y \in \mathbb{R}^n$ positiv definit ist. Eine positiv definite Matrix ist stets invertierbar. Eine symmetrische Matrix $A$ ist positiv definit genau dann, wenn es eine invertierbare Matrix $P$ mit $A = P^T P$ gibt (Koecher, 1985, S. 153 f.).

### 2.6.6 Skalarprodukt

Ist $V$ ein Vektorraum über dem Körper $\mathbb{R}$ der reellen Zahlen, so nennt man eine positiv definite symmetrische Bilinearform

$$\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{R},\ (x, y) \mapsto \langle x, y \rangle$$

ein Skalarprodukt in $V$. Ein reeller Vektorraum zusammen mit einem Skalarprodukt heißt euklidischer Vektorraum (Fischer, 1989, S. 184 f.). Für eine quadratische reelle Matrix $A$ und Vektoren $x, y$ gilt $\langle x, y \rangle_A := \langle Ax, y \rangle = \langle x, A^T y \rangle$.

#### 2.6.6.1 Kanonisches Skalarprodukt

Für Spaltenvektoren $x$ und $y$ im reellen Vektorraum $\mathbb{R}^n$ ist durch

$$\begin{aligned} \langle \cdot\,, \cdot \rangle \colon \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R}, \\ (x, y) \ &\mapsto \ \langle x, y \rangle := x^T y = \sum_{i=1}^{n} x_i\, y_i \end{aligned}$$

das sogenannte kanonische Skalarprodukt (euklidisches inneres Produkt, *dot product, inner product*) in $\mathbb{R}^n$ definiert (Fischer, 1989, S. 185).

Die Projektion eines Vektors $y$ auf einen anderen Vektor $x$ ist das Vielfache von $y$, dessen Länge durch das Lot von $y$ auf $x$ bestimmt wird. Geometrisch bedeutet das kanonische Skalarprodukt $\langle x, y \rangle$ eine Multiplikation der Länge des Vektors $x$ im ersten Argument mit der Länge der Projektion des Vektors $y$ im zweiten Argument auf den ersten Vektor. Das kanonische Skalarprodukt bestimmt die Länge der Projektion von $y$ auf $x$,

$$\langle x, y \rangle = \|x\| \cdot \|\text{Projektion von } y \text{ auf } x\|$$

(Abbildung 2.2). Das kanonische Skalarprodukt mit normierten Vektoren berechnet den Cosinus des Winkels (Abschnitt 2.9) zwischen den Vektoren.

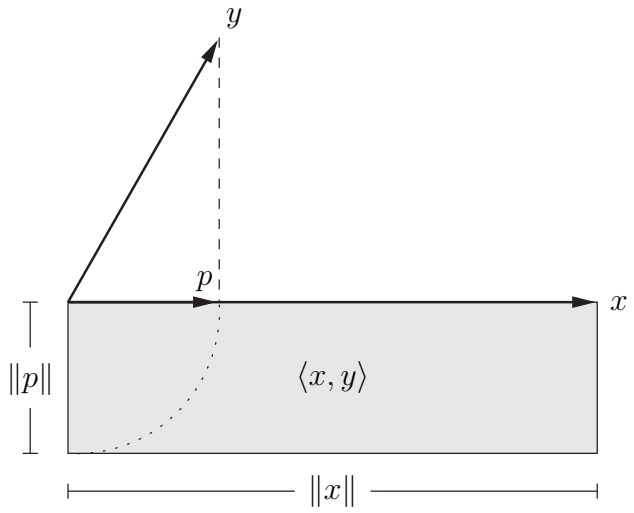



**Abbildung 2.2:** Kanonisches Skalarprodukt als Multiplikation der Länge des Vektors $x$ mit der Länge der Projektion $p$ von $y$ auf $x$

#### 2.6.6.2 Induziertes Skalarprodukt

Die Abbildung

$$\begin{aligned} &\langle \cdot\,,\cdot \rangle_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \\ &\qquad (x,y) \;\mapsto\; \langle x,y\rangle_A := \langle A^T x, y\rangle = \langle x, Ay\rangle = x^T A y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i\, a_{ij}\, y_j \end{aligned}$$

ist ein durch die gewichtende Matrix $A \in \mathbb{R}^{n\times n}$ induziertes Skalarprodukt im $\mathbb{R}^n$. Bei Gewichtung mit der Einheitsmatrix ergibt sich das kanonische Skalarprodukt.

### 2.6.7 Quadratische Form

Ist $f : V \times V \to K$ eine symmetrische Bilinearform, so heißt die Abbildung

$$q : V \to K, \quad v \mapsto q(v) := f(v,v)$$

die durch $f$ bestimmte quadratische Form. Wird $f$ als Summe einer symmetrischen Bilinearform $f_s$ und einer schiefsymmetrischen Bilinearform $f_{ss}$ dargestellt, ist $q$ allein durch die symmetrische Bilinearform vollständig bestimmt, $q(v) = f_s(v,v) + f_{ss}(v,v) = f_s(v,v)$. Jede symmetrische Bilinearform $f_s$ bestimmt eindeutig eine quadratische Form $q(v) = f_s(v,v)$. Gilt für den Skalarenkörper $1 + 1 \neq 0$, bestimmt jede quadratische Form eindeutig eine Bilinearform $f_s(u,v) = \frac{1}{2}[q(u+v) - q(u) - q(v)]$. Somit gibt es eine eineindeutige Zuordnung

zwischen symmetrischen Bilinearformen und quadratischen Formen. Die eindeutig bestimmte symmetrische Bilinearform $f_s$ einer quadratischen Form $q$ heißt Polarform von $q$ (Nering, 1970, S. 160 f.). Die darstellende Matrix $A$ einer quadratischen Form $q$ ist die darstellende Matrix der Polarform $f_s$ bezüglich einer Basis $B$. Der symmetrische Anteil von $A$ ist $\frac{1}{2}(A + A^T)$. Die quadratische Form läßt sich nun schreiben als $q(v) = [v]_B^T A [v]_B = [v]_B^T \frac{A+A^T}{2} [v]_B$. Eine aus einer symmetrischen Matrix $A$ erzeugte quadratische Form ist positiv definit genau dann, wenn alle Eigenwerte von $A$ positiv sind (Lay, 1997, S. 456).

## 2.7 Vektornormen

Im folgenden sei $\mathbb{K}$ stets der Körper der reellen oder komplexen Zahlen. Es sei $V$ ein $\mathbb{K}$-Vektorraum. Eine Abbildung

$$\|\cdot\| : V \to \mathbb{R}, \quad x \mapsto \|x\|$$

heißt (Vektor-)Norm auf $V$, wenn für alle $\lambda \in \mathbb{K}$ und alle $x, y \in V$

$$\|x\| = 0 \quad \Leftrightarrow \quad x = 0, \tag{2.3}$$

$$\|\lambda x\| = |\lambda| \, \|x\|, \tag{2.4}$$

$$\|x + y\| \leq \|x\| + \|y\| \tag{2.5}$$

gelten. Sie ist definit (2.3), absolut homogen (2.4) und erfüllt die Dreiecksungleichung (2.5). Nur der Nullvektor hat die Länge 0 (2.3). Aus (2.4) und (2.5) folgt $\|x\| \geq 0$ für jedes $x \in V$. Die nichtnegative reelle Zahl $\|x\|$ heißt Länge (Norm, Betrag) des Vektors $x$. Das Paar $(V, \|\cdot\|)$ aus dem $\mathbb{K}$-Vektorraum $V$ und einer Norm $\|\cdot\|$ auf $V$ heißt normierter Vektorraum. Vgl. Fischer (1989, S. 189 ff.).

### 2.7.1 Skalarproduktinduzierte Norm

Ist $\langle \cdot, \cdot \rangle_A$ ein Skalarprodukt auf einem euklidischen Vektorraum $V$, so ist durch $\|x\|_A := \sqrt{\langle x, x \rangle_A}$ für $x \in V$ eine Norm auf $V$ definiert. Man nennt sie euklidische Norm. Somit wird jeder euklidische Vektorraum $(V, \langle \cdot, \cdot \rangle_A)$ vermittels der positiv definiten quadratischen Form $\langle x, x \rangle_A$ für $x \in V$ zu einem normierten Raum $(V, \|\cdot\|_A)$ (Fischer, 1989, S. 191). Nicht jede Norm ist auf ein Skalarprodukt zurückzuführen (Koecher, 1985, S. 155).

### 2.7.2 $p$-Norm

Für ein $p \geq 1$ ist die $p$-Norm (MINKOWSKI-Norm) in $\mathbb{R}^n$ definiert durch

$$\|x\|_p := \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

(HORN und JOHNSON, 1985, S. 265). Von diesen sind die 1-, 2- und die $\infty$-Norm am wichtigsten. Die Definitheit (2.3) und die Homogenität (2.4) ergeben sich sofort aus den Eigenschaften des Betrages $|\cdot|$ und der allgemeinen Potenzen. Für diese Normen ist die Dreiecksungleichung (2.5) die MINKOWSKI-Ungleichung (BRONSTEIN et al., 1995, S. 21).

### 2.7.3 1-Norm

Für $p = 1$ ergibt sich die 1-Norm (Summennorm, Manhattan-Norm, *city block norm*)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

in $\mathbb{R}^n$. Sie mißt die Länge nur entlang der Koordinatenrichtungen. Sie kann nicht aus einem Skalarprodukt abgeleitet werden. Vgl. HORN und JOHNSON (1985, S. 265).

### 2.7.4 2-Norm

Die 2-Norm (euklidische Norm, Spektralnorm, Quadratsummennorm)

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{x^T x}$$

in $\mathbb{R}^n$ ist die wohl bekannteste Norm, da $\|x - y\|_2$ den euklidischen Abstand der beiden Punkte $x, y \in \mathbb{R}^n$ mißt. Sie wird vom kanonischen Skalarprodukt induziert. Vgl. HORN und JOHNSON (1985, S. 264 f.).

### 2.7.5 $\infty$-Norm

Für $p \to \infty$ erhält man die $\infty$-Norm (Supremumnorm, Maximumnorm, Tschebyscheff-Norm, *Chebyshev norm*)

$$\|x\|_\infty := \lim_{p\to\infty} \|x\|_p = \max_{1\le i\le n} |x_i|$$

in $\mathbb{R}^n$. Vgl. Horn und Johnson (1985, S. 265) und Lütkepohl (1996, S. 109).

### 2.7.6 Einheitsvektor

Ein Vektor $x$ wird Einheitsvektor (*unit vector*) genannt, wenn seine Norm 1 ist, $\|x\| = 1$. Jeder von Null verschiedene Vektor $x \neq 0$ wird durch Division mit dem Wert seiner Norm zu einem Einheitsvektor $x/\|x\|$, denn

$$\left\| \frac{x}{\|x\|} \right\| = \frac{1}{\|x\|}\|x\| = 1.$$

Diese Operation, bei der der Vektor auf die Hülle der Einheitskugel projiziert wird, nennt man Normierung (*normalization*) eines Vektors. Ein Vektor und seine normierte Form zeigen in dieselbe Richtung (Lay, 1997, S. 371).

## 2.8 Metriken

Es sei $X$ eine nichtleere Menge. Eine Abbildung

$$d : X \times X \to \mathbb{R},\ (x, y) \mapsto d(x, y),$$

die je zwei Punkten $x$ und $y$ aus $X$ eine reelle Zahl $d(x, y)$ zuordnet, heißt Metrik auf $X$, wenn sie folgende Axiome erfüllt:

(M1) $d(x, y) \geq 0$ für alle $x, y \in X$.

(M2) $d(x, y) = 0$ genau dann, wenn $x = y$.

(M3) $d(x, y) = d(y, x)$ für alle $x, y \in X$.

(M4) $d(x, z) \leq d(x, y) + d(y, z)$ für alle $x, y, z \in X$.

Metriken sind nichtnegativ (M1), symmetrisch (M3) und erfüllen die Identität (M2) und die Dreiecksungleichung (M4). Die Dreiecksungleichung besagt

anschaulich, daß die Länge einer Dreieckseite nicht größer als die Summe der beiden anderen Seiten ist. Der Wert $d(x,y)$ heißt Abstand oder Distanz der Elemente $x$ und $y$. Die in der Definition geforderten Axiome (M1) bis (M4) lassen sich auf die beiden folgenden Axiome reduzieren:

(M1′) $d(x,y) = 0$ genau dann, wenn $x = y$.

(M2′) $d(x,y) \leq d(x,z) + d(y,z)$ für alle $x, y, z \in X$.

Eine nichtleere Menge $X$ zusammen mit einer Metrik heißt metrischer Raum $(X,d)$. Zwei verschiedene Metriken auf derselben Menge erzeugen zwei verschiedene metrische Räume. Es ist also möglich, eine Menge auf verschiedene Arten zu metrisieren.

Die Abbildung $d$ heißt Ultrametrik, wenn anstelle der Dreiecksungleichung (M4) die Ultrametrikungleichung

$$\forall_{x,y,z \in X} \; d(x,z) \leq \max\big(d(x,y), d(y,z)\big)$$

gilt. Sie besagt, daß die Länge einer Dreieckseite nicht größer ist als die längste der anderen beiden. Ist die ultrametrische Ungleichung erfüllt, gilt auch die Dreiecksungleichung. Äquivalent zur Ultrametrikungleichung ist die Aussage, daß für beliebige $x, y, z \in X$ die beiden größten der Distanzen $d(x,y), d(x,z)$ und $d(y,z)$ gleich sind. Zum Beweis sei $a := d(x,y)$, $b := d(x,z)$ und $c := d(y,z)$. Es gelte die Ultrametrikungleichung und somit

$$a \leq \max(b,c), \tag{2.6}$$

$$b \leq \max(a,c),$$

$$c \leq \max(a,b). \tag{2.7}$$

Ohne Beschränkung der Allgemeinheit sei $b \leq c$. Dann ist wegen (2.6) auch $a \leq c$ und aus beiden Ungleichungen zusammen folgt $\max(a,b) \leq c$. Nach (2.7) muß somit $c = \max(a,b)$ sein. Nun ist entweder $a \leq b = c$ oder $b \leq a = c$. In beiden Fällen sind die beiden größten Distanzen gleich. Die Rückrichtung ergibt sich sofort durch Einsetzen in die Ultrametrikungleichung. Vgl. Eckes und Rossbach (1980, S. 38 f.), Fischer (1989, S. 189 f.) und Gordon (1999, S. 71).

### 2.8.1 Norminduzierte Metrik

Ist $\|\cdot\|$ irgendeine Norm auf einem $\mathbb{R}$-Vektorraum $V$, so ist durch $d(x,y) := \|x-y\|$ für $x,y \in V$ stets eine Metrik auf $V$ erklärt (Fischer, 1989, S. 190). Hervorzuheben sind hierbei Metriken, denen Normen zugrunde liegen, die wiederum aus durch quadratische reelle Matrizen $A$ gewichteten Skalarprodukten $\langle\cdot,\cdot\rangle_A$ erzeugt werden,

$$\begin{aligned} d_A &\colon V \times V \to \mathbb{R} \\ &(x,y) \mapsto d_A(x,y) := \|x-y\|_A := \sqrt{\langle x-y, x-y\rangle_A} = \sqrt{(x-y)^T A (x-y)}. \end{aligned}$$

### 2.8.2 Triviale Metrik

Aus jeder nichtleeren Menge $X$ wird vermittels der trivialen Metrik

$$d(x,y) := \begin{cases} 0, & \text{falls } x = y, \\ 1, & \text{falls } x \neq y \end{cases}$$

ein metrischer Raum $(X,d)$.

### 2.8.3 Euklidische Metrik

Mit der Einheitsmatrix $E$ wird die euklidische Metrik (euklidischer Abstand)

$$d_2(x,y) := \|x-y\|_2 = \sqrt{(x-y)^T (x-y)} = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

durch die euklidische Norm und somit letztlich durch das kanonische Skalarprodukt in $\mathbb{R}^n$ generiert. Sie entspricht der gewöhnlichen geometrischen Anschauung des Abstandes. Sie ist translationsinvariant, das heißt invariant bezüglich einer Verschiebung des Ursprungs, und invariant unter orthogonalen linearen Transformationen (Drehung, Spiegelung) der Vektoren. Sie ist jedoch nicht invariant unter beliebigen nichtsingulären Transformationen. Deswegen müssen die Einheiten der Komponenten der Vektoren vergleichbar und unkorreliert sein (Steinhausen und Langer, 1977, S. 58).

### 2.8.4 Mahalanobis-Metrik

Setzt man für $A$ die Inverse der Kovarianzmatrix

$$S := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \in \mathbb{R}^{n\times n}$$

ein, wobei $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ das arithmetische Mittel über alle $n$ Vektoren ist, ergibt sich die MAHALANOBIS-Metrik

$$\begin{aligned} d_{S^{-1}}(x,y) &= \|x-y\|_{S^{-1}} = \sqrt{(x-y)^T S^{-1}(x-y)} \\ &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}(x_i - y_i)(x_j - y_j)}, \end{aligned}$$

wobei $s_{ij}$ die Komponenten von $S^{-1}$ sind. Es wird vorausgesetzt, daß die Kovarianzmatrix $S$ symmetrisch und positiv definit ist, um invertierbar sein zu können. Die MAHALANOBIS-Distanz oder auch generalisierte Distanz ist invariant unter nichtsingulären Transformationen, skaleninvariant und eliminiert etwaige Korrelationen zwischen den Vektoren (STEINHAUSEN und LANGER, 1977, S. 60).

### 2.8.5 Minkowski-Metrik

Eine weitere unendliche Familie von Distanzfunktionen im $\mathbb{R}^n$ wird durch die MINKOWSKI-Normen definiert. Mit der 1-Norm wird die Manhattan-Metrik (*city block metric, taxicab metric*)

$$d_1(x,y) := \|x-y\|_1 = \sum_{i=1}^{n}|x_i - y_i|$$

gebildet. Mit der 2-Norm ergibt sich die euklidische Metrik. Für die Maximumnorm erhält man die Maximummetrik

$$d_\infty(x,y) := \|x-y\|_\infty = \max_{1\le i\le n}|x_i - y_i|.$$

Zwischen den MINKOWSKI-Metriken gilt die Beziehung

$$\underset{1\le p\le q\le\infty}{\forall} \; d_q(x,y) \le d_p(x,y).$$

## 2.9 Winkel

Es sei $V$ ein euklidischer Vektorraum mit einem Skalarprodukt $\langle\cdot,\cdot\rangle$ und der daraus induzierten Norm $\|\cdot\|$. Der Winkel $\measuredangle(x,y)$ im Bogenmaß zwischen zwei von Null verschiedenen Vektoren $x,y \in V$ ist die eindeutig bestimmte reelle Zahl $\theta$ definiert durch den Richtungscosinus (*direction cosine*)

$$\cos\theta = \frac{\langle x,y\rangle}{\|x\|\,\|y\|} = \left\langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \right\rangle \qquad \text{und} \qquad 0 \le \theta \le \pi.$$

Nach der CAUCHY-SCHWARZschen Ungleichung (FISCHER, 1989, S. 11) liefert diese Festlegung stets einen zwischen $-1$ und $+1$ gelegenen Wert für $\cos\theta$. Ein Strecken oder Stauchen der Vektoren verändert den Winkel nicht, weil die Vektoren zu Einheitsvektoren normiert werden. Für $\lambda,\mu \in \mathbb{R}$ mit $\lambda\mu > 0$ gilt also $\measuredangle(\lambda x, \mu y) = \measuredangle(x,y)$. Die Vektoren heißen orthogonal genau dann, wenn $\theta = \pi/2$, also $\langle x,y\rangle = \cos\theta = 0$. Der Nullvektor steht definitionsgemäß auf alle Vektoren senkrecht. Die Vektoren sind linear abhängig genau dann, wenn $\theta = 0$ oder $\theta = \pi$. Der eine Vektor $y$ ist dann ein Vielfaches des anderen Vektors $x$, also $y = \rho x$ mit $\rho \in \mathbb{R}$.

Das kanonische Skalarprodukt läßt sich auch schreiben als das Produkt aus dem Cosinus des Winkels zwischen den Vektoren und deren Längen

$$\langle x,y\rangle = \cos\theta\ \|x\|\ \|y\|.$$

Sind die Vektoren bereits normiert, stimmt der Richtungscosinus mit dem kanonischen Skalarprodukt überein. Das kanonische Skalarprodukt mit normierten Vektoren berechnet den Winkel zwischen den Vektoren. Vgl. FISCHER (1989, S. 12) und KOECHER (1985, S. 156).

## 2.10 Singulärwerte

### 2.10.1 Orthogonalität

Sei $V$ ein euklidischer Vektorraum. Zwei Vektoren $x,y \in V$ heißen orthogonal oder senkrecht zueinander, $x \perp y$, wenn ihr Skalarprodukt $\langle x,y\rangle = 0$ ist. Ein Menge $M$ von Vektoren aus $V$ heißt orthogonal, wenn ihre Elemente paarweise orthogonal sind,

$$\underset{x,y \in M,\ x \neq y}{\forall} \langle x,y\rangle = 0.$$

Sie heißt orthonormal wenn sie orthogonal ist und jedes Element die Norm 1 hat,

$$\underset{x \in M}{\forall} \|x\| = 1.$$

Eine Teilmenge eines euklidischen Vektorraums wird Orthonormalbasis genannt, wenn sie orthonormal und eine Basis ist.

**Beispiel**

Die gewöhnliche Basis $\{(1,0,0), (0,1,0), (0,0,1)\}$ des euklidischen Raumes $\mathbb{R}^3$ mit dem kanonischen Skalarprodukt ist eine Orthonormalbasis.

### 2.10.2 Orthogonale Matrix

Eine invertierbare reelle Matrix $A$ heißt orthogonal, wenn ihre Inverse gleich ihrer Transponierten ist, $A^{-1} = A^T$ (Fischer, 1989, S. 200). Invertierbare Matrizen sind auch immer quadratisch. Für orthogonale Matrizen gilt $AA^T = A^TA = E$. Die Eigenwerte einer orthogonalen Matrix haben stets den Betrag 1 (Eisenreich, 1980, S. 248). Eine symmetrische $n \times n$-Matrix $A$ hat $n$ orthonormale Eigenvektoren. Zu verschiedenen Eigenwerten von $A$ gehörige Eigenvektoren sind stets orthogonal (Eisenreich, 1980, S. 255).

Bildet man aus einer reellen orthogonalen $n \times n$-Matrix $A$ eine lineare Abbildung $x \mapsto Ax$ von $\mathbb{R}^n$ nach $\mathbb{R}^n$, so erhält diese die Längen der Vektoren und die Winkel zwischen den Vektoren (Johnson et al., 1993, S. 275).

### 2.10.3 Orthogonal diagonalisierbare Matrix

Eine reelle quadratische Matrix $A$ heißt orthogonal diagonalisierbar, wenn es eine orthogonale Matrix $P$ und eine Diagonalmatrix $D$ gibt, so daß

$$A = PDP^{-1}$$

(Lay, 1997, S. 444). Eine reelle quadratische Matrix ist orthogonal diagonalisierbar genau dann, wenn sie symmetrisch ist (Lay, 1997, S. 445). Die Eigenwerte einer symmetrischen Matrix sind reell (Eisenreich, 1980, S. 252).

### 2.10.4 Singulärwerte und Singulärvektoren

Der Betrag eines Eigenwertes einer symmetrischen Matrix $A$ gibt an, wie stark $A$ einen Vektor streckt oder staucht. Ist $Ax = \lambda x$ mit $\|x\| = 1$, gilt $\|Ax\| = \|\lambda x\| = |\lambda| \, \|x\| = |\lambda|$. Ist $\lambda_1$ der Eigenwert mit dem größten Betrag, so gibt der zugehörige Einheitseigenvektor $v_1$ die Richtung an, in der die Streckung von $A$ am größten ist. Die Länge von $Ax$ ist maximal für $x = v_1$, denn $\|Av_1\| = |\lambda_1|$. Das gesamte geometrische Verhalten einer linearen Abbildung $x \mapsto Ax$ wird durch die quadratische Form $x^T(A^TA)x$ bestimmt (Lay, 1997, S. 467 f.). Für eine reelle Matrix $A$ sind die Matrizen $A^TA$ und $AA^T$ reell, symmetrisch und positiv definit. Somit sind alle Eigenwerte von $A^TA$ nichtnegativ (Lay, 1997, S. 468). Deshalb kann $A^TA$ orthogonal diagonalisiert werden.

Die Singulärwerte $\sigma_1, \sigma_2, \ldots, \sigma_p$ einer reellen $m \times n$-Matrix $A$ sind die nichtnegativen Quadratwurzeln der $p = \min\{m, n\}$ Eigenwerte $\lambda_1, \lambda_2, \ldots, \lambda_p$ von $A^TA$ oder $AA^T$. Die Anzahl der positiven Singulärwerte entspricht dem Rang $r$ der Matrix $A$. Die restlichen Eigenwerte sind alle 0. Es ist üblich, die Singulärwerte absteigend zu ordnen. Somit gilt

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0.$$

Manchmal werden auch nur die positiven Werte als Singulärwerte bezeichnet. Die geordneten Singulärwerte sind eindeutig durch $A$ bestimmt (Horn und Johnson, 1991, S. 146). Der Eigenvektor $u_i$ von $A^TA$ heißt rechter Singulärvektor und der Eigenvektor $v_i$ von $AA^T$ heißt linker Singulärvektor von $\sigma_i$. Die Matrix $A^TA$ hat dieselben von 0 verschiedenen Eigenwerte $\lambda_i$ wie $AA^T$,

$$\underset{1 \leq i \leq r}{\forall} A^TAu_i = \lambda_i u_i, \; AA^Tv_i = \lambda_i v_i.$$

Die Matrix $A$ und ihre Transponierte $A^T$ haben dieselben Singulärwerte. Zwischen den linken und rechten Singulärwerten besteht der Zusammenhang

$$\underset{1 \leq i \leq r}{\forall} Au_i = \sigma_i v_i, \; A^Tv_i = \sigma_i u_i$$

(Bronstein et al., 1995, S. 251).

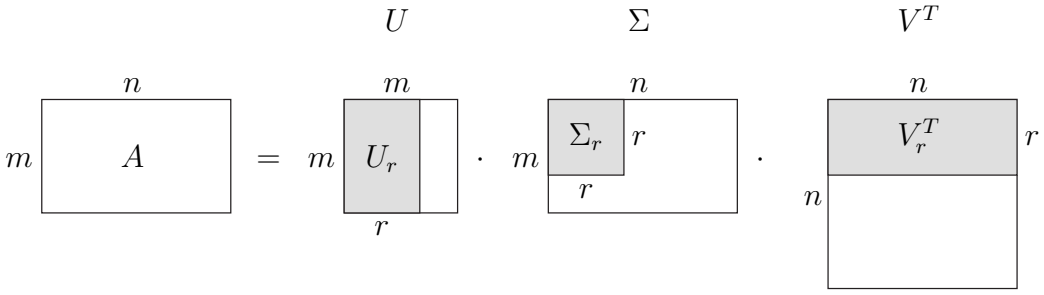



**Abbildung 2.3:** Reduzierte Singulärwertzerlegung

### 2.10.5 Singulärwertzerlegung

Es sei $A$ eine reelle $m \times n$-Matrix und $p = \min\{m, n\}$. Dann gibt es eine Diagonalmatrix $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}$ mit $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$ und zwei orthogonale Matrizen $U = (u_1, u_2, \ldots, u_m) \in \mathbb{R}^{m \times m}$ und $V = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^{n \times n}$, so daß

$$A = U \Sigma V^T.$$

Die Diagonalelemente $\sigma_i$ sind die Singulärwerte von $A$. Die $i$-te Spalten $u_i$ von $U$ ist der rechte und die $i$-te Spalten $v_i$ von $V$ der linke Singulärvektor zum Singulärwert $\sigma_i$. Eine solche Faktorisierung nennt man eine Singulärwertzerlegung (*singular value decomposition*) von $A$. Die Spalten von $U$ sind orthonormal und die von $V$ ebenfalls. Der Faktor $\Sigma$ ist stets eindeutig bestimmt. Die linken und rechten orthogonalen Faktoren $U$ und $V$ sind bis auf die Vorzeichen eindeutig bestimmt. Vgl. Horn und Johnson (1985, S. 414 f.) und Horn und Johnson (1991, S. 144 ff.).

Mit dem Rang $r$ von $A$ ist $\sigma_i > 0$ für alle $1 \leq i \leq r$ und $\sigma_i = 0$ für alle $r + 1 \leq i \leq p$. Es seien nun

$$\begin{aligned}
U_r &:= U(:, 1{:}r) = (u_1, u_2, \ldots, u_r) \in \mathbb{R}^{m \times r}, \\
\Sigma_r &:= \Sigma(1{:}r, 1{:}r) = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \in \mathbb{R}^{r \times r} \quad \text{und} \\
V_r &:= V(:, 1{:}r) = (v_1, v_2, \ldots, v_r) \in \mathbb{R}^{n \times r}
\end{aligned}$$

die entsprechend der positiven Singulärwerte reduzierten Faktoren. Dann ergibt sich (Abbildung 2.3) die reduzierte Singulärwertzerlegung (*reduced singular value decomposition, singular value decomposition expansion*)

$$A = U_r \Sigma_r V_r^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

(Golub und Van Loan, 1989, S. 71 f.).

Dimensionsreduktion wird dadurch erreicht, daß nur die ersten $k < r$ Singulärwerte $\sigma_1, \sigma_2, \ldots, \sigma_k$ beibehalten und alle anderen auf 0 gesetzt werden. Dann ist das Produkt

$$A_k := U \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k, 0, \ldots, 0)\, V^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

(*reduced dimension representation*) die beste Approximation von $A$ in einem $k$-dimensionalen Raum,

$$\min_{\operatorname{Rang}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

Es gibt keine Matrix vom Rang $k$ mit einer kleineren Distanz der kleinsten Quadrate zu $A$ als $A_k$ (Golub und Van Loan, 1989, S. 73).

## 2.11 Matrixnormen

Es sei $V$ ein $n$-dimensionaler normierter $\mathbb{K}$-Vektorraum, $\|\cdot\|_V$ eine Norm auf $V$, $W$ ein $m$-dimensionaler normierter $\mathbb{K}$-Vektorraum, $\|\cdot\|_W$ eine Norm auf $W$ und $A \in \mathbb{K}^{m\times n}$ die darstellende Matrix einer linearen Abbildung von $V$ nach $W$. Die Norm auf dem Vektorraum $\mathbb{K}^{m\times n}$ aller $m \times n$-Matrizen über dem Körper $\mathbb{K}$ ist definiert durch

$$|\!|\!| A |\!|\!|_{V,W} := \sup_{v\neq 0} \frac{\|Av\|_W}{\|v\|_V}$$

für $A \in \mathbb{K}^{m\times n}$. Die Norm $|\!|\!| A |\!|\!|_{V,W}$ von $A$ ist endlich und somit die kleinste reelle Zahl mit $\|Av\|_W \leq |\!|\!| A |\!|\!|_{V,W} \|v\|_V$ für alle $v \in V$. Wegen der Homogenität von Normen gilt

$$|\!|\!| A |\!|\!|_{V,W} = \sup_{\|v\|_V=1} \|Av\|_W.$$

Die Norm einer Matrix ist die Norm der zugehörigen linearen Abbildung. Vgl. Lax (1997, S. 190 f.).

### 2.11.1 Vektornorminduzierte Matrixnorm

Es seien $\|\cdot\|_\alpha$ auf $\mathbb{R}^n$ und $\|\cdot\|_\beta$ auf $\mathbb{R}^m$ beliebige Vektornormen und $A \in \mathbb{R}^{m\times n}$ die darstellende Matrix einer linearen Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^m$. Dann definiert

$$\|A\|_{\alpha,\beta} := \sup_{x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha}$$

die durch die Vektornormen $\|\cdot\|_\alpha$ und $\|\cdot\|_\beta$ induzierte Matrixnorm $\|\cdot\|_{\alpha,\beta}$ (*least upper bound norm*) in $\mathbb{R}^{m\times n}$. Also ist $\|A\|_{\alpha,\beta}$ die kleinste reelle Zahl mit $\|Ax\|_\beta \leq \|A\|_{\alpha,\beta}\|x\|_\alpha$ für alle $x \in \mathbb{R}^n$. Da die Hülle der Einheitskugel $\{x \in \mathbb{R}^n : \|x\|_\alpha = 1\}$ kompakt und $\|\cdot\|_\beta$ stetig ist, folgt

$$\|A\|_{\alpha,\beta} = \max_{\|x\|_\alpha=1} \|Ax\|_\beta$$

(Golub und Van Loan, 1989, S. 57).

### 2.11.2 $p$-Norm

Die Matrix-$p$-Normen

$$\|A\|_p := \|A\|_{p,p} = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{x\neq 0} \left\| A\frac{x}{\|x\|_p} \right\|_p = \max_{\|x\|_p=1} \|Ax\|_p$$

in $\mathbb{R}^{m\times n}$ werden durch die Vektor-$p$-Normen in $\mathbb{R}^m$ und $\mathbb{R}^n$ induziert (Golub und Van Loan, 1989, S. 56).

### 2.11.3 Spaltensummennorm

Die durch die Vektor-1-Norm induzierte Matrix-1-Norm heißt auch Spaltensummennorm. Die Spaltensummennorm einer $m \times n$-Matrix ist die größte Spaltensumme der Beträge der Komponenten,

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$$

(Golub und Van Loan, 1989, S. 57; Horn und Johnson, 1985, S. 294).

### 2.11.4 Spektralnorm

Die durch die Vektor-2-Norm induzierte Matrix-2-Norm heißt Spektralnorm (HORN und JOHNSON, 1985, S. 295). Die Spektralnorm einer reellen $m \times n$-Matrix ist ihr größter Singulärwert,

$$\|A\|_2 = \max\{\sigma_i \text{ Singulärwert von } A : 1 \leq i \leq \min\{m, n\}\}$$

(GOLUB und VAN LOAN, 1989, S. 72).

### 2.11.5 Zeilensummennorm

Die durch die Vektor-$\infty$-Norm induzierte Matrix-$\infty$-Norm nennt man Zeilensummennorm, da sie für eine $m \times n$-Matrix einfach durch

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$$

bestimmt wird. Die Spaltensummennorm einer Matrix ist gleich der Zeilensummennorm ihrer Transponierten, $\|A\|_1 = \|A^T\|_\infty$ (GOLUB und VAN LOAN, 1989, S. 57; HORN und JOHNSON, 1985, S. 295).

### 2.11.6 Frobenius-Norm

Die FROBENIUS-Norm (SCHUR-Norm, HILBERT-SCHMIDT-Norm) einer $m \times n$-Matrix

$$\|A\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

ist die in der numerischen linearen Algebra am häufigsten verwendete Matrixnorm. Sie ist die Vektor-2-Norm angewandt auf Matrizen, $\|A\|_F = \|A\|_2 = \sqrt{\text{Spur}(A^T A)}$. Das Quadrat der FROBENIUS-Norm einer $m \times n$-Matrix ist gleich der Summe der Quadrate ihrer Singulärwerte $\sigma_1, \sigma_2, \ldots \sigma_p$, $p = \min\{m, n\}$,

$$\|A\|_F^2 = \sum_{i=1}^{p} \sigma_i^2$$

(GOLUB und VAN LOAN, 1989, S. 56, 72; HORN und JOHNSON, 1985, S. 291).

## 2.12 Deskriptiv-statistische Grundbegriffe

### 2.12.1 Untersuchungseinheit

Die Objekte, auf die sich eine statistische Analyse beziehen, heißen Untersuchungseinheiten $\omega$ (statistische Einheiten, Merkmalsträger). Die Menge aller Untersuchungseinheiten heißt Grundgesamtheit $\Omega$ (Untersuchungsgesamtheit, Auswahlgesamtheit, Grundmenge, Population, statistische Masse). Die Anzahl der Merkmalsträger nennt man Umfang der Erhebung.

### 2.12.2 Merkmal und Ausprägung

Eine Eigenschaft einer Untersuchungseinheit heißt Merkmal (statistische Variable, Attribut). Jeder möglichen Ausprägung (Merkmalsausprägung, Abstufung, Wert) eines Merkmals wird genau eine Zahl einer Skala zugeordnet, so daß die Verhältnisse der Zahlen die Verhältnisse der Ausprägungen wiedergeben. Diese Menge aller möglichen Ausprägungen heißt Merkmalsraum (Zustandsraum, statistische Daten, Datensatz). Die an einer Untersuchungseinheit tatsächlich beobachteten Ausprägungen heißen Beobachtungen oder Realisationen eines Merkmals. Ein Merkmal heißt extensiv, falls die Summe der Merkmalsausprägungen sachlich sinnvoll interpretiert werden kann und intensive, falls nur Mittelwerte, nicht aber Summen interpretierbar sind.

### 2.12.3 Messung

Eine Messung ist die Zuordnung einer Merkmalsausprägung $x$ aus einer Skala $S$ zu einer Untersuchungseinheit $\omega \in \Omega$ aufgrund einer durch Regeln definierten Operation. Ein Merkmal wird als eine Abbildung $X : \Omega \to S$, $\omega \mapsto X(\omega) = x$ modelliert. Werden mehrere Merkmale $X_1, X_2, \ldots, X_n$, $n \in \mathbb{N}$, gleichzeitig untersucht, faßt man sie zu einem Vektor

$$\begin{aligned} X := (X_1, X_2, \ldots, X_n) : \Omega &\to S_1 \times S_2 \times \cdots \times S_n, \\ \omega &\mapsto \big(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\big) = (x_1, x_2, \ldots, x_n) =: x \end{aligned}$$

zusammen. Dann spricht man von einem multivariaten (mehrdimensionalen) Datensatz. Wird nur ein einziges Merkmal untersucht, heißen die Daten univariat.

## 2.13 Lagemaße

Datensätze werden durch einige wenige Kenngrößen (Maßzahlen, Parameter) charakterisiert, die die wesentlichen Eigenarten der Daten widerspiegeln. Lageparameter sind Maßzahlen, die die Lage des gesamten Datensatzes auf der Skala der Merkmalsausprägungen kennzeichnen.

### 2.13.1 Arithmetisches Mittel

Es seien $x_1, x_2, \ldots, x_n$ beliebige reelle Zahlen. Jedes $x_i$ sei gewichtet mit einer nichtnegativen reellen Zahl $w_i$, so daß die Summe aller Gewichte größer 0 ist. Das gewogene arithmetische Mittel ist

$$\bar{x} := \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \frac{w_i}{\sum_{i=1}^{n} w_i} x_i.$$

Mit normierten Gewichten, $\sum_{i=1}^{n} w_i = 1$, ist es

$$\bar{x} = \sum_{i=1}^{n} w_i x_i.$$

Sind alle Gewichte gleich groß, ergibt sich das (gewöhnliche) arithmetische Mittel

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Das arithmetische Mittel ist stabil gegenüber linearen Transformationen, das heißt, der linear transformierte Mittelwert ist gleich dem arithmetischen Mittel der auf gleiche Weise transformierten Werte. Die Gewichte dürfen dabei nicht verändert werden. Sie werden nicht transformiert.

Das gewöhnliche arithmetische Mittel minimiert die Quadratsumme der Abweichungen von den Einzelwerten,

$$\bar{x} = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} (x_i - a)^2.$$

Die Summe der quadrierten Abweichungen der Werte von ihrem arithmetischen Mittel ist kleiner als die Summe der quadrierten Abweichungen von jedem anderen Wert (Diehl und Kohr, 1983, S. 74). Siehe auch Ferschl (1978, S. 48 f.).

### 2.13.2 Geometrisches Mittel

Es seien $x_1, x_2, \ldots, x_n$ positive reelle Zahlen. Jedes $x_i$ sei gewichtet durch eine nichtnegative reelle Zahl $w_i$. Zusätzlich seien die Gewichte so gewählt, das ihre Summe größer 0 ist. Das gewogene geometrische Mittel ist

$$\bar{x}_G := G := \left(\prod_{i=1}^{n} x_i^{w_i}\right)^{\frac{1}{\sum_{i=1}^{n} w_i}}.$$

Sind die Gewichte normiert, $\sum_{i=1}^{n} w_i = 1$, so gilt

$$G = \prod_{i=1}^{n} x_i^{w_i}.$$

Sind alle Gewichte gleich, ergibt sich das (gewöhnliche) geometrische Mittel

$$G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Der Logarithmus des geometrischen Mittels ist gleich dem arithmetischen Mittel der logarithmierten Einzelwerte. Das geometrische Mittel ist für Werte geeignet, von denen das Produkt interpretierbar ist. Es wird für Verhältniszahlen verwendet. Mit ihm werden meist durchschnittliche Wachstumsraten berechnet. Vgl. Ferschl (1978, S. 58 f.).

### 2.13.3 Harmonisches Mittel

Es seien $x_1, x_2, \ldots, x_n$ positive reelle Zahlen. Jedes $x_i$ sei gewichtet mit einer nichtnegativen reellen Zahl $w_i$. Wenigstens eines der Gewichte jedoch sei größer als 0. Das gewogene harmonische Mittel ist

$$\bar{x}_H := H := \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Mit normierten Gewichten, $\sum_{i=1}^{n} w_i = 1$, ist es

$$H = \frac{1}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Sind alle Gewichte gleich, ergibt sich das (gewöhnliche) harmonische Mittel

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Der Reziprokwert des gewogenen harmonischen Mittels ist gleich dem gewogenen arithmetischen Mittel der Reziprokwerte der Einzelwerte. Die Gewichte bleiben unverändert. Vgl. Ferschl (1978, S. 61).

### 2.13.4 Quantile

Es seien $x_1, x_2, \ldots, x_n$ reelle Zahlen. Ferner sei $\alpha$ eine reelle Zahl echt zwischen 0 und 1. Eine jede reelle Zahl $\tilde{x}_\alpha$ heißt ein $\alpha$-Quantil ($\alpha$-Fraktil, $\alpha$-Punkt), wenn ein Anteil von mindestens $\alpha$ der Zahlen kleiner oder gleich $\tilde{x}_\alpha$ und ein Anteil von mindestens $1 - \alpha$ der Zahlen größer oder gleich $\tilde{x}_\alpha$ ist (Heiler und Michels, 1994, S. 96). Quantile sind nicht eindeutig bestimmt. Zu jedem $\alpha$ gibt es eine Menge von $\alpha$-Quantilen. $\tilde{x}_{0,5}$ heißt ein Median, $\tilde{x}_{0,25}$ ein unteres Quartil, $\tilde{x}_{0,75}$ ein oberes Quartil, $\tilde{x}_{0,1}, \tilde{x}_{0,2}, \ldots, \tilde{x}_{0,9}$ heißen Dezile und $\tilde{x}_{0,01}, \tilde{x}_{0,02}, \ldots, \tilde{x}_{0,99}$ Perzentile.

**Beispiele**

Es seien $x_1 \leq x_2 \leq \ldots \leq x_n$ der Größe nach aufsteigend geordnete reelle Zahlen und $0 < \alpha < 1$.

(a) Durch $\tilde{x}_\alpha := x_{[n\alpha]+1}$ wird eine der gegebenen Zahlen als $\alpha$-Quantil gewählt, wobei $[x]$ die größte ganze Zahl, die kleiner oder gleich $x$ ist, bezeichnet (Abschnitt 2.1.3).

(b) Im Falle eines ganzzahligen $n\alpha$ ist jeder Wert aus dem Intervall $[x_{n\alpha}, x_{n\alpha+1}]$ ein $\alpha$-Quantil. Mit

$$\tilde{x}_\alpha := \begin{cases} x_{[n\alpha]+1}, & \text{falls } n\alpha \text{ keine ganze Zahl ist,} \\ \dfrac{x_{n\alpha} + x_{n\alpha+1}}{2}, & \text{falls } n\alpha \text{ ganzzahlig ist,} \end{cases}$$

wird die Intervallmitte gewählt.

### 2.13.5 Median

Einer allgemeinen, die Definition (Abschnitt 2.13.4) einschränkenden Konvention folgend ist der Median $\tilde{x}$ der nach Größe geordneten Werte $x_1 \leq x_2 \leq \ldots \leq x_n$ bei ungerader Anzahl $n = 2m + 1$, $m \in \mathbb{N}$, der mittlere Wert $\tilde{x} = x_{m+1}$ und bei gerader Anzahl $n = 2m$ das arithmetische Mittel der beiden mittleren Werte $\tilde{x} = (x_m + x_{m+1})/2$. Er halbiert die Menge der Werte. Der Median minimiert die Summe der absoluten Abweichungen,

$$\min_{a \in \mathbb{R}} \sum_{i=1}^{n} |x_i - a| = \sum_{i=1}^{n} |x_i - \tilde{x}|.$$

Bei ungeradem $n$ ist der Median die eine eindeutige Lösung. Bei geradem $n$ sind alle Zahlen des zentralen Intervalls $[x_m, x_{m+1}]$ Lösungen des Minimierungsproblems (Ferschl, 1978, S. 70).

### 2.13.6 Modus

Der Modus $\hat{x}$ (Modalwert, *mode*) der Werte $x_1, x_2, \ldots, x_n$ ist derjenige, der am häufigsten vorkommt. Gibt es derer mehrere, ist der Modus nicht eindeutig bestimmt. Etwas allgemeiner definiert ist der Modus derjenige Punkt, bei dem eine Eigenschaft $f$ am deutlichsten ausgeprägt ist. Die Funktion $f$ hat an der Stelle $\hat{x}$ ein Extremum. Der Modus ist also eine Extremstelle. Bei der klassischen

Definition mißt die Funktion die Häufigkeit des Vorkommens eines jeden Wertes. Eine Menge, bei der der Modus das einzige lokale und damit globale Extremum ist, heißt unimodal.

## 2.14 Streuungsmaße

Streuungsmaße charakterisieren die Abstände zwischen den Elementen einer Menge. Sie messen, wie eng die Elemente zusammen liegen bzw. wie weit verstreut sie sind.

### 2.14.1 Spannweite

Die Spannweite (Variationsweite, *range*) einer nichtleeren Menge $X$ reeller Werte ist die Differenz zwischen dem größten Wert $x_{\max}$ und dem kleinsten Wert $x_{\min}$ der Menge, $\operatorname{range}(x) := x_{\max} - x_{\min}$.

### 2.14.2 Varianz

Es seien $x_1, x_2, \ldots, x_n$ beliebige reelle Werte. Die Varianz $\operatorname{var}(x)$ der Werte ist das arithmetische Mittel[1] der quadrierten Abweichungen der Werte von ihrem arithmetischen Mittel,

$$s^2 := \operatorname{var}(x) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(Ferschl, 1985, S. 92 ff.). Sie ist ein Maß für die Streuung (*spread, dispersion*) um den Mittelwert. Ein kleiner Wert der Varianz gibt an, daß die Werte dicht am Mittelwert konzentriert sind. Ein großer Wert besagt, daß sie weit um den Mittelwert verstreut liegen.

[1] Die Varianz aller Elemente der Grundgesamtheit wird bestimmt. Handelt es sich bei den Elementen hingegen um eine Stichprobe und soll ein Schätzwert für die Varianz der Grundgesamtheit berechnet werden, wird $n-1$ anstelle von $n$ im Nenner verwendet. Dann wird durch die Anzahl der voneinander unabhängigen Abweichungen dividiert. Nach der Berechnung des Mittelwertes sind von den $n$ Einzelwerten nämlich nur noch $n-1$ frei wählbar. Die Zahl $n-1$ nennt man deshalb Freiheitsgrad. Mit $n-1$ ist es ein erwartungstreuer (unverzerrter, *unbiased*) Schätzer für die unbekannte Varianz einer wahrscheinlichkeitstheoretischen Verteilung.

### 2.14.3 Standardabweichung

Die Varianz beschreibt die quadrierten Fehler der Werte bezüglich des Mittelwertes. Durch das Wurzelziehen wird die Quadrierung „rückgängig gemacht“, so daß die Standardabweichung die gleiche Maßeinheit wie die Daten selbst hat. Die positive Quadratwurzel der Varianz heißt Standardabweichung

$$s = \sqrt{s^2} = \sqrt{\operatorname{var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Mittelwert und Standardabweichung sind die wichtigsten Maßzahlen, um eine Aussage über die Einheitlichkeit einer Menge vergleichbarer Objekte durch die Verteilung der Werte der sie charakterisierenden Merkmalsausprägungen zu machen.

## 2.15 Konzentration

Die Messung der Konzentration setzt ein extensives Merkmal mit nichtnegativen Ausprägungen voraus. Konzentrationsparameter beschreiben, wie die Gesamtsumme der Merkmalsausprägungen auf die einzelnen Merkmalsträger verteilt ist. Absolute Konzentration bezieht die Anteile an der Merkmalssumme auf eine Anzahl von Merkmalsträgern. Relative Konzentration (Disparität) bezieht die Anteile an der Merkmalssumme auf einen Anteil von Merkmalsträgern an deren Gesamtanzahl.

### 2.15.1 Lorenzkurve

Es seien $0 \leq x_1 \leq x_2 \leq \ldots \leq x_n$ der Größe nach aufsteigend geordnete nichtnegative reelle Zahlen. Ihre Summe sei echt größer als 0. Der kumulierte Anteil der $k$ kleinsten Werte an der Gesamtsumme ist

$$v_k := \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i}.$$

Der Anteil der $k$ Werte an ihrer Gesamtzahl $n$ ist $u_k := k/n$. Der Anteilswert $v_k$ wird über dem Anteil $u_k$ in ein $(u, v)$-Koordinatensystem abgetragen. Den Streckenzug durch die $n + 1$ Punkte

$$(0, 0) =: (u_0, v_0), (u_1, v_1), \ldots, (u_n, v_n) = (1, 1)$$

heißt LORENZkurve. Da sich die Anzahl diskret verändert, können strenggenommen nur die Punkte $(u_k, v_k)$ inhaltlich interpretiert werden. Der Punkt $(u_k, v_k)$ besagt, daß auf $u_k$ der kleinsten Werte $v_k$, auf die übrigen $1 - u_k$ hingegen $1 - v_k$ der Gesamtsumme entfallen.

## 2.16 Standardisierung

Die Transformation unterschiedlich skalierter Zahlenwerte in einen einheitlichen Wertebereich heißt Standardisierung ($z$-Transformation) und die transformierten Werte Standardabweichungseinheiten (Standardwerte, *standard scores*). Es seien $x_1, x_2, \ldots, x_n$ reelle Zahlen. Die standardisierte Größe $z_i$ zum Rohwert $x_i$ ist

$$z_i := \frac{x_i - \bar{x}}{s_x}.$$

Die Abweichungen der Werte vom Mittelwert $\bar{x}$ werden in Standardabweichungseinheiten ausgedrückt, das heißt in Vielfachen der Standardabweichung $s_x$. Das arithmetische Mittel standardisierter Daten ist 0, deren Varianz 1.

## 2.17 Ähnlichkeitsmaße

Es sei $X$ eine nichtleere Menge. Ein Ähnlichkeitsmaß (*proximity measure*) ist eine Abbildung

$$s : X \times X \to \mathbb{R},\ (x, y) \mapsto s(x, y),$$

die je zwei Punkten $x$ und $y$ aus $X$ eine reelle Zahl $s(x, y)$ zuordnet. Zumindest eine dieser Zahlen ist besonders ausgezeichnet. Sie beschreibt größtmögliche Ähnlichkeit oder kleinstmöglichen Abstand. Je näher ein Funktionswert dieser ausgezeichneten Zahl ist, desto ähnlicher sind sich die Punkte. Jedoch hängt die Ähnlichkeit stets auch von der globalen Homogenität der Grundgesamtheit ab. Nimmt ein Paar von Punkten einen solchen Wert an, so sind diese

lediglich bezüglich dieses Maßes gleich. Im allgemeinen darf daraus nicht geschlossen werden, daß die Punkte aufeinander fallen. Für Vektoren setzt ein Ähnlichkeitsmaß die Übereinstimmungen zu den Abweichungen in den Komponenten in Beziehung. Ähnlichkeitsmaße sollten anschaulich interpretierbar sein. Es werden Distanzmaße, Ähnlichkeitsmaße im engeren Sinne[2] und Assoziationsmaße unterschieden.

### 2.17.1 Distanzmaße

Es sei $X$ eine nichtleere Menge. Eine Funktion

$$d : X \times X \to \mathbb{R},\ (x, y) \mapsto d(x, y),$$

die jedem Paar von Elementen aus $X$ eine reelle Zahl $d(x, y)$ zuordnet, heißt Distanzmaß (Distanzfunktion, Abstandsfunktion) auf $X$, wenn folgende Axiome[3] gelten:

(D1) $d(x, y) \geq 0$ für alle $x, y \in X$.

(D2) $d(x, x) = 0$ für jedes $x \in X$.

(D3) $d(x, y) = d(y, x)$ für alle $x, y \in X$.

Der nichtnegative reelle Wert $d(x, y)$ heißt Abstand oder Distanz der Elemente $x$ und $y$. Nach (D1) ist die kleinstmögliche Distanz 0. Wegen (D1) und (D2) ist $d$ minimal, falls beide Argumente gleich sind. Distanzmaße sind symmetrisch (D3). Eine obere Schranke für den Wertebereich von $d$ gibt es im allgemeinen nicht. Ein Distanzmaß wird zu einer Metrik (Abschnitt 2.8) und heißt metrisch, falls zusätzlich gilt:

(D4) Aus $d(x, y) = 0$ folgt $x = y$.

(D5) $d(x, z) \leq d(x, y) + d(y, z)$ für alle $x, y, z \in X$.

(D4) fordert, daß der kleinste Wert des Distanzmaßes immer nur dann angenommen wird, wenn beide Argumente gleich sind. (D2) und (D4) zusammen bilden das Identitätsaxiom. (D5) ist die Dreiecksungleichung. Ist darüber hinaus noch die ultrametrische Ungleichung,

[2] Das Wort Ähnlichkeitsmaß dient auch als Oberbegriff (*proximity measure*) für Ähnlichkeitsmaße (*similarity measure*), Distanzmaße und Assoziationsmaße.

[3] Die Axiome können allgemeiner mit einem beliebigen, aber festen Wert $d_0 \in \mathbb{R}$ für minimale Distanz anstelle von 0 formuliert werden.

(D6) $d(x,z) \leq \max\big(d(x,y), d(y,z)\big)$ für alle $x,y,z \in X$,

erfüllt, handelt es sich um eine Ultrametrik. Vgl. ECKES und ROSSBACH (1980, S. 38) und SPÄTH (1975, S. 14 f.).

### 2.17.2 Ähnlichkeitsmaße

Es sei $X$ eine nichtleere Menge. Eine Funktion

$$s : X \times X \to \mathbb{R},\ (x,y) \mapsto s(x,y),$$

heißt Ähnlichkeitsmaß (Ähnlichkeitsfunktion) auf $X$, wenn folgende Axiome[4] gelten:

(S1) $s(x,y) \leq 1$ für alle $x,y \in X$.

(S2) $s(x,x) = 1$ für alle $x \in X$.

(S3) $s(x,y) = s(y,x)$ für alle $x,y \in X$.

Wegen (S1) und (S2) wird $s$ maximal, wenn das Paar aus gleichen Elementen besteht. Ähnlichkeitsmaße sind symmetrisch (S3). Der Wert $s(x,y)$ heißt Ähnlichkeit der Elemente $x$ und $y$. Ein Ähnlichkeitsmaß heißt metrisch, falls gilt:

(S4) Aus $s(x,y) = 1$ folgt $x = y$.

(S5) $s(x,z) \geq s(x,y) \cdot s(y,z)$ für alle $x,y,z \in X$.

Nach (S4) besitzen nur zwei gleiche Elemente maximale Ähnlichkeit. (S2) und (S4) zusammen bilden das Identitätsaxiom. Die Forderung (S5) ist gerade so eingerichtet, daß Analogie zu (D5) besteht. Die der ultrametrischen Ungleichung entsprechenden Eigenschaft lautet

(S6) $s(x,z) \geq \min\big(s(x,y), s(y,z)\big)$ für alle $x,y,z \in X$.

Es ist üblich, Ähnlichkeitsmaße so zu konstruieren, daß sie nur Werte aus dem Einheitsintervall annehmen. 0 gibt dann die kleinste und 1 die größte Ähnlichkeit an. Vgl. ECKES und ROSSBACH (1980, S. 39) und SPÄTH (1975, S. 15).

[4] Die Axiome können auch allgemeiner mit einem beliebigen, aber festen Wert $s_0 \in \mathbb{R}$ für maximale Ähnlichkeit anstelle von 1 formuliert werden.

### 2.17.3 Assoziationsmaße

Es sei $X$ eine nichtleere Menge. Ein Assoziationsmaß (Kontingenz-, Korrelationskoeffizient)

$$r : X \times X \to \mathbb{R},\ (x, y) \mapsto r(x, y),$$

muß symmetrisch,

(R1) $r(x, y) = r(y, x)$ für alle $x, y \in X$,

sein. Es beschreibt die Anziehung oder Abstoßung zweier Elemente[5]. Zwei Elemente stoßen sich ab, wenn ihnen ein negativer Wert zugeordnet ist. Sie ziehen sich an, wenn sie auf einen positiven Wert abgebildet werden. Je größer der Betrag des Funktionswertes ist, desto stärker ist die jeweilige Ausprägung. Ist er 0, halten sich Abstoßung und Anziehung die Waage. Der Wertebereich ist häufig auf $[-1, 1]$ beschränkt, wobei 0 das Equilibrium, $-1$ eine perfekte Abstoßung und $+1$ eine perfekte Anziehung charakterisieren.

### 2.17.4 Skalentransformation

Beschränkte Distanz-, Ähnlichkeits- und Assoziationsmaße können ineinander überführt werden. Es seien hierzu $X := [x_{\min}, x_{\max}]$ und $Y := [y_{\min}, y_{\max}]$ abgeschlossene Intervalle und $x_{\text{span}} := x_{\max} - x_{\min}$ und $y_{\text{span}} := y_{\max} - y_{\min}$ deren Längen. Das Einheitsintervall $E := [0, 1]$ hat die Länge 1. Die Skalierung eines abgeschlossenen Intervalls auf das Einheitsintervall heißt Normierung (*normalization*) des Intervalls. Die lineare Transformation[6]

$$t_{EX} : X \to [0, 1],\ x \mapsto \frac{x - x_{\min}}{x_{\text{span}}}$$

[5] Eigentlich mißt ein empirisches Assoziationsmaß den Zusammenhang zwischen zwei statistischen Variablen anhand von Beobachtungsvektoren. Für jede Variable werden die Meßwerte der Fälle in Vektoren festgehalten und verglichen. Die statistischen Variablen sind die Merkmale und die Fälle die Untersuchungseinheiten. Assoziationsmaße werden hier ein wenig anders verwendet. Es werden nicht die Merkmale, sondern die Untersuchungseinheiten verglichen. Die Fälle sind nun nicht mehr die Untersuchungseinheiten, sondern die Merkmale. Merkmale und Untersuchungseinheiten tauschen ihre Rollen. Rein formal ändert sich nichts. Es werden stets zwei reelle Vektoren verglichen.

[6] Der Schreibweise bei Verknüpfungen $\circ$ wegen steht im Index des Funktionsnamens erst das Ziel und dann die Quelle.

normiert das Intervall $X$ auf das Einheitsintervall. Die lineare Transformation

$$t_{YE} : [0,1] \to Y,\ x \mapsto x\, y_{\text{span}} + y_{\min}$$

skaliert das Einheitsintervall auf das Intervall $Y$. Die Skalierung von $X$ auf $Y$ erfolgt durch die Verknüpfung

$$t_{YX} : X \to Y,\ x \mapsto (t_{YE} \circ t_{EX})(x) = \frac{x - x_{\min}}{x_{\text{span}}} y_{\text{span}} + y_{\min}.$$

Zunächst wird $X$ auf das Einheitsintervall und dann das Einheitsintervall auf $Y$ übertragen. Die streng monoton fallende Bijektion

$$u_{EE} : [0,1] \to [0,1],\ x \mapsto 1 - x$$

bildet kleine Werte auf große und große auf kleine so ab, daß der Wert der Summe von Bild $u_{EE}(x)$ und Urbild $x$ stets 1 ist. Sie „dreht“ das Einheitsintervall „um.“ Gleiches macht

$$u_{XY} : X \to Y,\ x \mapsto (t_{YE} \circ u_{EE} \circ t_{EX})(x) = \left(1 - \frac{x - x_{\min}}{x_{\text{span}}}\right) y_{\text{span}} + y_{\min}$$

mit beliebigen abgeschlossenen Intervallen. Etwas weniger allgemein geben Gordon (1981, S. 13 ff.) und Steinhausen und Langer (1977, S. 52 ff.) einige Beispiele an.

### 2.17.5 Tanimoto

Tanimoto (1958) definiert die Ähnlichkeit von zwei nichtleeren endlichen Mengen $A$ und $B$ als das Verhältnis der Anzahl der Elemente, die in beiden Mengen vorkommen, zu der Anzahl der Elemente, die nur in einer der beiden Mengen enthalten sind,

$$(A,B) \mapsto \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \in [0,1].$$

Zwei disjunkten Mengen wird 0 zugeordnet, weil minimale Ähnlichkeit vorliegt. Sind die Mengen gleich, liegt also maximale Ähnlichkeit vor, nimmt das Maß 1 an.

Sind den Elementen eineindeutig Indizes zugeordnet, so lassen sich die Komponenten entsprechender Vektoren als den Elementen zugeordnete reelle Gewichte begreifen. Man erhält das allgemeinere Maß

$$(x,y) \mapsto \frac{\langle x, y\rangle}{\|x\|^2 + \|y\|^2 - \langle x, y\rangle} = \frac{\langle x, y\rangle}{\langle x, x\rangle + \langle y, y\rangle - \langle x, y\rangle} \in [0, \infty)$$

mit dem kanonischen Skalarprodukt und der aus diesem induzierten euklidischen Norm. Minimale Ähnlichkeit wird durch 0 angezeigt. Eine obere Schranke für maximale Ähnlichkeit gibt es nicht. Für binäre Vektoren stimmen beide Definitionen überein.

### 2.17.6 Kovarianz

Es seien $x := (x_1, \ldots, x_n)$ und $y := (y_1, \ldots, y_n)$ die reellen Beobachtungsvektoren zweier Merkmale. Die Kovarianz $\operatorname{cov}(x, y)$ von $x$ und $y$ ist das arithmetische Mittel der Produkte der Abweichungen der zu Paaren zusammengefaßten Meßpunkte $(x_i, y_i)$ von ihren jeweiligen Mittelwerten $\bar{x}$ und $\bar{y}$,

$$s_{xy} := \operatorname{cov}(x, y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Die Beobachtungsvektoren $x$ und $y$ heißen unkorreliert, falls $\operatorname{cov}(x, y) = 0$. Die Merkmale sind dann maximal fremd, weil keines auf das anderen einwirkt. Stets ist $\operatorname{cov}(x, x) = \operatorname{var}(x)$. Aus der CAUCHY-SCHWARZschen-Ungleichung ergibt sich für den Betrag der Kovarianz $|\operatorname{cov}(x, y)| \leq \operatorname{var}(x)\operatorname{var}(y)$. Die Kovarianz ist ein Assoziationsmaß.

### 2.17.7 Korrelationskoeffizient

#### 2.17.7.1 Einfache lineare Regression

Eine Korrelationsanalyse bestimmt den Grad der linearen Abhängigkeiten zwischen zwei Merkmalen anhand von Meßwerten. Es seien $(x_i, y_i)$, $i = 1, \ldots, n$, die Ausprägungen der Merkmale $X$ und $Y$. Das Merkmal $Y$ sei linear abhängig von $X$. Dieser lineare Zusammenhang sei durch die Gerade $Y = a + bX$ beschrieben. Der Steigungsparameter $b$ gibt an, um wieviel Einheiten sich $Y$ verändert, wenn $X$ um eine Einheit zunimmt. Der Achsenabschnitt $a$ zeigt an, welchen Wert $Y$ an der Stelle $X = 0$ annimmt. Die Abweichung vom unterstellten

linearen Zusammenhang wird durch die Fehlervariable $R$ beschrieben, $Y = a + bX + R$. Für die Beobachtungen wird dies durch $y_i = a + bx_i + r_i$ ausgedrückt. Die den linearen Zusammenhang beschreibende Regressionsgerade ist so zu bestimmen, daß der Fehler minimal ist. Als Fehlerfunktion wird nicht die Summe der Beträge, sondern die Summe der Quadrate der Abweichungen (Residuen) verwendet, weil sie stetig differenzierbar ist. Die Ableitungen der mit Beträgen gebildeten Fehlerfunktion können nicht analytisch behandelt werden, weil sie nicht stetig sind. Allerdings haben bei der Verwendung von Quadraten Ausreißer einen unverhältnismäßig großen Einfluß auf die Anpassung. Zur Minimierung des Fehlers werden $a$ und $b$ so ermittelt, daß die Gesamtsumme der quadrierten Abstände zwischen den gegebenen $y_i$ und deren durch $a$ und $b$ beschriebenen Idealpositionen so klein wie möglich sind. Nach dem Prinzip der kleinsten Quadrate minimieren

$$\hat{b} := \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2} \qquad \text{und} \qquad \hat{a} := \bar{y} - b\bar{x}$$

die Summe der quadrierten Fehler

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Die Beziehung lautet dann $\hat{y}_i := \hat{a} + \hat{b}x_i$. Zu jedem $x_i$ gehört nicht nur ein Datenwert $y_i$, sondern auch ein systematischer Wert $\hat{y}_i$. Die systematischen Werte $\hat{y}_i$ liegen auf der von $\hat{a}$ und $\hat{b}$ bestimmten Gerade. Die Kleinstquadrateresiduen sind nun $\hat{r}_i := y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i = (y_i - \bar{y}) - \hat{b}(x_i - \bar{x})$. Der Koeffizient $\hat{b}'$ für die Regression der linearen Abhängigkeit von $X$ von $Y$ mit $X = a' + b'Y$ ist $\hat{b}' = \operatorname{cov}(x,y)/\operatorname{var}(y)$. Vgl. Heiler und Michels (1994, S. 244 ff.).

#### 2.17.7.2 Korrelationskoeffizient

Der empirische Korrelationskoeffizient $\operatorname{cor}(x, y)$ der Meßwertvektoren $x := (x_1, x_2, \ldots, x_n)$ und $y := (y_1, y_2, \ldots, y_n)$ ist definiert als das geometrische Mittel $r := \operatorname{cor}(x, y) := \sqrt{\hat{b}\hat{b}'}$ der beiden Regressionskoeffizienten,

$$\operatorname{cor}(x, y) = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \in [-1, 1].$$

Er mißt den Grad des linearen Zusammenhangs der konkreten Meßwertvektoren für zwei Merkmale. Der Korrelationskoeffizient bestimmt somit den Cosinus des Winkels zwischen den Vektoren der standardisierten Abweichungen der Meßwerte von ihren jeweiligen arithmetischen Mitteln. Durch die Standardisierung (Abschnitt 2.16) werden die Daten erst vergleichbar. Der Korrelationskoeffizient ist unabhängig sowohl vom Ursprung als auch von der Skalierung. Deswegen ist der Korrelationskoeffizient ein brauchbares Maß, wenn die „Größe“ weniger wichtig ist als die „Form“.

Der Korrelationskoeffizient ist ein Assoziationsmaß. Er gibt an, wie stark zwei Variablen sich gleichartig oder entgegengesetzt verändern. Variieren sie gleichartig, spricht man von positiver Korrelation. Je stärker ein Merkmal ausgeprägt ist desto stärker ist auch das andere. Die Steigung $r$ der Geraden, die diese Beziehung repräsentiert, ist positiv. Bei maximaler positiver Korrelation ist $r = 1$. Verändern sie sich gegenläufig, spricht man von negativer Korrelation. Je stärker ein Merkmal ausgeprägt ist desto schwächer ist das andere. Die Gerade hat eine negative Steigung. Bei stärkster negativer Korrelation ist $r = -1$. Sind beide Merkmale voneinander unabhängig, so ist $r = 0$.

### 2.17.8 Cosinus

Es seien $x, y \in \mathbb{R}^n$, $x, y \neq 0$. Das Cosinusmaß

$$r_{\text{Cosinus}}(x, y) := \frac{\langle x, y\rangle}{\|x\|\,\|y\|} = \left\langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \right\rangle \in [-1, 1]$$

bestimmt den Cosinus des Winkels zwischen den Vektoren (Abschnitt 2.9). Es ist $-1$ für $x = -y$, $0$ für $x \perp y$ und $1$ für $x = y$. Das Cosinusmaß ist keine Metrik, weil es das Identitätsaxiom (M1′) nicht erfüllt. Es ist ein Assoziationsmaß.

### 2.17.9 Ähnlichkeitsmaße für binäre Vektoren

Vektoren, deren Komponenten nur die Werte 0 und 1 annehmen, heißen binäre Vektoren. Alle Ähnlichkeitsfunktionen für Vektoren reeller Komponenten gelten auch für binäre Vektoren. Ähnlichkeitsmaße für binäre Vektoren sind recht eingängig durch den Vergleich von Mengen motiviert. Die Ähnlichkeit von zwei nichtleeren endlichen Mengen wird definiert als die Anzahl der Elemente, die in beiden Mengen vorkommen. Die Elemente der Grundmenge, in der diese beiden Mengen liegen, seien eindeutig indiziert. Dann können die Mengen als Vektoren aus 0 und 1 beschrieben werden, wobei 0 angibt, daß das zu dieser Komponente gehörige Element nicht in der Menge liegt, und 1, daß es in der Menge liegt. Die Ähnlichkeiten berechnen sich aus den Anzahlen der übereinstimmenden und abweichenden Komponenten der binären Vektoren. Mit diesen Anzahlen wird eine vergleichende Differenz oder ein vergleichender Quotient gebildet.

#### 2.17.9.1 Kanonisches Skalarprodukt

Das kanonische Skalarprodukt

$$n_{11} := \langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \ \in \{0, 1, \ldots, n\}$$

binärer Vektoren $x, y \in \{0,1\}^n$ zählt die Komponenten, die bei beiden Vektoren 1 sind. Die Anzahl der Komponenten, die bei beiden Vektoren 0 sind, ist $n_{00} := \langle 1-x, 1-y \rangle$. Hierbei steht 1 für das $n$-Tupel $(1, 1, \ldots, 1)$ und $1-x$ somit für den Vektor $(1-x_1, 1-x_2, \ldots, 1-x_n)$. Die Anzahl der Komponenten, bei denen $x$ den Wert 0 und $y$ den Wert 1 hat, ist $n_{01} := \langle 1-x, y \rangle$. Die Anzahl derer, bei denen $x$ den Wert 1 und $y$ den Wert 0 hat, ist $n_{10} := \langle x, 1-y \rangle$.

Die Anzahl der 0 in $x$ ist $n_{0\cdot} := n_{00} + n_{01} = \sum_{i=1}^{n}(1-x_i)$. Die Anzahl der 1 in $x$ ist $n_{1\cdot} := n_{10} + n_{11} = \sum_{i=1}^{n} x_i$. Analog sind $n_{\cdot 0} := n_{00} + n_{10}$ und $n_{\cdot 1} := n_{01} + n_{11}$ erklärt. Diese Häufigkeiten heißen Randhäufigkeiten. Sie sagen nur etwas über einen einzelnen Vektor aus.

#### 2.17.9.2 Übereinstimmende Komponenten

Die Anzahl der Paare übereinstimmender Komponenten (*matched pairs*) ist $m := n_{00} + n_{11}$, die der nicht übereinstimmenden (*unmatched pairs*) $u := n_{01} + n_{10}$. Es gilt $n = n_{0\cdot} + n_{1\cdot} = n_{\cdot 0} + n_{\cdot 1} = n_{00} + n_{01} + n_{10} + n_{11} = m + u$ (Tabelle 2.1). Aus diesen Werten sind nahezu alle Ähnlichkeitsmaße für binäre

| | | $y$ | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| $x$ | 0 | $n_{00}$ | $n_{01}$ | $n_{0\cdot}$ |
| | 1 | $n_{10}$ | $n_{11}$ | $n_{1\cdot}$ |
| | | $n_{\cdot 0}$ | $n_{\cdot 1}$ | $n$ |

**Tabelle 2.1:** Kontingenztabelle der Anzahl übereinstimmender Komponenten zweier binärer Vektoren

Vektoren aufgebaut. Die Ähnlichkeitsmaße werden danach unterschieden, ob sie die negativen Übereinstimmungen einschließen oder nicht. Sokal und Sneath (1963, S. 125 ff.) vergleichen mehrere solcher Ähnlichkeitsmaße.

#### 2.17.9.3 Simple-Matching-Coefficient

Das Verhältnis der Anzahl der übereinstimmenden Merkmale zur Gesamtzahl aller Merkmale ist

$$s_{\text{simple matching}} := \frac{m}{n} = \frac{n_{00} + n_{11}}{n} = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} \in [0, 1].$$

Der Koeffizient ist 0, wenn es nicht ein einziges Paar gleicher Komponenten gibt, $m = 0$. Er ist 1, wenn es nur solche Paare gibt, $m = n$. Vgl. Sokal und Sneath (1963, S. 133).

#### 2.17.9.4 Jaccard und Tanimoto

Es sei $n_{00} < n$. Das Weglassen der negativen Übereinstimmungen im Simple-Matching-Coefficient ergibt

$$\begin{aligned} s_{\text{Jaccard}} &:= \frac{n_{11}}{u + n_{11}} = \frac{n_{11}}{n_{01} + n_{10} + n_{11}} \\ &= \frac{n_{11}}{n_{01} + n_{11} + n_{10} + n_{11} - n_{11}} \\ &= \frac{n_{11}}{n_{\cdot 1} + n_{1\cdot} - n_{11}} =: s_{\text{Tanimoto}} \in [0, 1]. \end{aligned}$$

Der Koeffizient ist 0, falls $n_{11} = 0$ und 1, falls $u = 0$. Vgl. Abschnitt 2.17.5 und Sokal und Sneath (1963, S. 133).

#### 2.17.9.5 Hamann

Der Assoziationskoeffizient von HAMANN

$$r_{\text{HAMANN}} := \frac{m-u}{n} = \frac{n_{00}+n_{11}-(n_{01}+n_{10})}{n} = \frac{n_{00}+n_{11}-n_{01}-n_{10}}{n_{00}+n_{01}+n_{10}+n_{11}} \in [-1,+1]$$

ergibt sich aus dem Simple-Matching-Coefficient, wenn der Zähler um die Anzahl der Nichtübereinstimmungen vermindert wird. Für $m = 0$ ist er $-1$. Für $m = u$ ist er 0 und für $u = 0$ ergibt sich 1. Vgl. CLIFFORD und STEPHENSON (1975, S. 54) und SOKAL und SNEATH (1963, S. 134).

#### 2.17.9.6 Cosinus

Es seien $x, y \neq 0$.

$$\begin{aligned} s_{\text{Cosinus}}(x,y) &:= \frac{n_{11}}{\sqrt{n_{1\cdot}\, n_{\cdot 1}}} = \frac{n_{11}}{\sqrt{(n_{10}+n_{11})\,(n_{01}+n_{11})}} \\ &= \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k^2}} = \cos\big(\measuredangle(x,y)\big) \in [0,1] \end{aligned}$$

gibt den Cosinus des Winkels zwischen den binären Vektoren $x$ und $y$ an. Zwei Vektoren aus $x, y \in \{0{,}1\}^n$ sind orthogonal, wenn $x_k = |1 - y_k|$ für alle $k = 1, 2, \ldots, n$ gilt. Der Koeffizient ist 0, wenn die Vektoren orthogonal sind, denn aus $n_{10} + n_{01} = n$ folgt $n_{11} = 0$. Er nimmt 1 an, wenn die Vektoren gleich sind, $n_{10} = n_{01} = 0$. Der Cosinus-Koeffizient ist das kanonische Skalarprodukt normiert durch das geometrische Mittel der Anzahl der Komponenten des einen und der des anderen Vektors. Für Vektoren mit beliebigen reellen Komponenten ist das Cosinusmaß ein Assoziationsmaß (Abschnitt 2.17.8). Hingegen ist er für binäre Vektoren ein Ähnlichkeitsmaß, weil alle Vektoren im ersten Quadranten liegen und somit kein Winkel größer als $\pi/2$ sein kann. Für nichtleere endliche Mengen $A$ und $B$ ist er

$$(A,B) \mapsto \frac{|A \cap B|}{\sqrt{|A||B|}}.$$

Vgl. SOKAL und SNEATH (1963, S. 130).

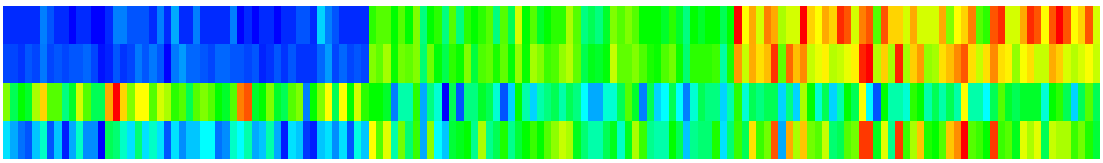
# Kapitel 3

# Clusteranalyseverfahren

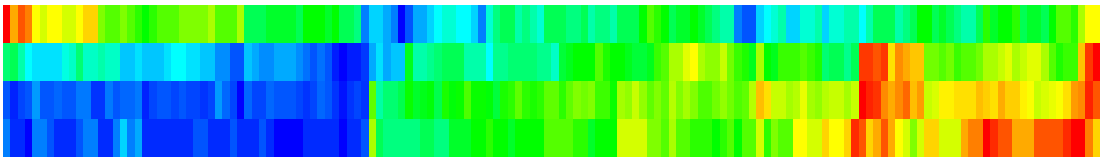
## 3.1 Datenbilder

Ein Datenbild (*data image*) (MINNOTTE und WEST, 1998) ist eine Visualisierung einer Matrix. Die Werte der Matrix werden auf Farbtöne eines Gradienten so abgebildet, daß ähnliche Farbtöne ähnlichen Werten entsprechen. Die Komponenten der Matrix werden als kleine farbig ausgefüllte Rechtecke zu einem großen Abbild der Matrix zusammengesetzt. Ein Datenbild vermittelt visuell einen Eindruck davon, wie häufig einzelne Werte vorkommen und wie ähnlich sie sind.

Werden die Zeilen und Spalten der Matrix so umgeordnet (*seriation*), daß möglichst viele benachbarte Komponenten möglichst kleine Differenzen aufweisen, entstehen zusammenhängende farbige Flächen. Bei einigen Datensätzen ist durch bloßes Betrachten des Datenbildes der Beobachtungsmatrix eine Gruppierung der Untersuchungseinheiten auszumachen. Dies läßt sich eingängig an einem Standardbeispiel der Clusteranalyse, dem Iris-Datensatz (Schwertlilien) von ANDERSON (1935), veranschaulichen (Abbildung 3.1). Er umfaßt jeweils vier Abmessungen der Blüten von 150 Iris-Exemplaren. Die Skalen der Merkmalsausprägungen sind unabhängig voneinander in das Einheitsintervall skaliert, so daß der kleinste Meßwert auf 0 und der größte auf 1 abgebildet wird. Das Einheitsintervall wird auf das Farbspektrum in Abbildung 3.1(c) übertragen. Der unteren Intervallgrenze ist blau und der oberen rot zugeordnet. Abbildung 3.1(a) zeigt das Datenbild einer manuellen Klassifikation der Pflanzen in die drei Unterarten Iris setosa, versicolor und virginica (von links nach rechts) anhand der vier Abmessungen der Blüten. Durch die Umordnung der Untersuchungseinheiten und Merkmale in der Beobachtungsmatrix nach



(a) Manuelle Klassifikation



(b) Automatische Klassifikation mit Average-Linkage

(c) Farbspektrum des Einheitsintervalls

**Abbildung 3.1:** Datenbilder des botanischen Iris-Datensatzes

Average-Linkage mit den euklidischen Abständen der Beobachtungsvektoren entsteht das Datenbild in Abbildung 3.1(b). Iris versicolor und virginica (rechts) bilden eigentlich nur eine von Iris setosa (links) linear getrennte Gruppe, sind aber intern nicht homogen, da Iris versicolor und virginica fließend ineinander übergehen.

## 3.2 Analyse

Eine Analyse ist eine Zergliederung eines Ganzen in seine Teile und die Untersuchung dieser Teile im Verhältnis zum Ganzen. Hierbei ist eine gewisse Vorstellung (Idee) über Art und mögliche Zusammenhänge von Teilen und Ganzem (Modell) leitend. Ziel ist es, die gegebenen Daten nach den ihnen zugrundeliegenden/sie charakterisierenden Eigenschaften oder signifikanten Beziehungen zu analysieren. Dies kann durch eine datengesteuerte oder eine modellgesteuerte Analyse erreicht werden. Eine modellgesteuerte Analyse überprüft, wie gut ein vorgegebenes Modell die Daten beschreibt. Das setzt voraus, daß die Struktur der Daten im wesentlichen bereits vor der Untersuchung bekannt ist. Einer datengesteuerten Analyse liegt hingegen nur eine vage Idee des Modells zugrunde. Die Struktur der Daten ist vor der Analyse weitgehend unbekannt. Es wird lediglich angenommen, daß eine Struktur existiert, die durch geeignete Verfahren aufgedeckt werden kann. Die Strukturen, die datengesteuerte Analysen erkennen können, sind im allgemeinen gröber

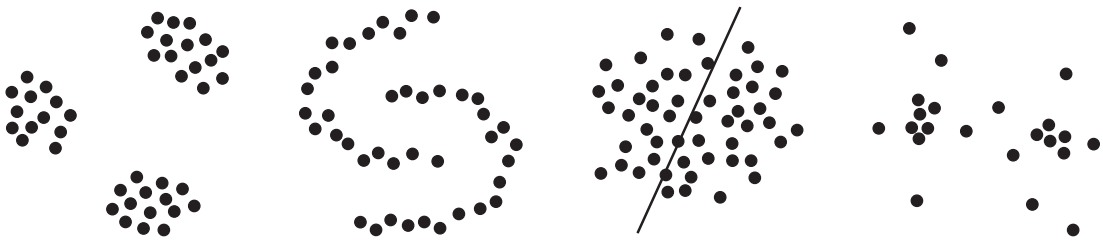



**Abbildung 3.2:** (a) drei kohäsive und isolierte Gruppen, (b) zwei isolierte, aber nicht kohäsive Gruppen, (c) zwei kohäsive, aber nicht isolierte Gruppen, (d) zwei Gebiete hoher Punktdichte ineinander übergehend durch umliegende verstreute Punkte

als solche, die modellgesteuerte Analysen zu identifizieren vermögen. Dafür unterliegt die modellgesteuerte Analyse viel restriktiveren Ausgangsbedingungen. Ob und welche Strukturen in den Daten erkannt werden können, ist bei beiden Analysearten zudem auch durch deren mögliche Repräsentationsformen bestimmt.

Eine Clusteranalyse ist eine deskriptive und explorative Datenanalyse, die kein Wissen über die Struktur der zu untersuchenden Daten voraussetzt. Sie ist somit primär datengesteuert. In ersten datengesteuerten Analyseschritten wird untersucht, ob die Daten überhaupt eine – wenn auch nur grobe – Struktur aufweisen. Die im Laufe der Untersuchungen schrittweise gewonnenen Informationen erlauben es mitunter, das Strukturmodell so sehr zu verfeinern, daß letztlich modellgesteuerte Analysen die Strukturen in den Daten nur noch lokalisieren. Dann sind Art, Form und Größe der in den Daten enthaltenen Typen von Clustern bekannt. Die Datenpunkte werden nur noch zu konkreten Clustern dieser Typen zusammengefaßt.

Clusterverfahren gehören zu den Mitteln der wissenschaftlichen Beschreibung. Sie können eine Ordnung in den Daten etablieren, die jedoch immer durch das Verfahren bestimmt ist und keinen Wahrheitswert besitzt. Sie können auch nicht als Überprüfungsinstanz von Hypothesen gewertet werden. Folglich kann von einer Clusteranalyse allein kein Erkenntnisgewinn erwartet werden.

## 3.3 Cluster

Ein Cluster (Gruppe, Klasse) ist eine Menge von Einzelteilen, die als Einheit betrachtet wird. Eine Clusteranalyse zerlegt eine Menge von Objekten (Beobachtungen, Punkten, Individuen, *samples*) in Cluster. Die Objekte werden aufgrund ihrer Ähnlichkeiten (*similarities, alikenesses*), Distanzen (*distances, dissimilarities*) oder Assoziationen (*associations, affinities*) gruppiert. Die Objekte innerhalb einer Gruppe sollten möglichst ähnlich sein und die Elemente unterschiedlicher Gruppen möglichst verschieden. Abbildung 3.2 (a) zeigt einen Idealfall kompakter und getrennter Cluster. Die klassische Sichtweise interner Homogenität (*internal cohesion*) und äußerer Abgrenzung (externe Isolation/ Heterogenität) ist jedoch keineswegs zwingend. Ein Cluster braucht weder isoliert noch kohäsiv zu sein (Abbildung 3.2 (b) – (d)). Die Punkte in Abbildung 3.2 (c) bilden eigentlich nur ein Cluster, werden aber in zwei ineinander übergehende Gruppen eingeteilt[1]. Gewöhnlich sind die Daten der Art wie in Abbildung 3.2 (d), in denen kleine, dichte Punkthäufungen in weiten, dünnbesiedelten Gebieten Clusterschwerpunkten entsprechen. Die Form eines Clusters hängt letztlich nur von der Struktur der konkret vorliegenden Daten und dem Untersuchungsziel ab.

Sowohl die Anzahl der Cluster als auch die Zuordnung der Elemente zu den Clustern werden bestimmt. Somit gehören Clusteranalysen zur unüberwachten Klassifikation. Bei der unüberwachten Klassifikation (*unsupervised classification*) wird eine gegebene Anzahl von Objekten in Gruppen eingeteilt, wodurch Klassen entstehen. Im Gegensatz dazu ordnen überwachte Klassifikationen Objekte in vorgegebene Klassen ein.

Das Klassifikationsproblem besteht bei der Kategorisierung einer Menge von Objekten darin, die Objekte einer angemessen kleinen Anzahl von Clustern zuzuordnen, so daß alle Elemente einer Untergruppe ausreichend gleich sind, um ihre individuellen Unterschiede ignorieren zu können. Im allgemeinen ist Klassifikation ein Optimierungsproblem der Vereinfachung, welches zwischen dem Verlust von Genauigkeit und dem Gewinn an Einheitlichkeit ausgleichen muß. Beides ist die Folge der Ersetzung von Individuen durch typische, durchschnittliche oder charakteristische Repräsentanten. Vgl. Eckes und Rossbach (1980, S. 9) und Gordon (1999, S. 3 f.).

[1] etwa der Übergang von Iris virginica zu Iris versicolor

## 3.4 Kriterium und Verfahren

Ein Cluster*kriterium* definiert axiomatische, numerische Bedingungen, deren Optimierung die ideal Zerlegung der Datenmenge liefert oder das Aussehen günstigster Cluster beschreibt. Ein *Verfahren* ist ein Algorithmus, durch den ein solches Optimum erreicht werden kann. Ein *Kriterium* kann durch *Verfahren* verschiedener Klassen implementiert werden. Diese Unterscheidung wird in der Literatur für gewöhnlich nicht getroffen, wodurch verschiedene Autoren Verfahren in unterschiedliche Klassen einordnen. GORDON (1981, S. 41–46) gibt sowohl einen agglomerativen Algorithmus als auch einen iterativen Austauschalgorithmus für ein Verfahren an, dessen Kriterium die Minimierung der Summe der quadrierten Abstände ist. Es wird gemeinhin aber als Partitionierungsverfahren gehandelt. Single-Linkage zählt zu den hierarchisch-agglomerativen Verfahren (SNEATH, 1957), wurde aber ursprünglich von FLOREK et al. (1951) als ein nicht-hierarchisches und nicht-iteratives Verfahren definiert.

Kein Clusterkriterium und kein Ähnlichkeitsmaß ist universell anwendbar. Verschiedene Kriterien können dieselben Daten auf unterschiedliche Weise beschreiben. Daten und Clusterkriterium müssen aufeinander abgestimmt sein. Alle Kriterien zwingen den Daten eine Struktur auf, weil die Bedingungen über Form und Größe der Gruppen implizit in der Definition des Kriteriums enthalten sind. Die Verfahren zerlegen die Daten entsprechend des Kriteriums und verzerren sie zu dem durch das Kriterium bestimmte formale Optimum hin. Die mittels eines Verfahrens gefundenen Gruppen können durchaus optimal im Sinne eines vorgegebenen Kriteriums sein, müssen deswegen aber durchaus noch nicht der den Daten eigenen Struktur entsprechen. Ein ungeeignetes Kriterium kann Artefakte erzeugen und die inhärente Struktur bis zur Unkenntlichkeit verzerren.

Im allgemeinen ist es nicht möglich, im voraus zu bestimmen, welche Kombinationen von Objektattributen, Ähnlichkeitsmaßen und Clusterkriterien zu interessanten und informativen Klassifikationen führen. Deshalb sind Vorwissen und Intuition notwendig, um geeignete Werte für freie Parameter angeben zu können. Auf diese Weise ist die Auswahl eines bestimmten Kriteriums zumindest teilweise subjektiv und stets offen für Fragen. Deswegen darf das Ergebnis eines einzelnen Clusterprozesses nie unkritisch akzeptiert werden. Weitere Analysen sind notwendig, um das Ergebnis zu stützen oder zu verwerfen.

Ist die Struktur der Daten bekannt, können geeignete Kriterien gewählt werden. Im allgemeinen trifft das jedoch nicht zu, da die Untersuchung die

Strukturen erst bestimmen soll. Die Strukturen werden meist schrittweise ermittelt. Bei diesem Herantasten besteht die Gefahr, daß Parameter so lange angepaßt werden, bis das gewünschte Ergebnis erzielt wird, anstatt zuvor ein Kriterium zu definieren, das die berechneten Ergebnisse als solche zu interpretieren erlaubt.

## 3.5 Arten

Es existieren zahlreiche unterschiedliche Clusteranalyseverfahren, die in eine oder mehrere der im folgenden aufgezählten Kategorien eingeordnet werden. Eine *partitionierende* Klassifikation besteht aus genau einer *Partition* der Datenmenge. Eine *hierarchische* Klassifikation ist eine *Familie von Partitionen.* Hierarchische Verfahren werden eingeteilt in *agglomerative* Verfahren, die schrittweise durch Zusammenfassen von Clustern aus der feinsten die gröbste Partition erzeugen, und in *divisive* Verfahren, die umgekehrt sukzessive die gröbste in die feinste Partition überführen. Somit ist eine hierarchische Klassifikation eine Folge von ineinander eingebetteten partitionierenden Klassifikationen. *Serielle* (sequentielle) Verfahren verarbeiten eine Distanz zwischen Objekten nach der anderen, während *simultane* (globale) Verfahren stets auf der ganzen Datenmenge operieren. Eine *deterministische* (disjunkte, scharfe, *hard, crisp*, exklusive) Klassifikation ist eine Partition der Menge der Objekte. Jedes Objekt wird genau einem Cluster zugeordnet. Verfahren werden *überlappend* (*clumping*) genannt, wenn ein Objekt in mehr als einem Cluster liegt. Dazu gehören die *probabilistischen* und *possibilistischen* Clusteranalyseverfahren. Bei den *probabilistischen* (stochastischen) wird zu jedem Datum die Wahrscheinlichkeit angegeben, mit der es zu einem durch eine Wahrscheinlichkeitsverteilung beschriebenen Cluster gehört. Bei den *possibilistischen* (unscharfen, *fuzzy*) Verfahren wird jedem Objekt ein Zugehörigkeitswert (Möglichkeitswert) zugeordnet, der den Grad der Zugehörigkeit zu einem Cluster angibt. Bei beiden Verfahren ist jedes Objekt mit jedem Cluster assoziiert. *Inkrementelle* Algorithmen bestimmen Klassifikationen durch Hinzufügen von Objekten zu bereits bestehenden Klassifikationen. Sie werden bei sehr großen Datenmengen benutzt. *Unvollständige* Verfahren untersuchen nur einen Teil der Merkmale. Das ist bei graphischen Verfahren der Fall, die sich meist auf drei Dimensionen beschränken, weil die Cluster aufgrund einer Visualisierung der Daten bestimmt

werden. Ein *polythetischer* Algorithmus berücksichtigt alle Merkmale gleichzeitig, ein *monothetischer* hingegen verarbeitet sie nacheinander.

## 3.6 Anforderungen

Von einer Klassifikation wird Stabilität und Objektivität gefordert. Die Verfahren sollten objektiv in dem Sinne sein, daß die Analyse derselben Datenmenge durch dieselbe Folge numerischer Methoden stets dieselbe Klassifikation liefert (Wiederholbarkeit). Unter Stabilität ist zu verstehen, daß die bereits klassifizierten Elemente nicht durch Hinzufügen weniger unterschiedlicher neuer Daten aus ihren Gruppen fallen. Eine Klassifikation darf nur geringfügig beeinflußt werden durch

(1) kleine Fehler bei der Erfassung der die Objekte beschreibenden Merkmale,

(2) das Hinzufügen weniger neuer Objekte zur Datenmenge,

(3) die Hinzunahme weniger neuer Merkmale zu jedem Objekt der Datenmenge,

(4) die Reihenfolge, in der die Objekte verarbeitet werden.

Die zweite Anforderung beinhaltet die Annahme, daß die zu untersuchenden Objekte eine repräsentative Auswahl einer größeren Datensammlung sind. Die dritte Bedingung fordert eine analoge Voraussetzung für die Merkmale. Vgl. GORDON (1981, S. 9).

## 3.7 Schritte

Eine Clusteranalyse durchläuft folgende Schritte, die – falls notwendig – ergänzt werden, um andere Merkmale oder ein anderes Ähnlichkeitsmaß zu wählen oder den Fokus auf ein bestimmtes Teilproblem zu lenken:

1. Auswahl der Objekte und deren Merkmale
2. Festlegung einer angemessenen Ähnlichkeitsfunktion
3. Bestimmung einer oder mehrerer geeigneter Methoden zur Gruppierung
4. Technische Durchführung
5. Validierung der Ergebnisse

6. Interpretation der Ergebnisse

Im letzten und sehr wichtigen Schritt einer Clusteranalyse wird das Ergebnis der Klassifikation bewertet. Es gilt zu erkennen, ob die Cluster natürlich oder künstlich sind, ob bessere Lösungen gefundenen werden können und ob eine überzeugende Interpretation gegeben werden kann. Die Interpretation des Ergebnisses einer Clusteranalyse ist wesentlich bestimmt durch Aufgabenstellung, Problemverständnis und persönliche Intuition. Häufig gilt eine Clusteranalyse dann als erfolgreich, wenn den erzeugten Clustern eine Bedeutung zukommt, die interpretiert werden kann. Es gibt keine optimale Strategie, weder zur Auswahl und Anwendung eines Clusterverfahrens noch zur Bewertung der Lösung.

## 3.8 Modellierung

### 3.8.1 Objekte

Clusteranalysen[2] gruppieren eine feste Anzahl $n$ verschiedener Objekte. Die Objekte werden beliebig, aber fest durch die ersten $n$ Elemente der Menge der natürlichen Zahlen $\mathbb{N}$ indiziert. Es seien $x_1, x_2, \ldots, x_n$ diese Objekte und $X := \{x_1, x_2, \ldots, x_n\}$ die aus ihnen gebildete Menge. Jedes Objekt wird durch mehrere Merkmale (Komponenten, *features, patterns, characteristics*) beschrieben. Die Merkmale können Eigenschaften in verschiedenen Skalen und mit unterschiedlichen Maßeinheiten quantifizieren (Everitt, 1993, S. 37–50; Gordon, 1981, S. 15–30; Steinhausen und Langer, 1977, S. 53–66). Die Merkmale sind voneinander unabhängig. Ihre Anzahl sei $p$. Auch die Merkmale werden beliebig, aber fest durch die ersten $p$ natürlichen Zahlen indiziert. Für die Fragestellung dieser Arbeit genügt es, Objekte mit reellen Merkmalsausprägungen zu betrachten. Im folgenden seien alle Objekte deswegen $p$-Vektoren aus $\mathbb{R}^p$.

[2] Für die Implementation von Clusteralgorithmen und die konkrete Durchführung von Clusterprozessen ist eine mathematische Modellierung, auch wenn sie formal sehr aufwendig erscheinen mag, unerläßlich. Dazu müssen die auf das System einflußnehmenden Größen quantifiziert werden. In der Literatur sind aber meist nur grundlegende Ideen zu finden, die für eine mathematische Modellierung nicht hinreichend präzisiert sind. Insbesondere die für die Clusteranalyse so bedeutenden Indizes erfreuen sich nachgerade durchgängiger Geringschätzung. In welch bemerkenswerter Weise die Beschreibungen der Verfahren unzulänglich sind, zeigt sich erst bei dem Unterfangen ihrer Implementation. Nur der Vergleich vieler Quellen hat es Verfasser vielfach ermöglicht, vollständige und richtige Formeln und Algorithmen zu erarbeiten. Der Einfachheit halber soll aber auch hier auf den mengentheoretischen Überbau verzichtet werden (siehe etwa Bandemer und Nähter, 1992; Bezdek, 1987, 1998; Höppner et al., 1997).

Damit gilt $X \subseteq \mathbb{R}^p$. Das $k$-te Objekt, $k = 1, 2, \ldots, n$, ist der Vektor $x_k := (x_{k1}, x_{k2}, \ldots, x_{kp}) \in \mathbb{R}^p$.

### 3.8.2 Distanzen

Die Gruppierung der Elemente durch eine Clustermethode beruht auf den paarweisen Abständen der Objekte. Die Abstände (Distanzen) geben an, wie unähnlich Objekte sind. Je kleiner der Abstand ist, desto ähnlicher sind sie. Werden anstelle von Distanzen Ähnlichkeiten verwendet, gelten alle Aussagen analog, denn große Ähnlichkeiten entsprechen kurzen Distanzen. Distanzen und Ähnlichkeiten lassen sich unter bestimmten Voraussetzungen ineinander überführen (Abschnitt 2.17.4).

Eine Metrik $d$ (Abschnitt 2.8) ordnet jedem der $n^2$ Paaren von Objekten $(x_k, x_l)$, $k, l = 1, 2, \ldots n$, genau eine nichtnegative reelle Zahl $d_{kl} := d(x_k, x_l)$ zu. Die Distanzen werden in der Distanzmatrix

$$D := (d_{kl})_{1 \le k,l \le n} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix} \in [0, \infty)^{n \times n}$$

zusammengefaßt. Die Komponente mit dem Index $(k, l)$ ist der Abstand der Objekte $k$ und $l$. Die Distanzmatrix ist wegen (M3) symmetrisch. Wegen (M1) sind alle Elemente auf der Hauptdiagonalen 0.

### 3.8.3 Partitionen

#### 3.8.3.1 Menge von Mengen

Eine Menge $\mathcal{U} := \{U_1, U_2, \ldots, U_c\}$ von nichtleeren Teilmengen $U_i$ von $X$ heißt Partition (Zerlegung) von $X$ in die $c$ Cluster $U_1, U_2, \ldots, U_c$, wenn die Vereinigung der Cluster die Datenmenge ergibt,

$$X = \bigcup_{i=1}^{c} U_i,$$

und die Cluster paarweise disjunkt sind,

$$\underset{1 \le i < j \le c}{\forall} U_i \cap U_j = \emptyset.$$

Eine Zerlegung in $c$ solche Teilmengen heißt $c$-Partition (scharfe $c$-Partition, *hard c-partition, crisp c-partition*).

#### 3.8.3.2 Matrix

Eine $c$-Partition kann auch als eine Matrix von Zugehörigkeitswerten dargestellt werden. Die Zeilen entsprechen den Clustern. Eine $c \times n$-Matrix

$$U := (u_{ik})_{1 \leq i \leq c,\, 1 \leq k \leq n}$$

beschreibt eine Zerlegung der $n$ Objekte in $c$ Cluster, wenn die folgenden Bedingungen gelten:

$$\underset{1 \leq i \leq c}{\forall} \; \underset{1 \leq k \leq n}{\forall} \; u_{ik} \in \{0, 1\}\,,$$

$$\underset{1 \leq k \leq n}{\forall} \; \sum_{i=1}^{c} u_{ik} = 1, \tag{3.1}$$

$$\underset{1 \leq i \leq c}{\forall} \; \sum_{k=1}^{n} u_{ik} \geq 1. \tag{3.2}$$

Jeder Cluster $i$ wird durch eine boolsche Zugehörigkeitsfunktion

$$u_i : X \to \{0, 1\}\,,\; x \mapsto \begin{cases} 1, & \text{falls } x \in \text{Cluster } i \\ 0, & \text{sonst} \end{cases}$$

unter Einhaltung obiger Bedingungen mit $u_{ik} := u_i(x_k)$ beschrieben. Die Zugehörigkeitsfunktionen entsprechen den Mengen der Partition. Die Komponente $u_{ik}$ ist 1, wenn das $k$-te Objekte zum $i$-ten Cluster gehört und 0, wenn es nicht Element des Clusters ist. Die Spaltensumme in (3.1) zählt die Cluster, in denen das Objekt $k$ vorkommt. So ist nach (3.1) jedes Objekt Element nur eines Clusters. Die Cluster sind somit paarweise disjunkt. (3.1) besagt auch, daß jedes Objekt einem Cluster zugewiesen wird, das heißt, die Partition ist erschöpfend. Die Summe in (3.2) zählt die Elemente des Clusters $i$. Wegen (3.2) enthält somit jeder Cluster mindestens ein, aber nie alle Objekte. Jedes Objekt ist also entweder Element oder nicht Element eines Clusters und liegt in genau einem Cluster.

#### 3.8.3.3 Tupel

Eine weitere Möglichkeit, eine Partition zu repräsentieren, ist ein Tupel $(u_1, u_2, \ldots, u_n)$. Die $k$-te Komponente ist der Index $u_k$, $1 \leq u_k \leq c$, des Clusters, in dem das $k$-te Objekt liegt.

### 3.8.4 Hierarchien

Für zwei Partitionen $\mathcal{U}, \mathcal{V} \subset \mathrm{Pot}(X)$ der Menge $X$ heißt $\mathcal{U}$ feiner als $\mathcal{V}$, $\mathcal{U} \subset \mathcal{V}$, wenn

$$\underset{V \in \mathcal{V}}{\forall} \; \underset{U \in \mathcal{U}}{\exists} \; U \subset V.$$

Die feinere Zerlegung enthält die kleineren Mengen. $\mathcal{V}$ heißt gröber als $\mathcal{U}$, wenn $\mathcal{U}$ feiner als $\mathcal{V}$ ist. Die feinste Partition von $X$ enthält nur die aus den einzelnen Elementen der Menge $X$ gebildeten einelementigen Mengen. Die gröbste Partition von $X$ enthält nur die Menge $X$ selbst.

Eine Hierarchie ist eine Familie $(\mathcal{U}_t)_{t=0}^{r}$ von Partitionen von $X$, wenn

$$\mathcal{U}_0 = \big\{\{x\} : x \in X\big\},$$

$$\mathcal{U}_r = \{X\}$$

und

$$\underset{0 \leq t \leq r-1}{\forall} \; \mathcal{U}_t \subsetneqq \mathcal{U}_{t+1}$$

gilt. $(\mathcal{U}_t)_{t=0}^{r}$ nennt man auch $(r+1)$-stufige Partition. $\mathcal{U}_t$ ist die Zerlegung der $t$-ten Stufe. $\mathcal{U}_0$ ist die feinste und $\mathcal{U}_r$ die gröbste Partition. Vgl. SCHADER (1981, S. 11).

## 3.9 Hierarchisch-agglomerative Verfahren

Eine hierarchische Clusterung ist eine Folge von Partitionen, in der jede Partition in die nächste eingebettet ist. Hierarchische Verfahren zeigen die Struktur der Daten auf mehreren verschiedenen Stufen. Sie lassen erkennen, wie die Mengen einer Partition schrittweise auseinander entstehen.

### 3.9.1 Algorithmus

Agglomerative Verfahren erzeugen eine endliche Folge von Partitionen, indem sie sukzessive die beiden ähnlichsten Cluster zu einem größeren zusammenfassen (Agglomeration, Fusion, Verschmelzung, Vereinigung, *join*, *amalgamation*). Sie beginnen mit der feinsten Partition, in der jedes Objekt ein einelementiges Cluster bildet, und enden mit der gröbsten, die aus einer alle Objekte umfassenden Menge besteht. Einmal demselben Cluster zugewiesene Objekte werden in keinem Schritt mehr getrennt.

Die hier genannten Verfahren operieren nur auf der Distanzmatrix, nicht auf den Objekten selbst. Im folgenden allgemeinen Algorithmus zur hierarchisch-agglomerativen Clusterung (*matrix updating algorithm*) von $n$ Objekten werden in jedem Schritt die zu den zu vereinigenden Gruppen gehörenden zwei Zeilen und zwei Spalten der Distanzmatrix durch jeweils eine für die neue vereinigte Gruppe ersetzt, wodurch die Matrix schrumpft.

(HA1) Beginne mit der feinsten Partition.

(HA2) Suche in der Distanzmatrix den kleinsten Wert. Gibt es mehrere, entscheidet ein willkürlich zu wählendes, dann aber festgelegtes Kriterium, etwa die zuerst gefundene kleinste Distanz. Die beiden zugehörigen Gruppen seien $i$ und $j$.

(HA3) Fusioniere die beiden Gruppen $i$ und $j$ zu der neuen Gruppe $(ij)$. Dadurch verringert sich die Anzahl der Gruppen um 1.

(HA4) Verkleinere die Distanzmatrix durch Ersetzen der Zeilen und Spalten $i$ und $j$ durch eine neue Zeile und eine neue Spalte für die fusionierte Gruppe $(ij)$. Berechne die Abstände zwischen der neuen Gruppe $(ij)$ und allen übrigen Gruppen $k$ neu.

(HA5) Beende nach $n-1$ Schritten, wenn die gröbste Partition vorliegt, ansonsten fahre beim zweiten Schritt (HA2) mit der geänderten Distanzmatrix fort.

Für diesen Algorithmus ist die Partition am besten als Tupel darzustellen. Der Index einer Komponente ist der Index des zugehörigen Objektes und die Komponente der Index des Clusters dieses Objektes. Anfänglich bildet jedes Objekt ein einelementiges Cluster. Index und Komponente sind gleich. Das initiale Tupel $(1, 2, \ldots, n)$ beschreibt die feinste Partition. In jedem Schritt werden zwei Cluster vereinigt. Dazu werden alle Vorkommen des einen Clusterindex im Tupel

durch den anderen ersetzt. Letztlich sind alle Komponenten des Tupels gleich. Dieses Tupel repräsentiert die gröbste Partition.

Die Hierarchie einer hierarchisch-agglomerativen Clusterung wird nicht als Tupel von Partitionen, sondern kürzer als $(n-1)$-Tupel von Tripeln dargestellt. Das $t$-te Tripel $(\mathit{from}_t, \mathit{to}_t, l_t)$ beschreibt die $t$-te von insgesamt $n-1$ Agglomerationen. Im $t$-ten Schritt wird der Cluster $\mathit{from}_t$ dem Cluster $\mathit{to}_t$ hinzugefügt. Alle Komponenten im Partitionstupel mit dem Wert $\mathit{from}_t$ werden durch $\mathit{to}_t$ ersetzt. Der Index $\mathit{from}_t$ tritt von da an nicht mehr auf. Die dritte Komponente $l_t$ heißt Agglomerationsniveau. Sie ist die Distanz zwischen den vereinigten Gruppen.

### 3.9.2 Rekursionsformel

Die agglomerativen Methoden unterscheiden sich durch verschiedene Distanzmaße, die den Abstand zwischen zwei Gruppen von Objekten bestimmen. Das Abstandsmaß beeinflußt wesentlich die Form der resultierenden Cluster. Die Abstände der einelementigen Cluster der feinsten Partition sind die Abstände der Objekte selbst. Alle weiteren Clusterabstände werden aus diesen abgeleitet. Lance und Williams (1967, S. 376) bestimmen die Distanz zwischen einer Gruppe $k$ und der aus der Fusion der Gruppen $i$ und $j$ entstandenen Gruppe $(ij)$ rekursiv durch

$$d_{k(ij)} := \alpha_i\, d_{ki} + \alpha_j\, d_{kj} + \beta\, d_{ij} + \gamma\, |d_{ki} - d_{kj}|,$$

wobei $\alpha_i$, $\alpha_j$, $\beta$ und $\gamma$ Parameter sind und $d_{ij}$ die Distanz zwischen den Gruppen $i$ und $j$ ist. Der Abstand der durch Fusion entstandenen Gruppe $(ij)$ zu einer anderen Gruppe $k$ ist die Summe aus dem gewichteten Abstand der Gruppen $i$ und $k$, dem der Gruppen $j$ und $k$, dem der beiden agglomerierten Gruppen $i$ und $j$ sowie dem Unterschied der Abstände von $i$ zu $k$ und $j$ zu $k$. Die kleinste aller Distanzen gibt an, welche Cluster auf der nächsten Stufe zu vereinigen sind. Das Distanzmaß ist somit eine lokale Zielfunktion. Für Ähnlichkeiten ist das lokale Kriterium zu maximieren. Der Aufbauprozeß der Hierarchie wird durch dieses Gütekriterium optimiert.

### 3.9.3 Eigenschaften

Ein hierarchisch-agglomeratives Verfahren heißt *monoton*, wenn die Stufe der nächsten Agglomeration immer größer oder gleich der jetzigen ist. Bei der Vereinigung der Cluster $i$ und $j$ zum Cluster $(ij)$ fordert die Monotonie also $d_{k(ij)} \geq d_{ij}$ für alle Cluster $k$. Keine Distanz in der neu berechneten Matrix ist kleiner als die kleinste in der vorigen. Die Ultrametrikungleichung (2.8) ist erfüllt. Ziel einer monotonen hierarchischen Clustermethode ist es, den Daten algorithmisch die Ultrametrikungleichung aufzuerlegen. Vereinigen sich zwei Cluster auf einem niedrigeren Niveau als die vorige Fusion, spricht man von einer Inversion (*reversal, crossover*). Das Verfahren ist dann nicht mehr monoton. Monotone Distanzmaße ergeben sich etwa für $\gamma = 0$ und $\alpha_i + \alpha_j + \beta \geq 1$ (Lance und Williams, 1967, S. 374). Ein Verfahren wird *kontrahierend* genannt, wenn es dazu neigt, noch entferntere Elemente einem Cluster zuzuordnen. Es heißt *dilatierend*, wenn es dazu tendiert, die Elemente in sehr kleine Gruppen zusammenzufassen. Es wird *konservativ* (raumerhaltend) genannt, wenn es keine dieser beiden Eigenschaften hat.

### 3.9.4 Kophänetische Distanzen

Hierarchische Verfahren bilden die Distanzmatrix auf die Matrix der sogenannten kophänetischen Distanzen ab. Die kophänetische[3] Distanz (*cophenetic proximity*) $c_{ij}$ zweier Objekte $i$ und $j$ gibt das Agglomerationsniveau an, an dem die beiden Objekte erstmalig in denselben Cluster fallen. Die $n \times n$-Matrix $C$ dieser Werte $c_{ij}$ heißt kophänetische Korrelationsmatrix (*cophenetic matrix*). Sie ist symmetrisch. Eine hierarchische Klassifikation ist somit eine Transformation, die die beobachteten Distanzen $d_{ij}$ auf die kophänetischen Distanzen $c_{ij}$ abbildet. Die resultierende Hierarchie kann als Baumgraph (Dendrogramm) repräsentiert werden. Ist das Verfahren monoton, so gilt für je drei Komponenten aus der kophänetischen Matrix die Ultrametrikungleichung. Zwei der drei Komponenten $c_{ij}$, $c_{ik}$ und $c_{kj}$ sind dann gleich, die dritte kleiner oder gleich (Abschnitt 2.8). Die kophänetischen Distanzen sind dann ultrametrische Distanzen. Das Dendrogramm einer monotonen Hierarchie ist ein ultrametrischer Baum.

[3] von Phänotyp (Ähnlichkeit), den körperlichen Merkmalen eines Organismusses, im Gegensatz zum Genotyp (Abstammung), der genetischen Ausstattung

Eine der $n$ Partitionen einer Hierarchie erhält man durch einen Schnitt im Dendrogramm oder durch Wahl einer Lösung aus der Familie der Zerlegungen. Die Wahl einer Partition aus einer Hierarchie entspricht der Bestimmung der Clusteranzahl. Vgl. JAIN und DUBES (1988, S. 68 ff.) und MIRKIN (1996, S. 332 f.).

### 3.9.5 Verfahren

Die Distanzmaße der folgenden hierarchisch-agglomerativen Clusterverfahren werden durch geeignete Wahl der Parameter $\alpha_i$, $\alpha_j$, $\beta$ und $\gamma$ bestimmt. Dadurch können all diese Verfahren durch einen einzigen Algorithmus beschrieben werden. Vgl. EVERITT (1993, S. 66 f.), GORDON (1981, S. 46), JAIN und DUBES (1988, S. 79 f.) und STEINHAUSEN und LANGER (1977, S. 75 ff.).

#### 3.9.5.1 Minimaler Spannbaum

Ein minimaler Spannbaum ist kein Ergebnis einer hierarchisch-agglomerativen Clusterung. Er ist aber mit dem im nächsten Abschnitt vorgestellten Single-Linkage eng verbunden, da aus ihm das Ergebnis einer Single-Linkage-Clusterung abgeleitet werden kann (GOWER und ROSS, 1969).

Ein Graph ist ein Tripel $G := (V, E, f)$ aus einer Menge von Knoten $V$ (*vertices, nodes*), einer Mengen von Kanten $E$ (*edges*) und einer Abbildung $f : E \to V \times V$, die jeder Kante die durch diese verbundenen zwei Knoten zuordnet. Ein Graph mit $n$ Knoten hat höchstens $n(n-1)/2$ Kanten. Ein Graph heißt zusammenhängend (*connected*), wenn es von jedem Knoten einen Pfad zu jedem anderen gibt. Eine Schleife (*cycle*) ist ein Pfad, dessen Start- und Endknoten identisch sind. Ein Baum ist ein zusammenhängender Graph ohne Schleifen. Ein Spannbaum ist ein Baum, der alle Knoten des Graphen enthält. Das Gewicht eines Graphen ist die Summe der Gewichte aller Kanten. Der minimale Spannbaum (*minimum/minimal spanning tree*) eines Graphen ist der Spannbaum mit dem geringsten Gewicht. Im allgemeinen ist ein minimaler Spannbaum nicht eindeutig bestimmt. Es kann verschiedene Bäume mit minimalem Gewicht geben. Ein minimaler Spannbaum für $n$ Knoten hat $n - 1$ Kanten. Die Verteilung der Längen der Kanten eines minimalen Spannbaums läßt die Kompaktheit der Daten erkennen (BACKER, 1995, S. 69).

Ein isolierter Punkt ist ein Knoten, der nicht durch eine Kante mit einem anderen Knoten verbunden ist. Der Algorithmus zur Erzeugung eines minimalen

Spannbaums von Prim (1957) fügt sukzessive isolierte Punkte und isolierte Fragmente zusammen, bis alle Punkte miteinander verbunden sind. Ihm liegen folgende zwei Prinzipien zugrunde.

- Jeder isolierte Punkt kann mit seinem nächsten Nachbarn verbunden werden.
- Jedes isolierte Fragment kann mit einem nächsten Nachbarn verbunden werden.

Jede Anwendung einer dieser beiden Prinzipien vermindert die Anzahl isolierter Punkte oder Fragmente um 1. Die $n$ Punkte werden in $n-1$ Schritten zu einem minimalen Spannbaum verbunden.

(MST1) Wähle einen beliebigen Punkt $x_i$, $1 \leq i \leq n$. Dieser bildet das einelementige initiale Fragment $F$. Alle Punkte, die nicht zum Fragment gehören, heißen isoliert. Zu jedem isolierten Punkt gibt es einen nächsten Nachbarn im Fragment. Fasse den Index $j$ eines jeden Punktes $x_j$, $1 \leq j \leq n$, $j \neq i$, dessen Abstand $d_{ij}$ zum Startpunkt $x_i$ und den Index des nächsten Nachbarn $k$ im Fragment zu einem Tripel $(j, d_{ij}, k)$ und alle Tripel zu eine Liste $L$ zusammen. Initial sind alle $k$ gleich $i$.

(MST2) Bestimme den kleinsten Abstandswert in der Liste. Der Index des zugehörigen Punktes sei $l$.

(MST3) Füge diesen und dessen Kante mit dem zugehörigen nächsten Nachbarn im Fragment dem Fragment hinzu. Lösche den zugehörigen Eintrag aus der Liste.

(MST4) Berechne zu dem neu in das Fragment aufgenommenen Punkt $x_l$ die Abstände zu allen verbleibenden isolierten Punkten. Ist ein Abstand kleiner als der entsprechende in der Liste, ersetze diesen durch den kleineren und den Index des zugehörigen nächsten Nachbarn im Fragment durch den des neuen Punktes im Fragment.

(MST5) Springe zu Schritt (MST2), wenn die Liste noch nicht leer ist. Ist sie leer, bildet $F$ einen minimalen Spannbaum.

#### 3.9.5.2 Single-Linkage

Beim Single-Linkage ist die Distanz zwischen zwei Gruppen definiert als der kleinste aller Abstände von je einem Element aus der einen und einem aus der anderen Gruppe. Dieses Verfahren ergibt sich für $\alpha_i := \alpha_j := 1/2$, $\beta := 0$ und $\gamma := -1/2$. Für $d_{ki} < d_{kj}$ ist

$$d_{k(ij)} = \frac{1}{2}d_{ki} + \frac{1}{2}d_{kj} - \frac{1}{2}\left(d_{kj} - d_{ki}\right) = d_{ki}.$$

Für $d_{kj} < d_{ki}$ ergibt sich analog $d_{k(ij)} = d_{kj}$. Der Abstand der Gruppe $k$ und der aus der Fusion der Gruppen $i$ und $j$ entstandenen Gruppe $(ij)$ ist somit

$$d_{k(ij)} = \min(d_{ki}, d_{kj}).$$

Für die Fusion zweier Cluster genügt also eine einzige Verbindung (*single link*) zweier Objekte. Single-Linkage verbindet alle Punkte, die unterhalb einer Distanzstufe (*threshold, level*) liegen. Auf einer Distanzstufe $h$ sind demnach all diejenigen Elemente zu einer Gruppe zusammengefaßt, die wenigstens zu einem der anderen Elemente der Gruppe eine Distanz kleiner oder gleich $h$ haben. Single-Linkage faßt auf relativ niedriger Stufe Objekte durch eine Reihe von dazwischenliegenden Knoten (*chaining points, intermediates*) zusammen. Diese Eigenschaft wird als Kettenbildung (*chaining*) bezeichnet. Die Methode vermag es nicht, Cluster zu unterscheiden, wenn einige wenige Objekte zwischen ihnen liegen. Aneinanderreihungen schlecht getrennter Cluster werden deswegen nicht aufgedeckt. Sind dagegen die Cluster durch leere Grenzzonen deutlich voneinander getrennt, kann Single-Linkage Gebilde jeder Form und Größe erkennen. Es kommt mehr auf den Zusammenhalt der Elemente als auf die Ähnlichkeiten der einzelnen Paare an. Single-Linkage ist angemessen für optimal verbundene Cluster (*optimally connected clusters*). Ein Nachteil der Verkettung ist, daß Objekte eines Clusters weiter voneinander entfernt sein können als Objekte aus verschiedenen Gruppen. Die Lösungen sind unter monotonen Transformationen der Distanzmatrix invariant. Gleiche Rangordnungen der Distanzwerte führen also zu identischen hierarchischen Partitionen. Single-Linkage setzt damit lediglich ordinalskalierte Distanzen voraus. Single-Linkage ist kontrahierend und monoton.

Alle Informationen für eine Single-Linkage-Clusterung einer Menge von Punkten sind in deren minimalen Spannbaum enthalten. Eine Single-Linkage-

Hierarchie kann deshalb aus einem minimalen Spannbaum abgeleitet werden. Umgekehrt jedoch ist dies nicht möglich. Single-Linkage-Clusterungen zum Level $h$ ergeben sich durch Löschen aller Kanten, die länger als $h$ sind, aus dem minimalen Spannbaum. Vgl. Florek et al. (1951), Sneath (1957) und auch Backer (1995, S. 67), Everitt (1993, S. 57–60), Gordon (1999, S. 53, 80), Steinhausen und Langer (1977, S. 76 f.).

#### 3.9.5.3 Complete-Linkage

Complete-Linkage (*clique, furthest neighbor*) ist das Gegenteil von Single-Linkage in dem Sinne, daß die Distanz zwischen zwei Gruppen definiert ist als die der am weitesten auseinanderliegenden Individuen aus unterschiedlichen Gruppen. Auf einer Distanzstufe $h$ werden all jene Elemente zu einer Gruppe vereinigt, die höchstens den Abstand $h$ haben. Dieses Verfahren tendiert im Gegensatz zu Single-Linkage zur Bildung kleiner Gruppen, ist also dilatierend. Es ist invariant unter monotonen Transformationen der Distanzmatrix. Complete-Linkage ist monoton. Mit den Parametern $\alpha_i := \alpha_j := 1/2$, $\beta := 0$ und $\gamma := 1/2$ ist

$$d_{k(ij)} = \max(d_{ki}, d_{kj})$$

der Abstand der Gruppen $k$ und $(ij)$. In der Tat ist für $d_{ki} < d_{kj}$

$$d_{k(ij)} = \frac{1}{2}d_{ki} + \frac{1}{2}d_{kj} + \frac{1}{2}\left(d_{kj} - d_{ki}\right) = d_{kj}.$$

Analog ergibt sich $d_{k(ij)} = d_{kj}$ für $d_{kj} < d_{ki}$. Vgl. Steinhausen und Langer (1977, S. 78).

#### 3.9.5.4 Average-Linkage

Beim Average-Linkage ist die Distanz zweier Gruppen definiert als der Durchschnitt aller Distanzen zwischen je einem Objekt aus dem einen und einem aus dem anderen Cluster. Beim Agglomerationsschritt fusionieren die beiden Klassen, deren mittlere Distanz am kleinsten ist. Es sei $n_i$ die Größe des $i$-ten Clusters. Dann ist $\alpha_i := n_i/(n_i + n_j)$ der Anteil der $i$-ten Gruppe an der Größe der vereinigten Gruppe und $\alpha_j := n_j/(n_i + n_j)$ der von $j$. Mit $\beta := \gamma := 0$ ist

$$d_{k(ij)} = \frac{n_i d_{ki} + n_j d_{kj}}{n_i + n_j}.$$

Das Verfahren ist konservativ und monoton. Es ist zwischen den Extremen Single-Linkage und Complete-Linkage einzuordnen. Vgl. GORDON (1999, S. 79, als *group average link*) sowie STEINHAUSEN und LANGER (1977, S. 78 f., als *weighted average linkage*).

#### 3.9.5.5 Ungewichtetes Average-Linkage

Bei der ungewichteten[4] Variante des Average-Linkage gehen beide Gruppen zu gleichen Teilen in die fusionierte Gruppe ein. Objekte in kleinen Clustern wiegen deshalb schwerer als Objekte in großen Clustern. Es ist $\alpha_i := \alpha_j := 1/2$, $\beta := \gamma := 0$ und somit

$$d_{k(ij)} = \frac{d_{ki} + d_{kj}}{2}.$$

Das Verfahren ist konservativ und monoton. Vgl. GORDON (1999, S. 79, als *weighted average link*), KAUFMAN und ROUSSEEUW (1990, S. 234, als *weighted average linkage*), STEINHAUSEN und LANGER (1977, S. 78, als *average linkage*).

#### 3.9.5.6 Zentroid

Beim Zentroid-Verfahren wird jede Gruppe durch das arithmetische Mittel aus allen Elementen der Gruppe repräsentiert. Dieser Mittelpunkt (Schwerpunkt, Zentroid) ist im allgemeinen kein Element der Gruppe. Die Gruppen werden entsprechend ihrer Größe gewichtet. Die rekursive Definition durch $\alpha_i := n_i/(n_i + n_j)$, $\alpha_j := n_j/(n_i + n_j)$, $\beta := -n_i n_j/(n_i + n_j)^2$, $\gamma := 0$ und

$$d_{k(ij)} = \frac{n_i d_{ki} + n_j d_{kj}}{n_i + n_j} - \frac{n_i n_j}{(n_i + n_j)^2} d_{ij}$$

ist nur bei quadrierter euklidischer Distanz zwischen den Mittelpunkten anschaulich interpretierbar. Es werden dann jeweils die beiden Gruppen vereinigt, deren Schwerpunkte den geringsten Abstand haben. Das Verfahren ist konservativ, aber nicht monoton. Mit abnehmender Anzahl von Clustern vergrößert sich der Abstand zwischen den Gruppen nicht immer. Es können Inversionen auftreten, das heißt, die Cluster $(ij)$ und $k$ agglomerieren auf niedrigerer Stufe als $i$ und $j$ im vorangegangenen Fusionsschritt. Zur Agglomeration zweier Cluster genügt es, daß ihre Objekte im Mittel hinreichend ähnlich sind. Der Einfluß einzelner unähnlicher Objektpaare wird durch größere

[4] In der englischsprachigen Literatur heißt eine Methode „weighted“, wenn alle Gruppen gleich gewichtet sind. Methoden werden „unweighted“ genannt, wenn die Cluster entsprechend ihrer Größe gewichtet sind. In der deutschsprachigen Literatur ist dies meistens umgekehrt.

Ähnlichkeiten anderer ausgeglichen. Die Clustermittelpunkte sind im allgemeinen keine Objekte der Datenmenge. Bei zwei Clustern stark unterschiedlicher Größe liegt der Schwerpunkt des Fusionsclusters sehr viel näher am Schwerpunkt des größeren Clusters, möglicherweise sogar in diesem. Die charakterisierende Eigenschaft des kleineren Clusters geht dadurch verloren. Vgl. EVERITT (1993, S. 62), STEINHAUSEN und LANGER (1977, S. 79).

#### 3.9.5.7 Median

Beim Median-Verfahren (ungewichtetes Zentroid-Verfahren) haben alle Cluster das gleiche Gewicht. Die Distanzen zwischen den Clustern werden mit $\alpha_i := \alpha_j := 1/2$, $\beta := -1/4$ und $\gamma := 0$ durch

$$d_{k(ij)} = \frac{1}{2}\left(d_{ki} + d_{kj}\right) - \frac{1}{4}d_{ij}$$

bestimmt. Auch hier ist das Ergebnis nur bei quadrierter euklidischer Distanz anschaulich interpretierbar. Nur dann werden die Fusionscluster repräsentiert durch den Median (Zentralwert) der Verbindungslinie der zusammengeführten Cluster. Der Schwerpunkt der neuen Gruppe liegt zwischen den Schwerpunkten der beiden vereinigten Gruppen. Das Verfahren ist ebenfalls konservativ, aber nicht monoton. Vgl. GOWER (1967), EVERITT (1993, S. 65), STEINHAUSEN und LANGER (1977, S. 79).

#### 3.9.5.8 Ward

Das WARD-Verfahren vereinigt in jedem Schritt die beiden Cluster, deren Fusion den geringsten Zuwachs an Heterogenität ergibt. Die Heterogenität des $i$-ten Clusters ist definiert als die Summe der $n_i$ quadrierten euklidischen Distanzen der Elemente $x_k := (x_{k1}, x_{k2}, \ldots, x_{kp})$, $k = 1, \ldots, n_i$, des Clusters zu deren arithmetischem Mittel $\bar{x}_i := (\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{ip})$ mit $\bar{x}_{ir} = \left(\sum_{k=1}^{n_i} x_{kr}\right)/n_i$, $r = 1, \ldots, p$,

$$\begin{aligned}
\mathrm{ss}_i &:= \sum_{k=1}^{n_i} \left\|x_k - \bar{x}_i\right\|^2 \\
&= \sum_{k=1}^{n_i} \sum_{r=1}^{p} \left(x_{kr} - \bar{x}_{ir}\right)^2 \\
&= \sum_{k=1}^{n_i} \sum_{r=1}^{p} \left(x_{kr}^2 - 2x_{kr}\bar{x}_{ir} + \bar{x}_{ir}^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_k \sum_r x_{kr}^2 - 2 \sum_r \bar{x}_{ir} \sum_k x_{kr} + \sum_r \sum_k \bar{x}_{ir}^2 \\
&= \sum_k \sum_r x_{kr}^2 - 2 \sum_r \bar{x}_{ir} n_i \bar{x}_{ir} + \sum_r n_i \bar{x}_{ir}^2 \\
&= \sum_{k=1}^{n_i} \sum_{r=1}^{p} x_{kr}^2 - n_i \sum_{r=1}^{p} \bar{x}_{ir}^2 \\
&= \sum_{k=1}^{n_i} \|x_k\| - n_i \|\bar{x}_i\|.
\end{aligned}
$$

WARD hat dieses Verfahren nur für reelle Zahlen, nicht für Vektoren definiert. Ziel dieses Verfahrens ist es, die Kostenfunktion $\mathrm{ss}_i$ (*sum of squares*) zu minimieren. Mit $\alpha_i := (n_k + n_i)/(n_k + n_i + n_j)$, $\alpha_j := (n_k + n_j)/(n_k + n_i + n_j)$, $\beta := -n_k/(n_k + n_i + n_j)$ und $\gamma := 0$ berechnet

$$
d_{k(ij)} = \frac{(n_k + n_i)\, d_{ki} + (n_k + n_j)\, d_{kj} - n_k d_{ij}}{n_k + n_j + n_i}
$$

die Zuwächse der Streuung. Die Distanzmatrix enthält nicht die Distanzen zwischen Gruppen, sondern die jeweiligen Zuwächse an Heterogenität bei deren Verschmelzung. Das WARD-Verfahren tendiert zur Bildung etwa gleichgroßer Cluster. Es ist konservativ und monoton.

Die Kostenfunktion ist hier ein lokales Gütemaß, weil sie in jedem Schritt minimiert wird. Die weiter hinten beschriebenen $c$-Means-Methoden verwenden dieses Varianzkriterium als globales Gütemaß. Vgl. GORDON (1981, S. 39 ff.), STEINHAUSEN und LANGER (1977, S. 79) und natürlich auch WARD (1963).

#### 3.9.5.9 Flexible-Strategy

Durch $0 < \alpha := \alpha_i := \alpha_j$, $\beta := 1 - 2\alpha$, $\gamma := 0$ und

$$
d_{k(ij)} = \alpha\,(d_{ki} + d_{kj}) + (1 - 2\alpha)\, d_{ij}
$$

ist eine unendliche Familie von hierarchisch-agglomerativen Verfahren definiert. Durch den Parameter $\alpha$ lassen sich die Eigenschaften des Verfahrens weitgehend beeinflussen. Für $\alpha$ nahe 0 ist es stark kontrahierend und neigt wie Single-Linkage zur Kettenbildung. Für $\alpha = \frac{1}{2}$ ergibt sich die ungewichtete Variante von Average-Linkage. Für $\alpha > \frac{1}{2}$ ist es dilatierend. Auf niedrigem Distanzniveau werden dann wie bei Complete-Linkage viele relativ kompakte Cluster gebildet. STEINHAUSEN und LANGER (1977, S. 81) beschränken $\alpha$ auf das Intervall $[\frac{1}{2}, 1]$, weil dilatierende

| Verfahren | Parameter $\alpha_i$ | $\beta$ | $\gamma$ | Distanz $d_{k(ij)}$ |
|---|---|---|---|---|
| Single-Linkage | $\frac{1}{2}$ | $0$ | $-\frac{1}{2}$ | $\min(d_{ki}, d_{kj})$ |
| Complete-Linkage | $\frac{1}{2}$ | $0$ | $\frac{1}{2}$ | $\max(d_{ki}, d_{kj})$ |
| Average-Linkage | $\frac{n_i}{n_i + n_j}$ | $0$ | $0$ | $\frac{n_i d_{ki} + n_j d_{kj}}{n_i + n_j}$ |
| Ungewichtetes Average-Linkage | $\frac{1}{2}$ | $0$ | $0$ | $\frac{1}{2}(d_{ki} + d_{kj})$ |
| Zentroid | $\frac{n_i}{n_i + n_j}$ | $-\frac{n_i n_j}{(n_i + n_j)^2}$ | $0$ | $\frac{n_i d_{ki} + n_j d_{kj}}{n_i + n_j} - \frac{n_i n_j d_{ij}}{(n_i + n_j)^2}$ |
| Median | $\frac{1}{2}$ | $-\frac{1}{4}$ | $0$ | $\frac{1}{2}(d_{ki} + d_{kj}) - \frac{1}{4} d_{ij}$ |
| WARD | $\frac{n_k + n_i}{n_k + n_i + n_j}$ | $-\frac{n_k}{n_k + n_i + n_j}$ | $0$ | $\frac{(n_k + n_i) d_{ki} + (n_k + n_j) d_{kj} - n_k d_{ij}}{n_k + n_j + n_i}$ |
| Flexible-Strategy | $\alpha > 0$ | $1 - 2\alpha$ | $0$ | $\alpha (d_{ki} + d_{kj}) + (1 - 2\alpha) d_{ij}$ |

**Tabelle 3.1:** Parameter der hierarchisch-agglomerativen Verfahren

Verfahren die Trennung der Gruppen besser hervorheben, bei Werten größer als 1 aber zu kleine Gruppen gebildet werden. Geeignet seien $0{,}6 \leq \alpha \leq 0{,}7$. Das Verfahren ist monoton. Siehe auch EVERITT (1993, S. 67).

## 3.10 Verfahren zur Verbesserung einer Anfangspartition

### 3.10.1 Optimierung

Simultane Partitionierungsverfahren zerlegen die gesamte Datenmenge in die Partition, die bezüglich einer vorgegebenen globalen Kostenfunktion (Gütefunktion, Zielfunktion, Bewertungsfunktion) optimal ist. Die Kostenfunktion $J$ bestimmt für jede Zerlegung der $n$ Objekte in $c$ Gruppen einen Wert $J(n, c)$ für die Güte der Partition. Die meisten Partitionierungsverfahren bestimmen die Güte mit einer skalarbildenden Funktion aus einer der folgenden Streuungsmatrizen (*dispersion/ scatter matrix*). Das setzt voraus, daß die Objekte Elemente metrischer Räume mit annähernd normalverteilten Komponenten sind. Die Komponenten der Streuungsmatrizen sind die nicht-standardisierten Varianzen der jeweiligen Objekte. Es sei $x_k := (x_{k1}, \ldots, x_{kp}) \in \mathbb{R}^p$ der Meßwertvektor des $k$-ten Objektes. Es sei

$$\begin{aligned} v_i &:= (v_{i1}, \ldots, v_{ip}) \\ &:= \left( \frac{1}{n_i} \sum_{k=1}^{n_i} x_{i_k 1}, \ldots, \frac{1}{n_i} \sum_{k=1}^{n_i} x_{i_k p} \right) \\ &= \frac{1}{n_i} \sum_{k=1}^{n_i} (x_{i_k 1}, \ldots, x_{i_k p}) \\ &= \frac{1}{n_i} \sum_{k=1}^{n_i} x_{i_k} \in \mathbb{R}^p \end{aligned}$$

der Mittelwertvektor (*mean*) der $i$-ten Gruppe, wobei $x_{i_k}$ der $k$-te Punkt der $i$-ten Gruppe und $n_i$ die Gruppengröße ist. Weiterhin sei

$$v := \frac{1}{n} \sum_{k=1}^{n} x_k \in \mathbb{R}^p \tag{3.3}$$

der Mittelwertvektor über alle $n$ Objekte der Datenmenge (*pooled mean*). Die Streuungsmatrix der $i$-ten Gruppe ist

$$W_i := \sum_{k=1}^{n_i} (x_{i_k} - v_i)(x_{i_k} - v_i)^T \in \mathbb{R}^{p \times p}.$$

Die Matrix der Streuung innerhalb aller Gruppen (*within groups scatter matrix*, Intraclustervarianzmatrix) ist

$$W := \sum_{i=1}^{c} W_i.$$

Sie faßt die clusterspezifischen Streuungsmatrizen $W_i$ zu einer gemeinsamen Matrix zusammen. Die Matrix der Streuung zwischen den Gruppen (*between groups scatter matrix*, Interclustervarianzmatrix) ist

$$B := \sum_{i=1}^{c} n_i \, (v_i - v)(v_i - v)^T \in \mathbb{R}^{p \times p}.$$

Die Streuungsmatrix des gesamten Datensatzes (totale Streuungsmatrix, *total/ overall scatter matrix*)

$$T := B + W = \sum_{i=1}^{k} (x_i - v)(x_i - v)^T$$

setzt sich additiv zusammen aus der Intraclustervarianzmatrix $W$ und der Interclustervarianzmatrix $B$. Die Matrix $W$ beschreibt die interne Homogenität der Cluster und die Matrix $B$ deren wechselseitige Abgrenzung. Die Optimierungskriterien sind so aus den Strukturvorstellungen der internen Homogenität und externen Isolation der Cluster abgeleitet. Da die totale Streuungsmatrix eines Datensatzes stets konstant ist, entspricht die Minimierung der Intraclustervarianzmatrix der Maximierung der Interclustervarianzmatrix. Als skalarbildende Funktionen werden häufig Spur oder Determinante verwendet. Sie ersetzen eine Streuungsmatrix durch eine reelle Zahl als Gütewert. Jedes Kriterium geht dabei von ganz spezifischen Voraussetzungen bezüglich der Datenstruktur aus und garantiert nicht das Auffinden tatsächlich vorhandener Strukturen. Eine im Hinblick auf das gewählte Kriterium optimale Aufteilung der Objektmenge kann sich daher erheblich von der zweckdienlichsten Einteilung unterscheiden. Eckes und Rossbach (1980, S. 55 f.) und Everitt (1993, S. 92 f.) geben einige Beispiele an.

### 3.10.2 Vollständige Aufzählung

Das globale Optimum wird sicher gefunden, wenn alle Partitionen betrachtet werden. Rein theoretisch ist es möglich, für alle Zerlegungen der $n$ Objekte in $c$ Gruppen die Zielfunktion zu berechnen und die Partition zu wählen, für welche diese den besten Wert liefert. Jedoch ist eine vollständige Aufzählung aller möglichen Partitionen zu umfangreich, um praktisch berechnet werden zu können. Die Anzahl aller möglichen Zerlegungen von $n$ Objekten in $c$ paarweise disjunkte nichtleere Gruppen geben die STIRLINGschen Zahlen zweiter Art

$$\begin{aligned} S_2(n,c) &:= \frac{1}{c!}\sum_{i=1}^{c}(-1)^{c-i}\binom{c}{i}i^n \\ &= \frac{1}{c!}\sum_{i=0}^{c}(-1)^{i}\binom{c}{i}(c-i)^n \end{aligned}$$

an, wobei $c!$ die Fakultät von $c$ bezeichnet. Sie lassen sich rekursiv berechnen durch

$$S_2(n,c) = S_2(n-1,c-1) + c\,S_2(n-1,c)$$

mit $S_2(0,0) := 1$ und $S_2(n,0) := S_2(0,c) := 0$ für $n,c > 0$. Ferner ist $S_2(n,1) = S_2(n,n) = 1$ und $S_2(n,c) = 0$ für $n < c$. Für steigendes $n$ wächst $S_2$ exponentiell (Tabelle 3.2).

Läßt man auch noch $c$ variabel, erhält man die Anzahl aller Zerlegungen einer Menge von $n$ Elementen in paarweise disjunkte nichtleere Teilmengen mit den BELLschen Zahlen

$$B(n) := \sum_{c=1}^{n} S_2(n,c)$$

(Tabelle 3.3). Für $n \geq 0$ gilt die Rekursionsformel

$$B(n+1) = \sum_{c=0}^{n}\binom{n}{c}B(c)$$

mit $B(0) := 1$. Vgl. STEINHAUSEN und LANGER (1977, S. 17 f.).

Um diese kombinatorische Explosion zu vermeiden, versuchen Optimierungsverfahren durch eine spezifische Iterationsregel den Suchprozeß auf eine relativ geringe Anzahl guter Zerlegungen zu beschränken und die weniger guten gar nicht erst zuzulassen. Der Iterationsprozeß führt nicht notwendigerweise zu der optimalen Partition, da nur eine vergleichsweise geringe Anzahl aus der

| $n$ \ $c$ | 2 | 3 | 10 | 100 |
|---|---|---|---|---|
| 2 | 1 | 0 | 0 | 0 |
| 3 | 3 | 1 | 0 | 0 |
| 4 | 7 | 6 | 0 | 0 |
| 5 | 15 | 25 | 0 | 0 |
| 6 | 31 | 90 | 0 | 0 |
| 7 | 63 | 301 | 0 | 0 |
| 8 | 127 | 966 | 0 | 0 |
| 9 | 255 | 3025 | 0 | 0 |
| 10 | 511 | 9330 | 1 | 0 |
| 100 | $6{,}3 \cdot 10^{29}$ | $8{,}6 \cdot 10^{46}$ | $2{,}6 \cdot 10^{93}$ | 1 |
| 1000 | $5{,}4 \cdot 10^{300}$ | $2{,}2 \cdot 10^{476}$ | $2{,}8 \cdot 10^{993}$ | $1{,}1 \cdot 10^{1842}$ |

**Tabelle 3.2:** STIRLINGsche Zahlen zweiter Art

| $n$ | $B(n)$ |
|---:|---:|
| 1 | 1 |
| 2 | 2 |
| 3 | 5 |
| 4 | 15 |
| 5 | 52 |
| 6 | 203 |
| 7 | 877 |
| 8 | 4140 |
| 9 | 21147 |
| 10 | 115975 |
| 20 | $5{,}2 \cdot 10^{13}$ |
| 30 | $8{,}5 \cdot 10^{23}$ |
| 40 | $1{,}6 \cdot 10^{35}$ |
| 50 | $1{,}9 \cdot 10^{47}$ |
| 60 | $9{,}8 \cdot 10^{59}$ |
| 70 | $1{,}8 \cdot 10^{73}$ |
| 80 | $9{,}9 \cdot 10^{86}$ |
| 90 | $1{,}4 \cdot 10^{101}$ |
| 100 | $4{,}8 \cdot 10^{115}$ |
| 1000 | $3{,}0 \cdot 10^{1927}$ |

**Tabelle 3.3:** BELLsche Zahlen

Gesamtheit aller möglichen Partitionen betrachtet wird. Auch ist es möglich, daß Lösungen durch das Verschieben einzelner Objekte nicht weiter verbessert werden können. In den meisten Fällen ist dies nur eine suboptimale Lösung. Der Algorithmus ist dann in einem lokalen Optimum gefangen.

### 3.10.3 Problematik

Optimierungsverfahren sind Verfahren zur Verbesserung einer gegebenen Anfangspartition (*iterative relocation*, *alternating optimization*). Damit sind einige wesentliche Schwierigkeiten verbunden. Eine initiale Partition und damit eingeschlossen auch die Anzahl $c$ der Gruppen müssen vorher extern bestimmt werden. Die Anzahl der Gruppen ist ein konstanter Parameter. Sie wird durch den Algorithmus nicht an die Daten angepaßt. Deshalb werden für einen Datensatz stets mehrere partitionierende Clusterungen mit verschiedenen Ausgangskonfigurationen durchgeführt. Die nach jeder Clusterung erweiterte Einsicht in die Daten wird genutzt, eine vermeintlich besser geeignete Initialisierung vorzunehmen. Jedoch können unterschiedliche Ausgangskonfigurationen zu verschiedenen lokalen Optima führen. Die Clusterung sollte deshalb mit verschiedenen Initialisierungen wiederholt werden, in der Hoffnung, daß wenigstens einige Durchläufe den besten, das heißt globalen, Wert liefern. Die Güte des Ergebnisses hängt also in großem Maße von einer guten Initialisierung ab. Langsame Konvergenz und weit voneinander abweichende Gruppierungen infolge von verschiedenen initialen Partitionen lassen gewöhnlich darauf schließen, daß die Anzahl der Gruppen falsch gewählt wurde, das heißt insbesondere, daß nichts auf eine solche Clusterstruktur hinweist (Marriott, 1982, S. 418). Ebenfalls ein Indiz für eine ungeeignete vorgegebene Gruppenanzahl ist die Gleichheit zweier Mengen der berechneten Zerlegung.

### 3.10.4 Algorithmus

Der Algorithmus (*hill-climbing algorithm*) durchläuft folgende Schritte.

(AO1) Finde eine initiale Partition der Menge der Objekte.

(AO2) Berechne die Kosten, die durch Verschieben eines jeden Objektes aus seiner eigenen Gruppe in eine der anderen $c - 1$ Gruppen entsteht.

(AO3) Führe die günstigste dieser Veränderungen durch.

(AO4) Wiederhole (AO2) und (AO3) solange, bis keine weitere Verschiebung eines einzelnen Objektes mehr zu einer Verbesserung führt. Das kann dadurch bestimmt werden, daß die Differenz der Kosten von zwei aufeinanderfolgenden Schritten unterhalb einer gegebenen Schranke liegt, etwa $\varepsilon \in [0{,}0001, 0{,}01]$. Eine andere Terminierungsbedingung sieht vor, die Iteration zu stoppen, wenn $c$ mal hintereinander kein Element mehr die Gruppe gewechselt hat. Es ist auch üblich, den Algorithmus nur eine fest vorgegebene Anzahl von Iterationen durchführen zu lassen.

Das mathematische Modell der Iteration beschreibt eine konvergente Folge von Gütewerten. Von dem Prozeß, der diese Folge praktisch berechnet, ist zu hoffen, daß er nahe des theoretischen Konvergenzpunktes terminiert.

### 3.10.5 Hard-c-Means

Hard-$c$-Means (Duda und Hart, 1973) erkennt eine vorgegebene Anzahl etwa gleichgroßer, kugelförmiger Punktwolken im $p$-dimensionalen Raum. Jeder Cluster wird durch seinen Mittelpunkt dargestellt. Die Mittelpunkte werden Prototypen genannt, da sie als Stellvertreter aller zugeordneter Daten angesehen werden. Die Clustermittelpunkte werden bei gegebenen Zugehörigkeiten der Daten zu den Clustern in Form einer verallgemeinerten Mittelwertbildung bestimmt, woher der Algorithmus seinen Namen hat.

Die Initialisierung dieses Verfahrens erfordert große Kenntnis der zu untersuchenden Daten. Die Anzahl der Cluster, die Clusterprototypen oder die Zugehörigkeiten der Objekte zu den Clustern müssen vorgegeben werden. Die Cluster sollten annähernd gleich groß und so kompakt wie möglich sein. Diese *a-priori*-Annahmen sind notwendig, aber nicht hinreichend dafür, ein globales Minimum garantieren zu können.

Das Verfahren arbeitet anders als hierarchisch-agglomerative Verfahren nicht auf einer Matrix paarweiser Distanzen eines beliebigen Abstandsmaßes, sondern berechnet die quadrierten euklidischen Distanzen direkt aus den Daten. Dies setzt voraus, das die Objekte Elemente eines euklidischen Vektorraumes sind. Hard-$c$-Means minimiert die Summe der quadrierten Fehler zwischen den Objekten und den Clusterprototypen

$$\min_{U,V} \left\{ J_1(U,V) := \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}\, d^2(x_k, v_i) \right\},$$

wobei $n$ die Größe der Datenmenge $X$, $x_k \in \mathbb{R}^p$, $k = 1, \ldots, n$, deren Objekte, $c$ die Anzahl der Cluster, $U := (u_{ik})_{1 \le i \le c,\, 1 \le k \le n} \in \{0,1\}^{c \times n}$ eine scharfe $c$-Partition von $X$, $V := (v_1, v_2, \ldots, v_c) \in \mathbb{R}^{c \times p}$ die Matrix der Clusterprototypen $v_i \in \mathbb{R}^p$, $i = 1, \ldots, c$, und

$$\underset{1 \le i \le c}{\forall} \; \underset{1 \le k \le n}{\forall} \; d(x_k, v_i) = \|x_k - v_i\|_A = \sqrt{(x_k - v_i)^T A (x_k - v_i)}$$

die durch eine Matrix $A$ induzierte Vektornorm im $\mathbb{R}^p$ bestimmten Abstände zwischen den Objekten und den Prototypen sind. Wie gewöhnlich sei auch hier $A$ die Einheitsmatrix $E$ und folglich $d$ der euklidische Abstand in $\mathbb{R}^p$.

Liegt $x_k$ im $i$-ten Cluster, so ist $u_{ik} = 1$. Dann gibt $d^2(x_k, v_i)$ die quadrierte Abweichung der Repräsentation von $v_i$ durch $x_k$, an. Ist $x_k$ nicht Element des $i$-ten Clusters, wird der Beitrag zur Abweichungssumme $J_1$ wegen $u_{ik} = 0$ gleich 0 gesetzt. Der $i$-te Summand ist die Summe der quadrierten Fehler innerhalb des $i$-ten Clusters. Er gibt die Streuung des Clusters an. $J_1$ ist die Summe der quadrierten Intraclusterfehler. Die Streuungsmatrix des Clusters $u_i$ ist

$$W_i = \sum_{k=1}^{n} u_{ik} (x_k - v_i)(x_k - v_i)^T.$$

Mit $W = \sum_{i=1}^{c} W_i$ gilt wegen Verwendung der euklidischen Norm

$$\mathrm{Spur}(W) = J_1(U, V)$$

als Spezialfall für $m = 1$ von Gleichung (3.10). Hard-$c$-Means minimiert die Spur der Intraclustervarianzmatrix. $J_1$ ist ein Maß für die lokale Dichte. $J_1$ ist klein, wenn die Punkte in jedem Cluster $u_i$ dicht am Clusterzentrum $v_i$ liegen. Dies läßt für $J_1$ eine geometrische Interpretation zu. Wegen des euklidischen Abstandes können nur hypersphärische Gebilde modelliert werden. Die Cluster sind konvex.

$V^{(t)} := \left(v_1^{(t)}, \ldots, v_c^{(t)}\right)$ bezeichne die Matrix der Prototypen und $U^{(t)} := \left(u_{ik}^{(t)}\right)_{1 \le i \le c, 1 \le k \le n}$ die $c$-Partition des $t$-ten Iterationsschrittes, $t \in \mathbb{N} \cup \{0\}$. Die Prototypen $V^{(t)}$ im Schritt $t$ werden aus $U^{(t)}$ und $X$ durch

$$\underset{1 \le i \le c}{\forall} \; v_i^{(t)} := \frac{\sum_{k=1}^{n} u_{ik}^{(t)} x_k}{\sum_{k=1}^{n} u_{ik}^{(t)}} = \frac{\sum_{j=1}^{n_i} x_{i_j}}{n_i} = \bar{x}_i \tag{3.4}$$

bestimmt. Die Zugehörigkeiten $U^{(t+1)}$ des $(t+1)$-ten Schrittes berechnen sich aus den Prototypen $V^{(t)}$ des vorangegangen Schrittes $t$ und den Vektoren aus $X$ mit

$$\underset{1\le i\le c}{\forall}\ \underset{1\le k\le n}{\forall}\ u_{ik}^{(t+1)} := \begin{cases} 1, & \text{falls } \underset{j=1,\dots,c,\ j\neq i}{\forall} d\big(x_k, v_i^{(t)}\big) \le d\big(x_k, v_j^{(t)}\big) \\ 0, & \text{sonst.} \end{cases} \tag{3.5}$$

Der Algorithmus dieses Verfahrens durchläuft folgende Schritte:

(HCM1) Wähl $c$, $1 < c < n$, und bestimme den Wert $\varepsilon > 0$ für das Abbruchkriterium. Gib eine scharfe Partition $U^{(0)}$ vor. Setz $t := 0$.

(HCM2) Berechne die $c$ Mittelwertvektoren $v_i^{(t)}$ mit (3.4) aus $U^{(t)}$ und $X$.

(HCM3) Berechne die Zugehörigkeiten $U^{(t+1)}$ nach (3.5) aus $V^{(t)}$ und $X$.

(HCM4) Vergleiche $U^{(t)}$ und $U^{(t+1)}$ mit einer geeigneten Matrixnorm $\|\cdot\|$. Beende die Iteration, wenn $\|U^{(t+1)} - U^{(t)}\| \le \varepsilon$ oder fahre mit $t := t+1$ mit Schritt (HCM2) fort. Die nichtnegative reelle Zahl $\|U^{(t+1)} - U^{(t)}\|$ gibt den Abstand der beiden Matrizen an.

Wird der Prozeß mit Prototypen anstelle einer Partition initialisiert, sieht der Algorithmus wie folgt aus:

(HCM1′) Setz die Schranke $\varepsilon > 0$ für das Abbruchkriterium. Gib $c$, $1 < c < n$, Prototypen in Form einer Matrix $V^{(0)}$ vor.

(HCM2′) Berechne die Zugehörigkeiten $U^{(0)}$ nach (3.5) aus $V^{(0)}$ und $X$. Setz $t := 0$.

(HCM3′) Berechne die Mittelwertvektoren $V^{(t+1)}$ mit (3.4) aus $U^{(t)}$ und $X$.

(HCM4′) Berechne die Zugehörigkeiten $U^{(t+1)}$ nach (3.5) aus $V^{(t+1)}$ und $X$.

(HCM5′) Beende die Iteration, wenn $\|U^{(t+1)} - U^{(t)}\| \le \varepsilon$ oder fahre mit $t := t+1$ mit Schritt (HCM3′) fort.

Hard-$c$-Means weist einige Nachteile auf. Das Ergebnis hängt von der Anfangspartition oder den initialen Prototypen ab. Ebenfalls eine negativen Einfluß auf das Ergebnis üben diejenigen Objekte aus, deren Zuordnungen problematisch sind, weil sie weit außerhalb eines Clusters (Ausreißer) oder zwischen zwei Clustern liegen. Sind die Punkthäufungen in der Datenmenge von stark unterschiedlicher Größe, teilt Hard-$c$-Means die größeren auf mehrere kleine

Cluster auf. Ungeeignete Cluster werden nicht zusammengefaßt oder aufgeteilt. Die Clustergrenzen sind normierte Einheitskugeln. Die Formeln sind lediglich notwendige Kriterien für Extrema. Die Konvergenz zu einer minimalen Partition ist wegen lokaler Minima und Sattelpunkte nicht sichergestellt.

## 3.11 Unscharfe Verfahren zur Verbesserung einer Anfangspartition

Die scharfe Clusterung wird durch die gewöhnliche scharfe Zugehörigkeitsfunktion beschrieben. Jedes Objekt ist Element exakt eines Clusters. Gehört es zu einem Cluster, hat es den Zugehörigkeitswert 1, für alle anderen beträgt dieser Wert 0. Dabei werden wohldefinierte Grenzen zwischen den Clustern angenommen. Dieses Modell eignet sich nicht zur Beschreibung von Daten, in denen der Übergang von einem zum benachbarten Cluster eher graduell als abrupt ist und eine feiner abgestufte Beschreibung der Affinitäten der Objekte zu den Clustern notwendig macht.

Unscharfe Clustermethoden basieren auf der Theorie unscharfer Mengen (Zadeh, 1965). Hierbei ist die Bildmenge der Zugehörigkeitsfunktion das reelle Einheitsintervall $[0,1]$. Ein Wert nahe 1 zeigt einen höheren Grad der Zugehörigkeit als ein kleiner Wert nahe 0 an. Jedes Objekt ist mit jedem Cluster assoziiert.

### 3.11.1 Unscharfe c-Partition

Eine unscharfe $c$-Partition (*fuzzy c-partition*) ist definiert als $c \times n$-Matrix

$$U := (u_{ik})_{1 \le i \le c, 1 \le k \le n} \,,$$

die die Bedingungen

$$\underset{1 \le i \le c}{\forall} \; \underset{1 \le k \le n}{\forall} \; u_{ik} \in [0,1] \,, \tag{3.6}$$

$$\underset{1 \le k \le n}{\forall} \; \sum_{i=1}^{c} u_{ik} = 1, \tag{3.7}$$

$$\underset{1 \le i \le c}{\forall} \; 0 < \sum_{k=1}^{n} u_{ik} \tag{3.8}$$

erfüllt. Die Cluster werden durch Zugehörigkeitsfunktionen $u_i : X \rightarrow [0,1]$, die diesen Bedingungen mit $u_{ik} := u_i(x_k)$ genügen, beschrieben. Die Zugehörigkeitsfunktion eines jeden Clusters weist nun allgemeiner jedem Objekt $x_k$ einen Wert aus dem Einheitsintervall zu (vgl. Abschnitt 3.8.3.2). Nach (3.7) addieren sich die Zugehörigkeiten eines Objektes auf die einzelnen Cluster stets zu 1 auf. Die Summe in (3.8) wird als unscharfe Anzahl (*fuzzy number*) der Objekte im $i$-ten Cluster oder unscharfe Kardinalität (*fuzzy cardinality*) des $i$-ten Clusters bezeichnet. Wegen (3.8) ist kein Cluster leer, das heißt, wenigstens ein Objekt hat eine Zugehörigkeit größer 0, und es gibt keinen Cluster, welcher alle Objekte mit voller Zugehörigkeit enthält. Jedes Objekt ist also Element eines jeden Clusters mit einer Zugehörigkeit zwischen einschließlich 0 und 1. Eine $c$-Partition ist maximal unscharf, wenn alle Zugehörigkeitswerte $1/c$ sind.

Eine unscharfe $c$-Partition, die nur die Bedingungen (3.6) und (3.7) erfüllt, heißt *degeneriert.* Eine degenerierte $c$-Partition kann bei einer unscharfen Clusteranalyse auftreten, wenn ein Prototyp weitab aller natürlichen Cluster der Datenmenge liegt, und ihm deshalb kein Element zugewiesen wird. Solch ein Prototyp ist ein Indiz für eine ungeeignete Initialisierung des Verfahrens.

### 3.11.2 Verfahren

Unscharfe Clustermethoden sind Partitionierungsmethoden. Sie bestimmen die unscharfe Partition der Datenmenge, die bezüglich einer vorgegebenen Kostenfunktion optimal ist. Die Wahl guter Anfangsbedingungen ist eine der schwierigsten Fragen im Zusammenhang mit unscharfen Optimierungsverfahren, da unscharfe Clusteranalyseverfahren ziemlich sensibel bezüglich der Initialisierung sind. Auch unscharfe Clusteralgorithmen zerlegen die Daten in eine vorgegebene Anzahl von Clustern, gleichgültig, ob diese Cluster die inhärente Struktur der Daten adäquat wiedergeben oder nicht. Die Gültigkeit der Cluster muß nach der Clusterung separat ausgewertet werden. Typischerweise werden externe Kriterien verwendet.

#### 3.11.2.1 Fuzzy-c-Means

Fuzzy-$c$-Means ist die Erweiterung von Hard-$c$-Means auf unscharfe Partitionen (Dunn, 1974). Bezdek (1981) hat das Verfahren auf überabzählbar viele Bewertungsfunktionen $J_m(U,V)$ mit $1 \leq m < \infty$ verallgemeinert. Damit ist

DUNNS Funktion mit $m = 2$ nur noch ein Spezialfall. Fuzzy-$c$-Means minimiert die Kostenfunktion

$$J_m(U,V) := \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \, d^2(x_k, v_i)$$

einer unscharfen $c$-Partition $U$ und einer Matrix $V$ von Prototypen, wobei der Parameter $m \in [1, \infty)$ ein die Unschärfe steuernder gewichtender Exponent (*fuzzifier*) ist, $x_k \in \mathbb{R}^p$ das $k$-te Objekt, $v_i \in \mathbb{R}^p$ der Schwerpunkt des $i$-ten Clusters, $d$ eine innere Produktmetrik von $\mathbb{R}^p$, $n$ die Anzahl der Datenpunkte und $c$ die Anzahl der Cluster. Für $m = 1$, der euklidischen Metrik in $\mathbb{R}^p$ und scharfem $U$ entspricht die Methode dem Hard-$c$-Means. Für $m > 1$ werden die Cluster unscharf. Der Parameter $m$ bestimmt den Grad der Unschärfe. Je größer $m$ ist, desto stärker wird die Zugehörigkeit gemindert und die Unschärfe nimmt zu. Dann fallen die Ausprägungen von lokalen Extrema weniger deutlich aus. Sind die Cluster weit voneinander entfernt, werden sie bei einem $m$ nahe 1 schärfer getrennt. Bei nahezu nicht unterscheidbaren Clustern ist $m$ sehr groß zu wählen. Werte zwischen 1,1 und 5 sind üblich. Meist wird $m = 2$ und die euklidische Distanz verwendet.

Die unscharfe Streuungsmatrix (*fuzzy scatter matrix*) $W_i = \left(w_{rt}^{(i)}\right)_{1 \le r,t \le p}$ des Clusters $u_i$ ist definiert durch

$$W_i := \sum_{k=1}^{n} u_{ik}^m \, (x_k - v_i) \, (x_k - v_i)^T \in \mathbb{R}^{p \times p}. \tag{3.9}$$

und für ihre Komponenten gilt folglich

$$\underset{1 \le r,t \le p}{\forall} \; w_{rt}^{(i)} = \sum_{k=1}^{n} u_{ik}^m \, (x_{kr} - v_{ir})(x_{kt} - v_{it}).$$

Die unscharfe Intraclustervarianzmatrix

$$W := \sum_{i=1}^{c} W_i.$$

ist symmetrisch. Diese unscharfen Variante von $c$-Means ist durch die potenzierten Zugehörigkeitswerte gewichtet. Bei Verwendung des euklidischen Abstandes ist

die Spur der unscharfen Intraclustervarianzmatrix gleich der Kostenfunktion $J_m$, denn

$$\begin{aligned}
\operatorname{Spur}(W) &= \operatorname{Spur}\left(\sum_{i=1}^{c} W_i\right) \\
&= \sum_{i=1}^{c} \operatorname{Spur}(W_i) \\
&= \sum_{i=1}^{c} \operatorname{Spur}\left(\sum_{k=1}^{n} u_{ik}^m \left(x_k - v_i\right)\left(x_k - v_i\right)^T\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \operatorname{Spur}\left(\left(x_k - v_i\right)\left(x_k - v_i\right)^T\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \left(\sum_{r=1}^{p} \left(x_{kr} - v_{ir}\right)^2\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \, d^2(x_k, v_i) \\
&= J_m(U, V). \qquad (3.10)
\end{aligned}$$

Damit verallgemeinert die Kostenfunktion $J_m$ die Partitionierung aufgrund minimaler Varianzen (Bezdek, 1981, S. 78).

Die Matrix der unscharfen Punktprototypen (*fuzzy prototypes*) $V^{(t)}$ des $t$-ten Schrittes ergibt sich mit $U^{(t)}$ durch

$$\underset{1\le i\le c}{\forall} \quad v_i^{(t)} := \frac{\sum_{k=1}^{n} \left(u_{ik}^{(t)}\right)^m x_k}{\sum_{k=1}^{n} \left(u_{ik}^{(t)}\right)^m} \qquad (3.11)$$

und somit komponentenweise durch

$$\underset{1\le i\le c}{\forall} \quad \underset{1\le r\le p}{\forall} \quad v_{ir}^{(t)} = \frac{\sum_{k=1}^{n} \left(u_{ik}^{(t)}\right)^m x_{kr}}{\sum_{k=1}^{n} \left(u_{ik}^{(t)}\right)^m}.$$

Der Prototyp eines Clusters ist dessen durch die standardisierten Zugehörigkeiten $\left(u_{ik}^{(t)}\right)^m / \sum_{l=1}^{n} \left(u_{il}^{(t)}\right)^m$ gewichtetes arithmetisches Mittel. Die Berechnung der

unscharfen $c$-Partition $U^{(t+1)}$ des $(t+1)$-ten Iterationsschrittes aus den Prototypen $V^{(t)}$ des $t$-ten Schrittes erfordert ein wenig mehr Aufwand. Es sei

$$\underset{1\le k\le n}{\forall} \; I_k^{(t)} := \left\{ i : 1 \le i \le c,\; d\big(x_k, v_i^{(t)}\big) = 0 \right\}$$

die Menge der Clusterindizes, für die der Datenpunkt $x_k$ zum Prototypen $v_i^{(t)}$ den Abstand 0 hat, und

$$\underset{1\le k\le n}{\forall} \; \left(I_k^{(t)}\right)^c = \{1, 2, \ldots, c\} \setminus I_k^{(t)}$$

das Komplement von $I_k^{(t)}$. Nun werden die Komponenten von $U^{(t+1)}$ bestimmt durch

$$\underset{1\le k\le n}{\forall} \left( I_k^{(t)} = \emptyset \quad \Longrightarrow \quad \underset{1\le i\le c}{\forall} \; u_{ik}^{(t+1)} := \frac{\left(\dfrac{1}{d^2\big(x_k, v_i^{(t)}\big)}\right)^{\frac{1}{m-1}}}{\displaystyle\sum_{j=1}^{c} \left(\dfrac{1}{d^2\big(x_k, v_j^{(t)}\big)}\right)^{\frac{1}{m-1}}} \in [0,1], \right.$$

$$\left. I_k^{(t)} \ne \emptyset \quad \Longrightarrow \quad \underset{i\in\; I_k^{(t)}{}^c}{\forall} \; u_{ik}^{(t+1)} := 0 \quad \text{und} \quad \sum_{i\in I_k^{(t)}} u_{ik}^{(t+1)} = 1 \right) \tag{3.12}$$

(Bezdek, 1981, S. 68). Der zweite Fall tritt ein, wenn wenigstens ein Punkt zu einem der Prototypen den Abstand 0 hat. Die Zugehörigkeitswerte $u_{ik}^{(t+1)}$ für $i \in I_k^{(t)}$ müssen dann die Spaltenbedingung (3.7) erfüllen, sind ansonsten aber beliebig, etwa $u_{ik}^{(t+1)} := 1/|I_k^{(t)}|$. Die Bedingungen (3.11) und (3.12) sind notwendig, aber nicht hinreichend, um ein globales Minimum bestimmen zu können (Bezdek, 1981, S. 66).

Der Fuzzy-$c$-Means-Algorithmus durchläuft die gleichen Schritte wie der Hard-$c$-Means-Algorithmus. Er gilt nur für $m > 1$. Für $m = 1$ ist Hard-$c$-Means anzuwenden. Er lautet:

(FCM1) Leg die Anzahl der Cluster $c$, $1 < c < n$ und den Fuzzifier $m$, $1 < m < \infty$ fest. Setz den Wert $\varepsilon > 0$ für das Abbruchkriterium und/oder eine feste Anzahl von Iterationen $t_{\max}$. Wähl eine innere Produktmetrik $d$ für $\mathbb{R}^p$. Gib die initiale unscharfe Partition $U^{(0)}$ vor. Wiederhol für $t = 0, 1, \ldots$ folgende Schritte:

(FCM2) Berechne die Clusterprototypen $V^{(t)}$ aus (3.11) mit $U^{(t)}$ und $X$.

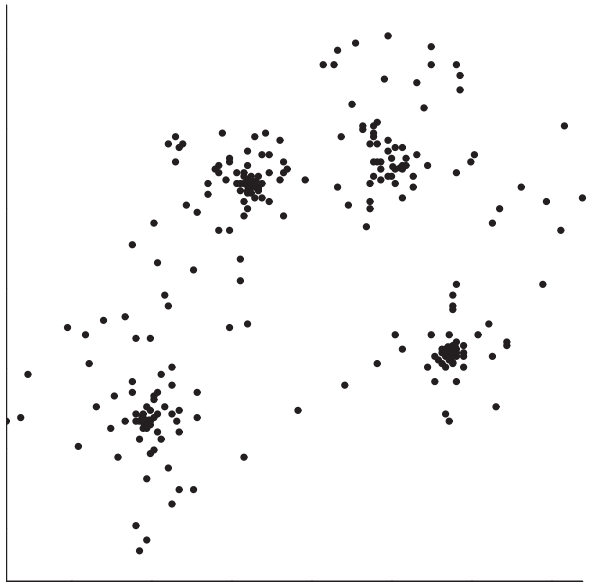


**Abbildung 3.3:** Vier ineinander übergehende Punkthäufungen aus 254 Punkten

(FCM3) Berechne die Matrix der unscharfen Zugehörigkeiten $U^{(t+1)}$ nach (3.12) aus $V^{(t)}$ und $X$.

(FCM4) Vergleiche $U^{(t)}$ und $U^{(t+1)}$ mit einer Matrixnorm $\|\cdot\|$. Beende die Iteration, wenn $\|U^{(t+1)} - U^{(t)}\| \leq \varepsilon$ oder die vorgegeben Anzahl $t_{\max}$ an Iterationen durchgeführt wurde. Ansonsten beginne den $(t+1)$-ten Schritt mit (FCM2).

Auch Fuzzy-$c$-Means und seine Varianten können anstatt mit einer Partition mit Prototypen initialisiert werden. Der Algorithmus ist dann, wie beim Hard-$c$-Means gezeigt, zu verändern.

Fuzzy-$c$-Means weist einige Nachteile auf. Die Anzahl der Cluster $c$ muß vorgegeben werden. Ungeeignete Cluster werden nicht zusammengefaßt oder aufgeteilt. Der Durchschnitt eines Clusters ist nicht immer dessen bester Repräsentant. Die Clustergrenzen sind normierte Einheitskugeln. Die zu minimierende Funktion $J_m$ ist nicht konvex, was zur Konvergenz zu einem lokalen Minimum oder einem Sattelpunkt führen kann. Cluster unterschiedlicher Form, Größe und Dichte werden nicht erkannt. Eine automatische Anpassung der Clusterform – möglichst für jeden Cluster einzeln – ist durch Fuzzy-$c$-Means nicht möglich. Fuzzy-$c$-Means hat aber den Vorteil, daß jede Iteration eine Verbesserung, das heißt eine Verkleinerung des Wertes von $J_m$ bewirkt. Sowohl Hard-$c$-Means als auch Fuzzy-$c$-Means minimieren die Varianz der Daten

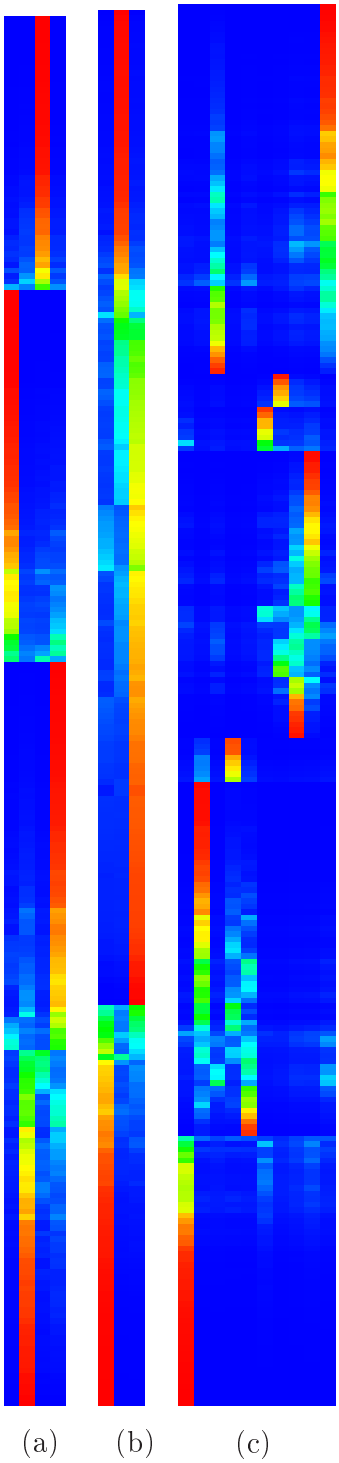



**Abbildung 3.4:** Datenbilder unscharfer Partitionen der vier Punkthäufungen aus Abbildung 3.3: (a) unscharfe 4-Partition, (b) falsch klassifizierte unscharfe 3-Partition, (c) falsch klassifizierte unscharfe 10-Partition. Die Zeilen und Spalten sind jeweils nach Average-Linkage der euklidischen Distanzen umgestellt.

innerhalb der Cluster. Vgl. Bezdek (1981, S. 65–70), Bezdek (1987), Bezdek (1998, S. F6.2:3 f.), Höppner et al. (1997, S. 35–41).

#### 3.11.2.2 Gustafson-Kessel

Gustafson und Kessel (1978) haben Fuzzy-$c$-Means so erweitert, daß es Cluster mit unterschiedlichen hyperellipsoiden Formen in derselben Datenmenge erkennt. Dazu wird jeder Cluster $i$ durch eine eigene Matrix $A_i \in \mathbb{R}^{p\times p}$ charakterisiert. Diese Matrizen induzieren für jeden Cluster eine eigene gewichtete Norm $\|x\|_{A_i} = \sqrt{x^T A_i x}$ in $\mathbb{R}^p$. Die Distanz zwischen dem Prototyp $v_i$ des $i$-ten Clusters, $i = 1, \ldots, c$, und dem $k$-ten Objekt, $k = 1, \ldots, n$, in dieser Norm ist

$$d\left(x_k, v_i\right) := \|x_k - v_i\|_{A_i} = \sqrt{\left(x_k - v_i\right)^T A_i \left(x_k - v_i\right)}. \tag{3.13}$$

In $A_i$ ist eine Drehung und eine Streckung enthalten. Dies ermöglicht beliebige ellipsoide Formen. Bei jeder Iteration verbessert der Algorithmus automatisch die Anpassung der Distanzfunktion an die Cluster. Der Gustafson-Kessel-Algorithmus ist also ein lokal adaptiver unscharfer Clusteralgorithmus. Um zu verhindern, daß beim Minimieren auch die Matrizen minimal werden, wird für jeden Cluster ein konstantes Volumen durch $\det(A_i) > 0$ gefordert. Somit ist nur die Clusterform, nicht aber die Clustergröße variabel. Die Wahl der Konstanten setzt Wissen über die Cluster voraus. Das Gustafson-Kessel-Modell ist durch das Problem

$$\min_{U,V,A} \left\{ J_m^{GK}(U,V,A) := \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \, \|x_k - v_i\|_{A_i}^2 \right\}$$

definiert, wobei alle Variablen und Parameter die gleichen wie beim Fuzzy-$c$-Means sind und $A := (A_1, A_2, \ldots, A_c)$ ein Vektor symmetrischer, positiv definiter, norminduzierender Matrizen ist (Abschnitt 2.6). Die zusätzliche Bedingung $\det(A_i) > 0$ stellt sicher, daß $A_i$ positiv definit ist. Eine weitere wesentliche Bedingung ist

$$\underset{1\le i\le c}{\forall} \; A_i = \left(\rho_i \det(C_i)\right)^{1/p} C_i^{-1}, \tag{3.14}$$

wobei

$$\underset{1\le i\le c}{\forall} \; C_i := \frac{\displaystyle\sum_{k=1}^{n} u_{ik}^m \; \left(x_k - v_i\right)\left(x_k - v_i\right)^T}{\displaystyle\sum_{k=1}^{n} u_{ik}^m} = \frac{W_i}{\displaystyle\sum_{k=1}^{n} u_{ik}^m} \tag{3.15}$$

die sogenannte unscharfe Kovarianzmatrix (*fuzzy covariance matrix*) von $u_i$ und $W_i$ die unscharfe Streuungsmatrix (3.9) von $u_i$ ist. Es ist

$$\det(A_i) = \left(\sqrt[p]{\rho_i \det(C_i)}\right)^p \det(C_i^{-1}) = \rho_i \det(C_i)\, \frac{1}{\det(C_i)} = \rho_i.$$

Die Prototypen (3.11) werden genau wie beim Fuzzy-$c$-Means berechnet. Ebenso, aber mit der für jeden Cluster einzeln angepaßten Norm $d\,(x_k, v_i) = \|x_k - v_i\|_{A_i}$, ergibt sich die unscharfe $c$-Partition mit (3.12). Die Aussagen (3.11), (3.12) und (3.14) sowie Symmetrie und positive Definitheit der norminduzierenden Matrizen sind notwendig, damit der Algorithmus ein globales Minimum bestimmen kann. Der Gustafson-Kessel-Algorithmus lautet:

(GK1) Lege die Clusteranzahl $c$, $1 < c < n$, den Fuzzifier $m \in (1, \infty)$, die das Volumen der Cluster beschränkenden Konstanten $\rho_i \in (0, \infty)$, $1 \leq i \leq c$, und den Schwellenwert für die Terminierung $\varepsilon > 0$ fest. Setze eine initiale unscharfe Partition $U^{(0)}$. Wiederhole für $t = 0, 1, \ldots$ die folgenden Schritte:

(GK2) Berechne die Prototypen $V^{(t)}$ mit (3.11) aus $U^{(t)}$ und $X$.

(GK3) Berechne die $c$ unscharfen Kovarianzmatrizen $C_i$ mit (3.15) aus $U^{(t)}$, $V^{(t)}$ und $X$, deren Determinanten und Inversen.

(GK4) Berechne die $c$ norminduzierenden Matrizen $A_i$ mit (3.14).

(GK5) Berechne die Distanzen $\big(d(x_k, v_i)\big)_{1 \leq i \leq c,\, 1 \leq k \leq n}$ der Datenpunkte zu den Prototypen mit (3.13).

(GK6) Berechne die neue Partitionsmatrix $U^{(t+1)}$ mit (3.12).

(GK7) Vergleiche $U^{(t)}$ und $U^{(t+1)}$ mit einer geeigneten Matrixnorm $\|\!|\cdot|\!\|$. Beende die Iteration, wenn $\|\!|U^{(t+1)} - U^{(t)}|\!\| \leq \varepsilon$ oder setze $t := t + 1$ und fahre mit Schritt (GK2) fort.

Das Gustafson-Kessel-Verfahren ist empfindlicher gegenüber Initialisierungen als Fuzzy-$c$-Means. Die Resultate von Fuzzy-$c$-Means sind gute Initialisierung für dieses Verfahren. Es werden meist kleine Fuzzifier verwendet, $1 < m \leq 2$. Ein häufig eingesetzter Wert ist $m = 1{,}5$. Vgl. Bezdek (1998, S. F6.2:9 f.).

#### 3.11.2.3 Gath-Geva

Der Algorithmus von GATH und GEVA (1989) ist eine Erweiterung des GUSTAFSON-KESSEL-Algorithmus. Er erkennt auch hyperellipsoide Cluster mit unterschiedlichen Größen und Dichten. Zu lösen ist

$$\min_{U,V} \left\{ J_m^{GG}(U,V) := \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \, d^2(x_k, v_i) \right\}$$

mit dem exponentiellen Distanzmaß

$$\begin{aligned} &\underset{1\le i\le c}{\forall} \; \underset{1\le k\le n}{\forall} \; d^2(x_k, v_i) \\ &\qquad := \frac{\sum_{k=1}^{n} \sum_{j=1}^{c} u_{jk}^m}{\sum_{k=1}^{n} u_{ik}^m} \sqrt{\det(C_i)} \; \exp\left( \frac{1}{2} (x_k - v_i)^T C_i^{-1} \, (x_k - v_i) \right). \end{aligned} \tag{3.16}$$

$C_i$ ist die unscharfe Kovarianzmatrix (3.15) des $i$-ten Clusters. Alle anderen Variablen und Parameter sind wie bei GUSTAFSON-KESSEL. Die Zugehörigkeiten werden genauso berechnet wie bei Fuzzy-$c$-Means (3.12). In der Formel (3.11) für die Prototypen werden die Distanzen mit (3.16) bestimmt. Diese Bedingungen sind notwendig, um ein globales Minimum erlangen zu können. Der Algorithmus zur Minimierung von $J_m^{GG}$ unter diesen Bedingungen läßt sich wie folgt beschreiben:

(GG1) Lege die Clusteranzahl $c$, $1 < c < n$, den Fuzzifier $m \in (1, \infty)$ und den Schwellenwert für die Terminierung $\varepsilon > 0$ fest. Initialisiere eine unscharfe Partition $U^{(0)}$. Wiederhole für $t = 0, 1, \ldots$ die folgenden Schritte:

(GG2) Berechne die $c$ unscharfen Prototypen $v_i^{(t)}$ mit (3.11) aus $U^{(t)}$ und $X$.

(GG3) Berechne die $c$ unscharfen Kovarianzmatrizen $C_i$ mit (3.15) aus $U^{(t)}$, $V^{(t)}$ und $X$, deren Determinanten und Inversen.

(GG4) Berechne die neue Partitionsmatrix $U^{(t+1)}$ nach (3.12) mit (3.16) aus $V^{(t)}$ und $X$.

(GG5) Vergleiche $U^{(t)}$ und $U^{(t+1)}$ mit einer geeigneten Matrixnorm $\|\cdot\|$. Beende, wenn $\|U^{(t+1)} - U^{(t)}\| \le \varepsilon$ gilt, ansonsten erhöhe $t$ um 1 und fahre mit Schritt (GG2) fort.

Wegen des Exponenten in der Distanzfunktion sucht GATH-GEVA das Optimum in einer engen lokalen Umgebung. Das Verfahren ist deswegen noch anfälliger gegenüber ungeeigneter Initialisierungen. Für die richtige Einteilung durch diesen Algorithmus müssen die initialen Prototypen bereits in der Nähe der endgültigen Prototypen liegen. Der GUSTAFSON-KESSEL-Algorithmus liefert gute Anfangsbedingungen. Im allgemeinen werden Clusteralgorithmen mit zunehmender Komplexität anfälliger für lokale Minima. In den Bereichen des Übergangs von einem Cluster zum nächsten wechseln die Zugehörigkeiten sehr schnell von 0 auf 1. Die Cluster werden so äußerst genau getrennt. Vgl. HÖPPNER et al. (1997, S. 48 ff.).

#### 3.11.2.4 Achsenparalleler Gath-Geva

Bei der vereinfachten achsenparallelen Variante des GATH-GEVA-Algorithmus werden die unscharfen Kovarianzmatrizen auf Diagonalmatrizen $C_i := \left(c_{rt}^{(i)}\right)_{1\leq r,t\leq p} \in \mathbb{R}^{p\times p}$, die nur die Achsen skalieren, ohne eine Drehung vorzunehmen, beschränkt. Deren Diagonalelemente sind

$$\underset{1\leq i\leq c}{\forall}\ \underset{1\leq r\leq p}{\forall}\ c_{rr}^{(i)} := \frac{\sum\limits_{l=1}^{n} u_{il}^{m}\,(x_{lr}-v_{ir})^2}{\sum\limits_{l=1}^{n} u_{il}^{m}}.$$

Sie beschreiben die unscharfe Streuung des $i$-ten Clusters entlang der $r$-ten Achse. Dies entspricht der unscharfen Streuung der $r$-ten Komponenten der Objektvektoren bezüglich $v_{ir}$. Somit werden nur achsenparallele Hyperellipsoiden erkannt. Die Inversen und Determinanten der Matrizen sind dadurch sehr viel einfacher zu berechnen. Die Determinante einer Diagonalmatrix ist das Produkt der Diagonalelemente. Die Inverse einer Diagonalmatrix ist die Diagonalmatrix der Kehrbrüche. Es gilt

$$\begin{aligned}
&\underset{1\leq i\leq c}{\forall}\ \underset{1\leq k\leq n}{\forall}\ (x_k-v_i)^T C_i^{-1}(x_k-v_i)\\
&\qquad = \left((x_{k1}-v_{i1})\,\frac{1}{c_{11}^{(i)}},\,(x_{k2}-v_{i2})\,\frac{1}{c_{22}^{(i)}},\,\ldots,\,(x_{kp}-v_{ip})\,\frac{1}{c_{pp}^{(i)}}\right)^T (x_k-v_i)\\
&\qquad = \sum_{r=1}^{p} \frac{1}{c_{rr}^{(i)}}(x_{kr}-v_{ir})^2.
\end{aligned}$$

Für (3.16) ergibt sich somit

$$\underset{1\le i\le c}{\forall}\ \underset{1\le k\le n}{\forall}\ d^2(x_k, v_i)$$
$$= \frac{\sum_{j=1}^{c}\sum_{l=1}^{n} u_{jl}^m}{\sum_{l=1}^{n} u_{il}^m} \sqrt{\prod_{r=1}^{p} c_{rr}^{(i)}}\ \exp\left(\frac{1}{2}\left(\sum_{l=1}^{n} u_{il}^m\right)\sum_{r=1}^{p}\frac{(x_{kr}-v_{ir})^2}{\sum_{l=1}^{n} u_{il}^m (x_{lr}-v_{ir})^2}\right).$$

## 3.12 Nachbarschaften

Bisher wurde die Ähnlichkeit zweier Objekte immer durch eine reelle Zahl beschrieben. Es ist auch möglich, Ähnlichkeit über die Umgebungen von Punkten zu definieren: Zwei Punkte sind ähnlich, wenn sie dieselbe Nachbarschaft haben. Bei dieser Definition der Ähnlichkeit werden nicht die beiden Punkte selbst, sondern nur ihre Nachbarschaften betrachtet. Die Nachbarschaften selbst können auf verschiedene Weisen definiert werden:

### 3.12.1 Radius

Für einen gegebenen Radius $\varepsilon > 0$ ist die $\varepsilon$-Nachbarschaft $N_\varepsilon$ eines Punktes $x_0$ die Menge aller Punkte, die einen Abstand $d$ kleiner oder gleich $\varepsilon$ zu $x_0$ haben, $N_\varepsilon(x_0) := \{x \mid d(x_0, x) \le \varepsilon\}$.

### 3.12.2 Rangfolge

Die Nachbarschaft $N$ eines Punktes $x_0$ ist ein nach Abständen zu $x_0$ aufsteigend sortiertes Tupel aller übrigen Punkte des Datensatzes, $N(x_0) := (x_1, x_2, \ldots, x_n)$ mit $d(x_0, x_1) \le d(x_0, x_2) \le \ldots \le d(x_0, x_n)$. Haben zwei Punkte denselben Abstand zu $x_0$, sind deren Ränge miteinander vertauschbar und das Tupel nicht mehr eindeutig bestimmt. Diese beiden Punkte sind zwar bezüglich dieser Nachbarschaftsdefinition nicht unterscheidbar, aber im allgemeinen nicht identisch.

### 3.12.3 k-Nachbarschaft

Die $k$-Nachbarschaft $N_k$ eines Punktes $x_0$ besteht aus den $k$ Punkten mit den kleinsten Abständen zu $x_0$ und auch denjenigen Punkten, die den gleichen

Abstand zu $x_0$ haben, wie der am weitesten entfernte unter den ersten $k$ Punkten. Für jeden Punkt, der nicht Element der aus diesen Punkten gebildeten Menge $N_k$ ist, ist der Abstand zu $x_0$ echt größer als der Punkt aus $N_k$ mit dem größten Abstand zu $x_0$, also $d(y, x_0) > \max_{x \in N_k} d(x, x_0)$ für alle $y \notin N_k$. Betrachtet man auch die Rangfolge der Punkte entsprechend ihrer Abstände zu $x_0$, kann die $k$-Nachbarschaft auch als Tupel der ersten $k$ Elemente der Rangfolge aller Punkte zuzüglich derjenigen, die den gleichen Abstand zu $x_k$ haben wie die $k$-te Komponente, betrachtet werden (Abschnitt 3.12.2). Letztere Punkte müssen zur Nachbarschaft gezählt werden, um einen künstlichen Schnitt inmitten einer Menge äquivalenter Elemente zu vermeiden. Die $k$-Nachbarschaft hat also nicht genau $k$, sondern mindestens $k$ Elemente.

Über die Abstände der benachbarten Punkte zu $x_k$ wird nichts ausgesagt. So können für einen Punkt alle Nachbarn innerhalb eines kleinen Radius liegen, für einen anderen mögen sie weit verstreut sein. Hierdurch findet eine automatische Anpassung an die lokale Dichte im Datensatz statt. In dünn besiedelten Gebieten sind die Nachbarn auf ein größeres Gebiet verteilt als in dicht besiedelten.

## 3.13 Clusterprototypen

Die Partitionierungsverfahren müssen mit der richtigen Zahl der Cluster und der ungefähren Lage der Prototypen oder einer geeigneten Partition initialisiert werden. Das Ergebnis eines Optimierungsverfahrens kann wesentlich durch die gewählte Ausgangszerlegung beeinflußt werden. Die anfängliche Clusterkonfiguration wird mit Hilfe vorhandenen Wissens oder zufällig gewählt. Sie kann auch das Ergebnis eines vorangegangenen anderen Clusterprozesses sein. Zu Beginn der Untersuchung ist im allgemeinen die Zahl der zu lokalisierenden Cluster nicht bekannt. Die Analyse wird daher meist für mehrere verschiedene Werte durchgeführt. Drei populäre Methoden zur Wahl initialer Prototypen sind:

(1) Nimm die ersten $c$ verschiedenen Punkte der Datenmenge.

(2) Wähle zufällig $c$ verschiedene Punkte eines Hyperquaders (*hyperbox*), der die Datenmenge enthält. Die Punkte gehören im allgemeinen nicht zur Datenmenge.

(3) Wähle $c$ Punkte, die gleichmäßig auf einer Diagonalen eines Hyperquaders liegen, der die Datenmenge umfaßt. Die Punkte liegen auch hier im allgemeinen nicht in der Datenmenge.

Die Ergebnisse dieser Methoden, die eher einem Raten als einer zielgerichteten Auswahl gleichen, sind zu ungenau, um mit ihnen ein bezüglich den Startparametern sensibles Verfahren anzustoßen. Die Bestimmung der Clusteranzahl bleibt zudem offen.

### 3.13.1 Mountain-Function

Eine Mountain-Function ist ein Maß für die physikalische Dichte der Umgebungen von Punkten. Es ist wesentlich, daß einem Punkt, in dessen Nähe viele Punkte liegen, durch eine Mountain-Function ein höherer Wert zugewiesen wird als einem mit wenigen benachbarten Punkten. Die Werte werden als Höhen von Ortspunkten in einem Gebirge interpretiert. Die Höhe eines Punktes wird berechnet als Summe aller Zugehörigkeitswerte einer unscharfen Relation eines jeden Punktes mit dem Punkt selbst. Die Zugehörigkeitswerte der unscharfen Relation werden aufgrund der Abstände der Punkte bestimmt. Die unscharfe Relation $R$ wird gebildet aus einer die Punktabstände $d$ gewichtenden unscharfen Menge $f$, $R(x,y) := f\big(d(x,y)\big)$. Sie gibt an, wie stark benachbarte Punkte zur Höhe eines Punktes beitragen. Sie ist so zu gestalten, daß kleine Abstände für die Generierung der Höhen wesentlich sind, während große keinen oder nur einen geringen Einfluß ausüben. Die Funktion $f$ überführt Distanzen in Ähnlichkeiten. Die Mountain-Function berechnet die Höhe eines Punktes, der selbst nicht zur Datenmenge gehören muß, durch Aufaddieren der Ähnlichkeiten dieses Punktes zu allen Punkten der Datenmenge. Gehört der Punkt zur Datenmenge, geht auch seine Ähnlichkeit zu ihm selbst in die Summe ein.

Die Dichte wird bestimmt durch einen die Nachbarschaft bestimmenden Radius, außerhalb dessen die Zugehörigkeiten verschwindend gering sind. Ist der Radius kleiner als der kleinste Abstand zwischen allen Datenpunkten, so steht jeder Punkt für sich allein (Abbildung 3.5). Je größer der Radius ist, desto mehr benachbarte Punkte üben einen Einfluß aus. Ist der Radius größer als der größte aller Punktabstände, trägt jeder Punkt zur Höhe eines jeden anderen bei (Abbildung 3.6). Sinnvolle Radien liegen somit zwischen dem kleinsten und dem größten Abstand (Abbildung 3.7). Da die Daten durch einen Hyperwürfel beschränkt sind, ist der Radius ein Bruchteil seiner Breite. Werden die Abstände so in das Einheitsintervall skaliert, daß die größte Distanz 1 ist, ist der Radius ein prozentualer Anteil der größten Distanz.

Der Radius wird durch den Vergleich klassierter Häufigkeitsverteilungen der berechneten Höhen bestimmt. Sind die häufigsten Höhen klein, wurde ein zu

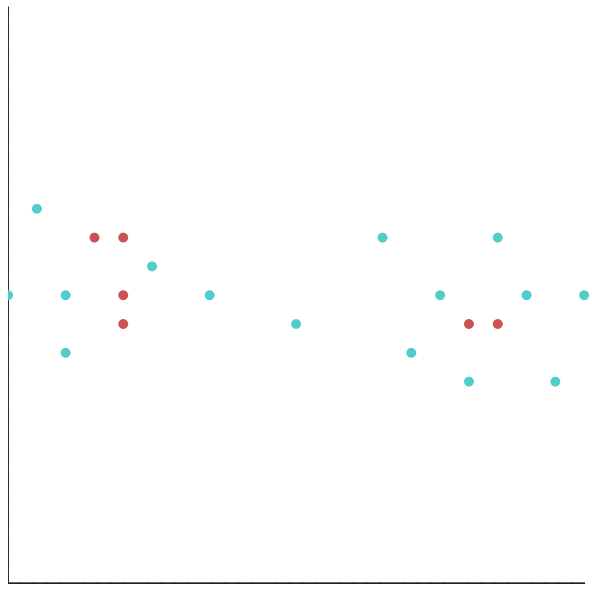


**Abbildung 3.5:** Zu kleiner Radius $c = 0{,}07$ von der größten Distanz für die Z-Funktion mit $a = 0$. Es gibt nur zwei verschiedene Höhen (Farbtöne). Nahezu jeder Punkt ist ein Gipfel.

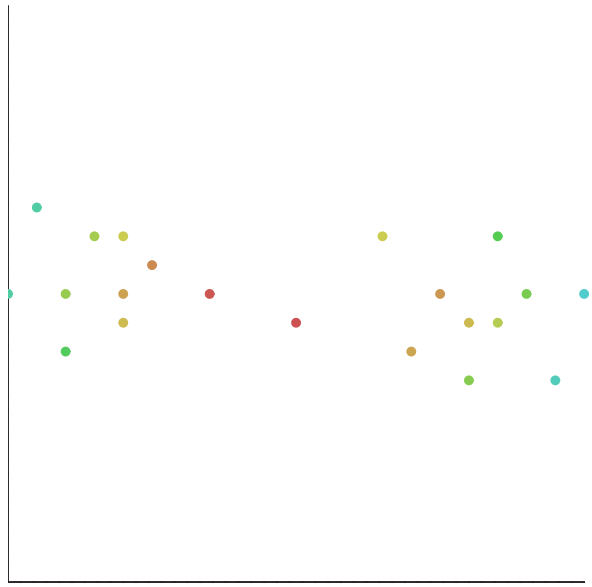


**Abbildung 3.6:** Zu großer Radius $c = 0{,}7$ von der größten Distanz für die Z-Funktion mit $a = 0$. Der zwischen den beiden Häufungen liegende Punkt ist am höchsten bewertet.

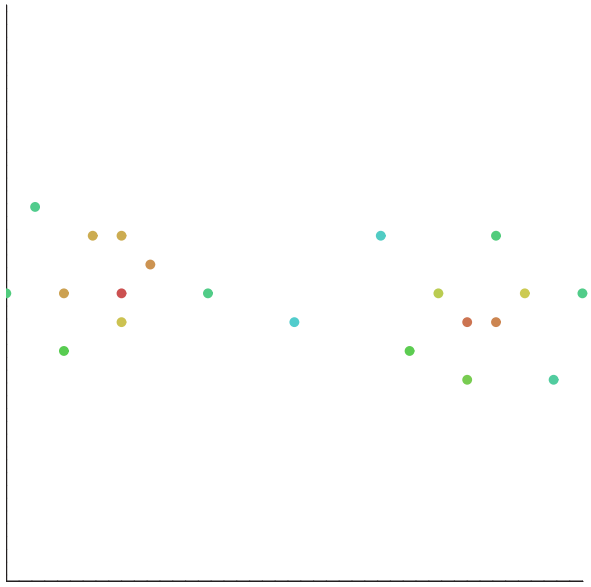


**Abbildung 3.7:** Geeigneter Radius $c = 0{,}2$ von der größten Distanz für die Z-Funktion mit $a = 0$. Die beiden Punkthäufungen lassen sich leicht als zwei Berge interpretieren.

kleiner Radius gewählt und somit zu wenig benachbarte Punkte betrachtet. Sind die meisten Höhen groß, wurden zu viele Punkte zur Nachbarschaft aufgrund eines zu großen Radius gezählt. Zur Bestimmung der Dichten ist derjenige Radius zu verwenden, bei dem die maximale Häufigkeit am niedrigsten ist. Dann überwiegen weder kleine noch große Radien. Ihr Verhältnis ist etwa ausgeglichen. Für die Konstruktion der unscharfen Relation eignen sich unscharfe Mengen für „klein“ oder unscharfe Zahlen für „0“.

#### 3.13.1.1 Triangular-Fuzzy-Number

Es sei $a < b < c$. Die Abbildung

$$f\colon \mathbb{R} \to [0,1],\ x \mapsto \begin{cases} 0, & \text{falls } x \leq a, \\ \dfrac{x-a}{b-a}, & \text{falls } a < x < b, \\ 1, & \text{falls } x = b, \\ \dfrac{c-x}{c-b}, & \text{falls } b < x < c, \\ 0, & \text{falls } x \geq c, \end{cases}$$

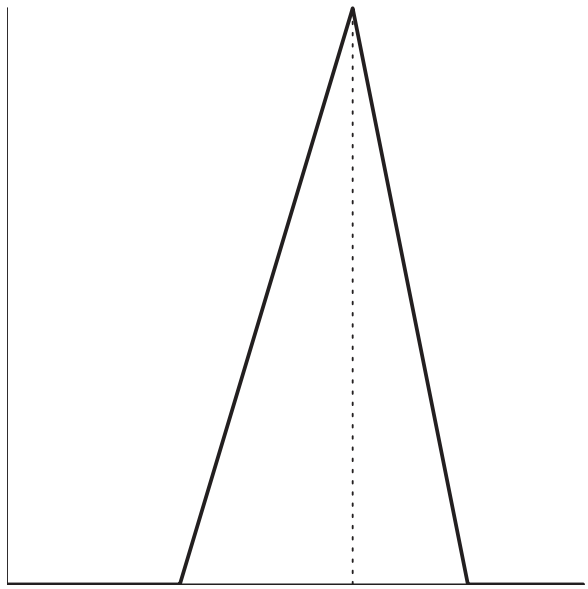


**Abbildung 3.8:** Triangular-Fuzzy-Number

heißt Triangular-Fuzzy-Number (Abbildung 3.8). Wird mit ihr als unscharfer Menge $f$ die unscharfe Relation $R$ definiert, ist $b := 0$ und $c$ der Radius. Da Abstände nie kleiner als 0 sind, kommt $a$ keine Bedeutung zu.

#### 3.13.1.2 Trapezoidal-Fuzzy-Number

Es seien $a < b \leq c < d$. Die Abbildung

$$f\colon \mathbb{R} \to [0,1],\ x \mapsto \begin{cases} 0, & \text{falls } x \leq a, \\ \dfrac{x-a}{b-a}, & \text{falls } a < x < b, \\ 1, & \text{falls } b \leq x \leq c, \\ \dfrac{d-x}{d-c}, & \text{falls } c < x < d, \\ 0, & \text{falls } x \geq d, \end{cases}$$

heißt Trapezoidal-Fuzzy-Number (Abbildung 3.9). Zur Definition der unscharfen Relation wird $b := 0$ gesetzt. Der Parameter $a$ ist dann ein beliebiger negativer Wert. Innerhalb des Radius $c$ werden alle benachbarten Punkte voll mit 1 gewichtet. Die Punkte außerhalb des Radius $d$ zählen gar nicht mehr.

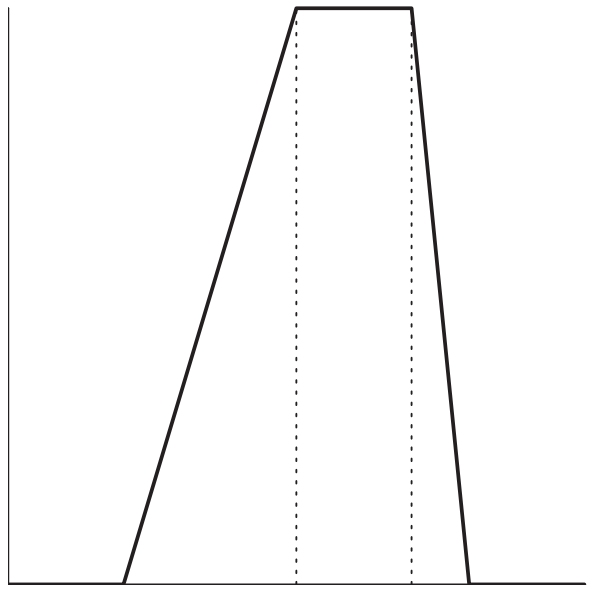


**Abbildung 3.9:** Trapezoidal-Fuzzy-Number

#### 3.13.1.3 Z-Funktion

Es sei $0 \leq a < c$ und $b := (a+c)/2$. In Anlehnung an die Form ihres Graphen wird die Abbildung

$$f\colon \mathbb{R} \to [0,1],\ x \mapsto \begin{cases} 1, & \text{falls } \ x \leq a, \\ 1 - 2\left(\dfrac{x-a}{c-a}\right)^2, & \text{falls } \ a < x \leq b, \\ 2\left(\dfrac{x-c}{c-a}\right)^2, & \text{falls } \ b < x < c, \\ 0, & \text{falls } \ x \geq c, \end{cases}$$

Z-Funktion genannt (Abbildung 3.10).

#### 3.13.1.4 Exponentialfunktion des Quadrates

Es sei $\alpha > 0$. Die Exponentialfunktion

$$f\colon \mathbb{R} \to (0,1],\ x \mapsto e^{-\alpha x^2}$$

nimmt den Wert 0 nie an. Deshalb kann lediglich ein die Nachbarschaft bestimmender Radius gewählt werden, außerhalb dessen die Zugehörigkeit

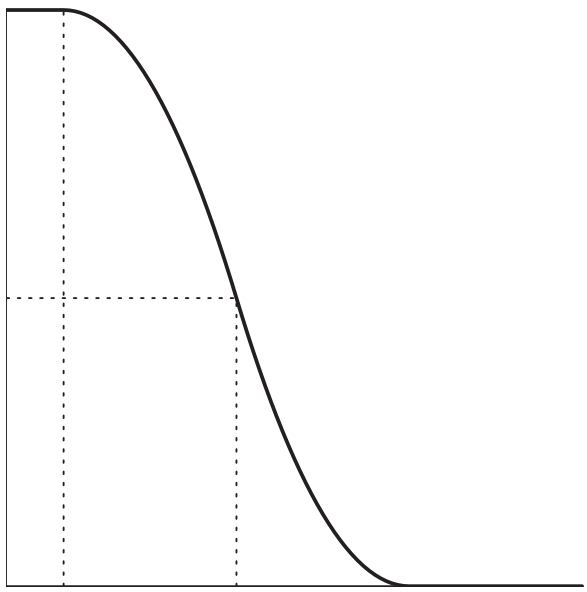


**Abbildung 3.10:** Z-Funktion

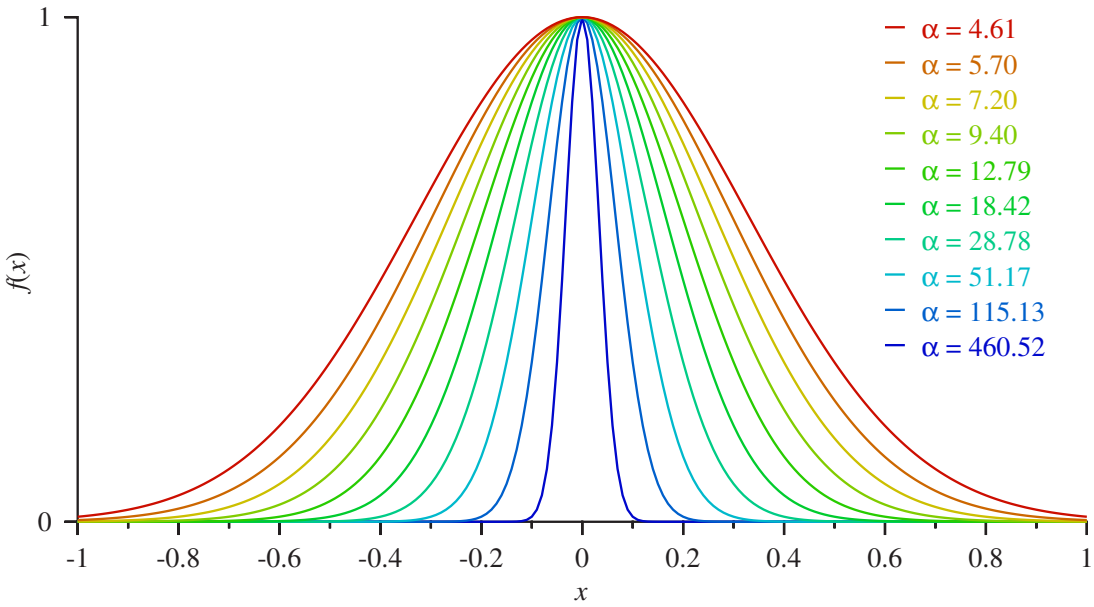



**Abbildung 3.11:** $f(x) = e^{-\alpha x^2}$, $x \in [-1, 1]$

| $\varepsilon = 0{,}01$ | | $\varepsilon = 0{,}001$ | |
|---|---|---|---|
| $r$ | $\alpha$ | $r$ | $\alpha$ |
| 0,05 | 1842,07 | 0,05 | 2763,10 |
| 0,10 | 460,52 | 0,10 | 690,78 |
| 0,15 | 204,67 | 0,15 | 307,01 |
| 0,20 | 115,13 | 0,20 | 172,69 |
| 0,25 | 73,68 | 0,25 | 110,52 |
| 0,30 | 51,17 | 0,30 | 76,75 |
| 0,35 | 37,59 | 0,35 | 56,39 |
| 0,40 | 28,78 | 0,40 | 43,17 |
| 0,45 | 22,74 | 0,45 | 34,11 |
| 0,50 | 18,42 | 0,50 | 27,63 |
| 0,55 | 15,22 | 0,55 | 22,84 |
| 0,60 | 12,79 | 0,60 | 19,19 |
| 0,65 | 10,90 | 0,65 | 16,35 |
| 0,70 | 9,40 | 0,70 | 14,10 |
| 0,75 | 8,19 | 0,75 | 12,28 |
| 0,80 | 7,20 | 0,80 | 10,79 |
| 0,85 | 6,37 | 0,85 | 9,56 |
| 0,90 | 5,69 | 0,90 | 8,53 |
| 0,95 | 5,10 | 0,95 | 7,65 |
| 1,00 | 4,61 | 1,00 | 6,91 |

**Tabelle 3.4:** Parameter $\alpha$ für Radius $r$ und $\varepsilon$ in $e^{-\alpha x^2}$

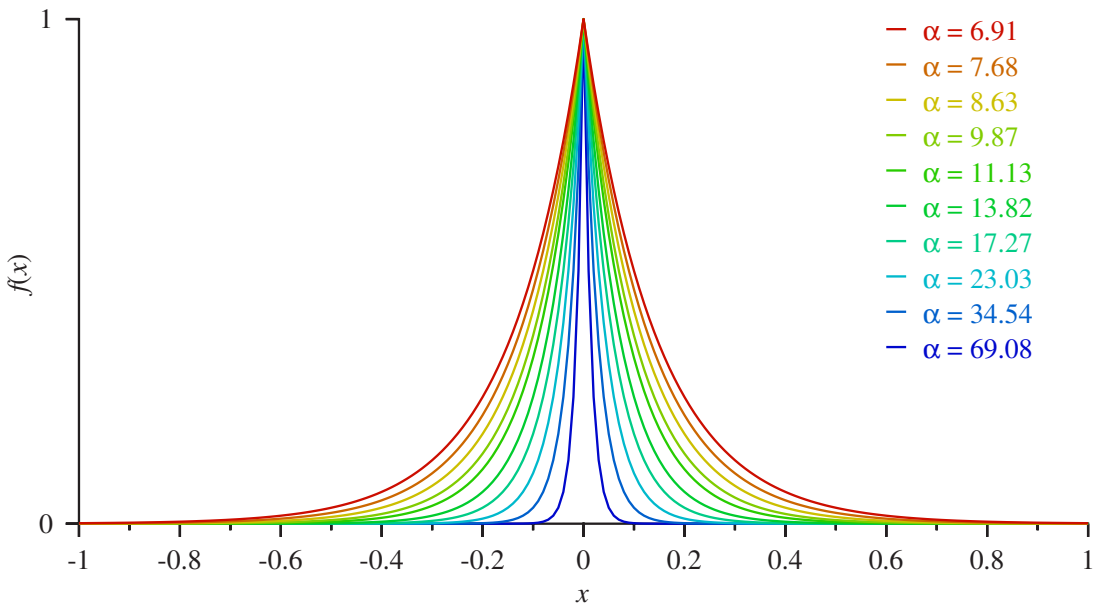



**Abbildung 3.12:** $f(x) = e^{-\alpha|x|}$, $x \in [-1, 1]$

| $\varepsilon = 0{,}01$ | | $\varepsilon = 0{,}001$ | |
|---|---|---|---|
| $r$ | $\alpha$ | $r$ | $\alpha$ |
| 0,05 | 92,10 | 0,05 | 138,16 |
| 0,10 | 46,05 | 0,10 | 69,08 |
| 0,15 | 30,70 | 0,15 | 46,05 |
| 0,20 | 23,03 | 0,20 | 34,54 |
| 0,25 | 18,42 | 0,25 | 27,63 |
| 0,30 | 15,35 | 0,30 | 23,03 |
| 0,35 | 13,16 | 0,35 | 19,74 |
| 0,40 | 11,51 | 0,40 | 17,27 |
| 0,45 | 10,23 | 0,45 | 15,35 |
| 0,50 | 9,21 | 0,50 | 13,82 |
| 0,55 | 8,37 | 0,55 | 12,56 |
| 0,60 | 7,68 | 0,60 | 11,51 |
| 0,65 | 7,08 | 0,65 | 10,63 |
| 0,70 | 6,58 | 0,70 | 9,87 |
| 0,75 | 6,14 | 0,75 | 9,21 |
| 0,80 | 5,76 | 0,80 | 8,63 |
| 0,85 | 5,42 | 0,85 | 8,13 |
| 0,90 | 5,12 | 0,90 | 7,68 |
| 0,95 | 4,85 | 0,95 | 7,27 |
| 1,00 | 4,61 | 1,00 | 6,91 |

**Tabelle 3.5:** Parameter $\alpha$ für Radius $r$ und $\varepsilon$ in $e^{-\alpha|x|}$

verschwindend gering ist. Gibt man einen kleinen Wert $0 < \varepsilon \leq 1$ und einen Radius $0 < r \leq 1$ vor und setzt $\alpha := -(\ln \varepsilon)/r^2$, so ist für $|x| \geq r$ der Funktionswert stets kleiner oder gleich $\varepsilon$ (Tabelle 3.4). An den Stellen $-r$ und $r$ ist er gleich $\varepsilon$. Je größer $\alpha$ ist, desto stärker fällt die Kurve nahe 0 und desto stärker ist die Trennung der Punkte (Abbildung 3.11).

#### 3.13.1.5 Exponentialfunktion des Betrages

Es sei $\alpha > 0$. Die Abbildung

$$f\colon \mathbb{R} \to (0,1],\ x \mapsto e^{-\alpha|x|}$$

nimmt den Wert 0 nicht an. Gibt man einen kleinen Wert $0 < \varepsilon \leq 1$ und einen Radius $0 < r \leq 1$ vor und setzt $\alpha := -(\ln \varepsilon)/|r|$, so ist für $|x| > r$ der Funktionswert stets kleiner $\varepsilon$ (Tabelle 3.5). An den Stellen $-r$ und $r$ ist er gleich $\varepsilon$. Die Funktion fällt nahe 0 deutlich stärker als $x \mapsto \exp(-\alpha x^2)$. Je größer $\alpha$ ist, desto kleiner ist der Radius (Abbildung 3.12).

### 3.13.2 Mountain-Clustering

Es seien $x_k$, $k = 1, 2, \ldots, n$, Datenpunkte. Für jeden möglichen Clusterschwerpunkt $v_i$, $i = 1, 2, \ldots, r$, wird der Wert einer Mountain-Function berechnet. Je höher der Wert ist, desto größer ist das Potential, ein Clusterschwerpunkt zu sein.

Es sei $C$ eine unscharfe Relation für die Konstruktion und $D$ eine unscharfe Relation für die Destruktion. Im ersten Schritt wird für jeden möglichen Schwerpunkt $v_i$ ein Wert (Dichte, Höhe)

$$F_1(v_i) := \sum_{k=1}^{n} R(v_i, x_k)$$

und in allen weiteren Schritten $2 \leq t \leq c \leq n$ jeweils ein Wert

$$F_t(v_i) := \max\left\{0,\ F_{t-1}(v_i) - F_{t-1}(v_{i_{t-1}}) D(v_i, v_{i_{t-1}})\right\}$$

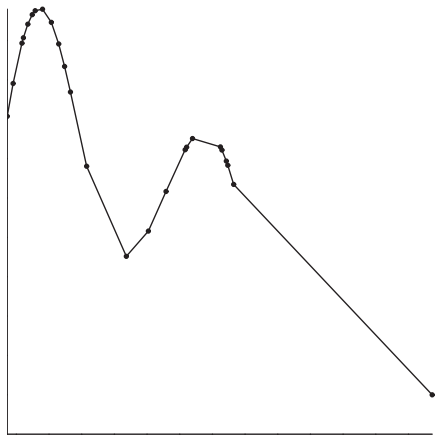
berechnet. In jedem Schritt wird der Punkt mit dem größten Wert zum Prototypen bestimmt. Gibt es mehr als ein Maximum, wird eines beliebig gewählt. Der Index des Prototyps im $t$-ten Schritt sei $i_t$. Der Prototyp ist somit $x_{i_t}$ und der zugehörige Wert $F_t(v_{i_t}) = \max_{1 \leq i \leq r} F_t(v_i)$. Alle weiteren Clusterprototypen werden durch eine iterative Zerstörung der Gipfel gewonnen. Die Höhen der

Punkte werden entsprechend ihrer Zugehörigkeiten zum zuletzt bestimmten Prototyp verringert. Die Höhen von Punkten nahe des Prototyps werden dabei stark reduziert und die des Prototyps wird 0. Der nächste Clusterschwerpunkt ist der mit der größten verbleibenden Höhe. Die Höhen nehmen in jedem Schritt ab und somit auch die Güte der Prototypen. Die Iteration wird beendet, wenn die gewünschte Anzahl $c$ an Prototypen berechnet ist oder das Verhältnis der Maxima des ersten und des letzten Schrittes $t$ einen vorgegebenen Schwellenwert $\delta < 1$ unterschreitet, $F_t(v_{i_t})/F_1(v_{i_1}) < \delta$, spätestens jedoch, wenn alle Höhen 0 sind. Der Schwellenwert bestimmt die Mindesthöhe eines Berges im Verhältnis zum größten Berg. Ist $\delta$ zu groß, werden zu wenig Punkte als Prototypen akzeptiert. Ist $\delta$ zu klein, werden zu viele Prototypen erzeugt.

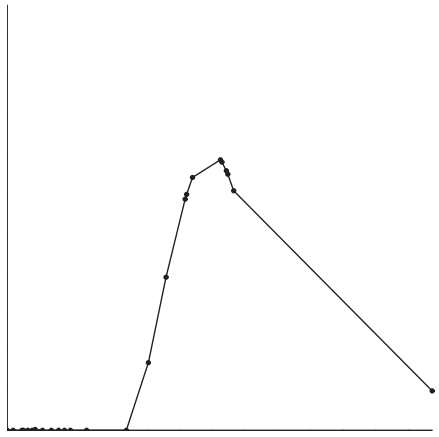
Abbildungen 3.13, 3.14 und 3.15 zeigen an einem einfachen Beispiel mit 24 Punkten auf der reellen Zahlengerade, wie die Höhen schrittweise verringert werden. Zur Berechnung der Höhen werden die Abstände zwischen den Punkten in das Einheitsintervall skaliert, weil es einfacher ist diese Skalierung auszuführen als für jeden Datensatz eine eigene angepaßte Mountain-Function zu definieren. Die die Radien bestimmenden Parameter beziehen sich deshalb auf das Einheitsintervall und nicht auf die eigentlichen Abstände. Die Abszissenwerte der Punkte in den Abbildungen sind die Datenpunkte und die Ordinatenwerte deren Höhen. Die Verbindungsstrecken der Punkte dienen lediglich der Veranschaulichung.

Idealerweise wird ein Berg so zerstört, daß die anderen davon unberührt bleiben. Ist der Radius jedoch zu groß gewählt, verlieren auch die umliegenden Berge an Höhe (Abbildung 3.15). Ist er zu klein, werden durch die nur teilweise Zerstörung eines Berges neue künstliche Berge erzeugt (Abbildung 3.14). Dadurch entstehen fälschlicherweise Prototypen dort, wo keine Clusterschwerpunkte sind. Ist der Funktionsgraph zu spitz, entsteht bei der Destruktion an der Stelle, wo vorher der Gipfel war, ein sich ebenso spitz nach unten verjüngendes Tal. Die verminderten Höhen der benachbarten Punkte sind verhältnismäßig groß, obwohl sie, genauso wie die Höhe des Gipfels, nahe 0 liegen sollten (Abbildung 3.14).

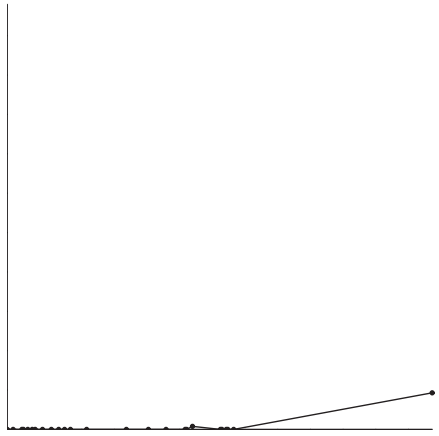
Um eng benachbarte Clusterprototypen zu vermeiden, wird für die Destruktion ein etwas größerer Radius verwendet. So wird nicht nur die Höhe des Prototyps auf 0 gesetzt, sondern auch die der nächstgelegenen Punkte. Gleiches kann auch erreicht werden durch Verwendung verschiedener unscharfer Relationen für die Konstruktion und die Destruktion. Bei der Trapezoidal-Fuzzy-Number



1. Schritt: $\alpha = 73{,}68$

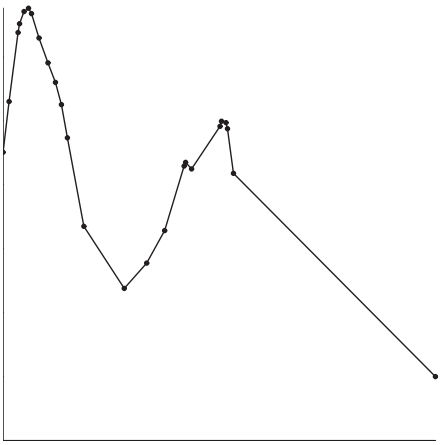


2. Schritt: $\alpha = 18{,}42$

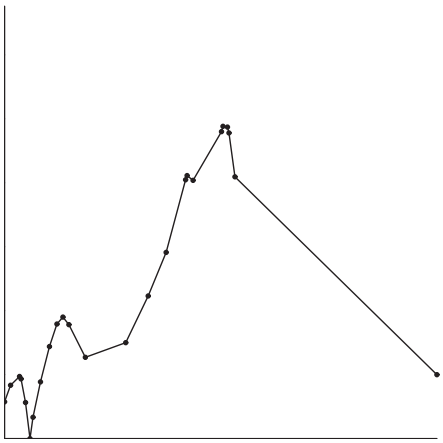


3. Schritt: $\alpha = 18{,}42$

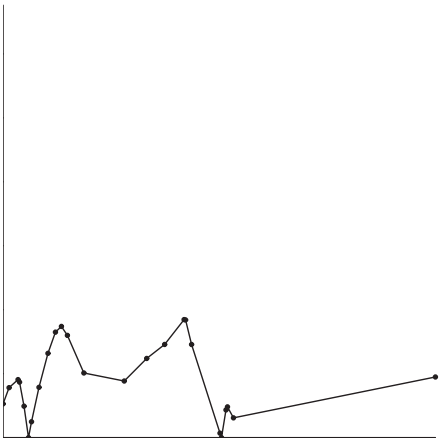
**Abbildung 3.13:** Höhen beim Subtractive-Clustering mit $\exp(-\alpha x^2)$



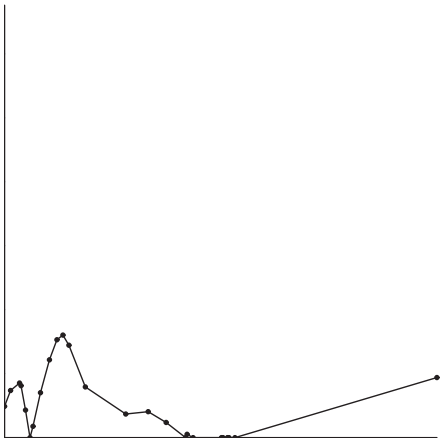
1. Schritt: $\alpha = 18{,}42$



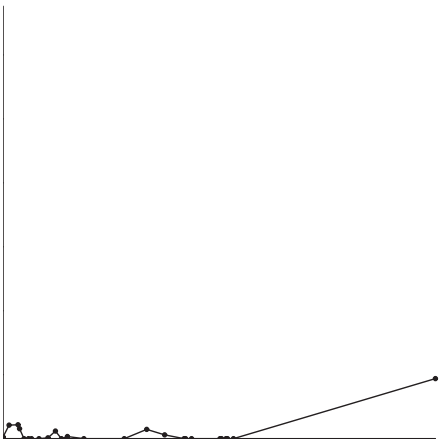
2. Schritt: $\alpha = 9{,}21$



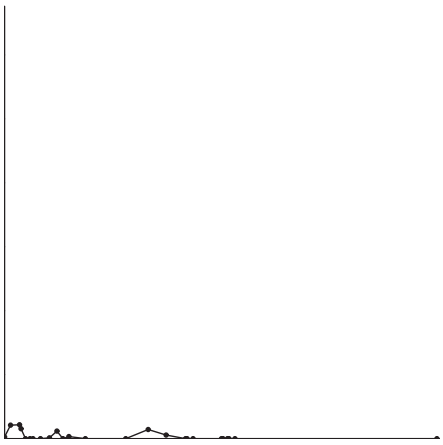
3. Schritt: $\alpha = 9{,}21$



4. Schritt: $\alpha = 9{,}21$

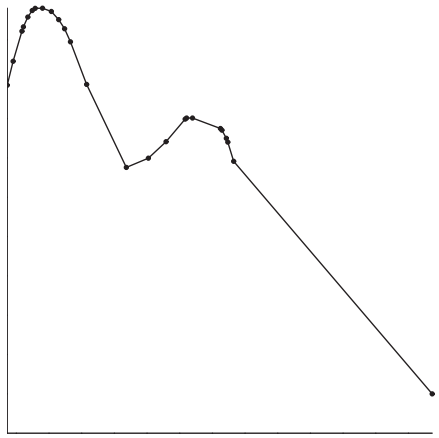


5. Schritt: $\alpha = 9{,}21$

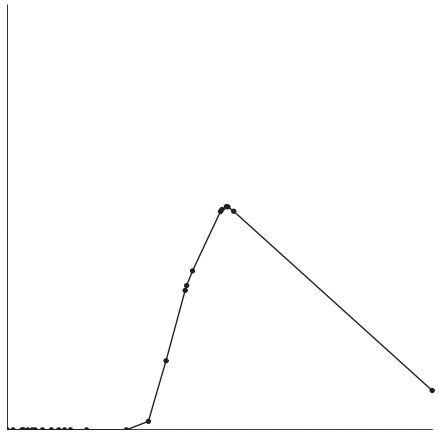


6. Schritt: $\alpha = 9{,}21$

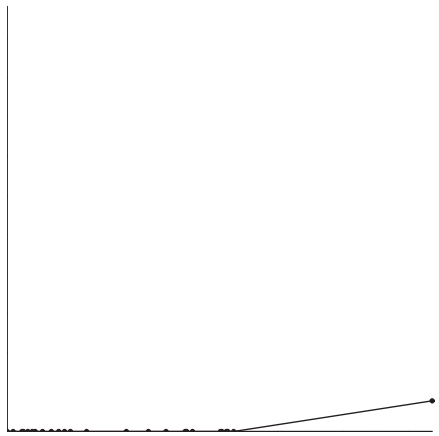
**Abbildung 3.14:** Höhen beim Subtractive-Clustering mit $\exp(-\alpha|x|)$



1. Schritt: Triangular-Fuzzy-Number mit $b = 0$ und $c = 0{,}25$



2. Schritt: Trapezoidal-Fuzzy-Number mit $b = 0$, $c = 0{,}1$ und $d = 0{,}5$



3. Schritt: Trapezoidal-Fuzzy-Number mit $b = 0$, $c = 0{,}1$ und $d = 0{,}5$

**Abbildung 3.15:** Höhen beim Subtractive-Clustering mit Triangular-Fuzzy-Number und Trapezoidal-Fuzzy-Number

etwa kann nicht nur die Breite des Fußes des Berges, sondern auch die Breite eines flachen Gipfels angegeben werden.

Der höchste Punkt ist immer ein Clusterschwerpunkt. Das trifft auf den zweithöchsten im allgemeinen schon nicht mehr zu. Er kann ein weiterer Punkt am Rücken des Berges, dessen Gipfel der höchste Punkt ist, sein oder die Spitze eines anderen Berges. Ersterer ist kein Prototyp, letzterer schon.

Der die unscharfe Relation bestimmende Radius gibt den Radius der Cluster an. Je kleiner der Radius ist, desto weniger Einfluß haben die umliegenden Punkte und desto schärfer ist die Trennung der Gruppen. Jedoch ist bei schärferer Trennung auch die Anzahl der lokalen Maxima und somit die der Prototypen größer.

#### 3.13.2.1 Yager

YAGER und FILEV (1994a,b) versehen den Datenraum mit einem nicht notwendigerweise gleichmäßigen Raster. Die Auflösung des Rasters bestimmt die Auflösung des Modells. Je feiner das Raster ist, desto genauer, aber aufwendiger ist die Berechnung. Jeder Rasterpunkt ist ein potentieller Clusterschwerpunkt.

#### 3.13.2.2 Chiu

CHIU (1994), dessen Variante Subtractive-Clustering genannt wird, betrachtet jeden Datenpunkt als einen potentiellen Clusterschwerpunkt. Nur für die Datenpunkte werden Dichten berechnet. Die Dimension des Datenraumes ist irrelevant. Das ist ein erheblicher Vorteil gegenüber YAGERs Methode, bei der sich die Zahl der Rasterpunkte mit jeder weiteren Dimension vervielfacht.

## 3.14 Ein selbststabilisierender k-Nearest-Neighbors-Algorithmus

Es sei $F$ eine Mountain-Function (Abschnitt 3.13.1), die jedem Punkt einen als Höhe interpretierten nichtnegativen reellen Wert zuordnet. Ein Modus ist hier ein lokales Maximum der Funktion $F$ (Abschnitt 2.13.6). Eine Menge heißt *unimodal,* wenn sie genau einen Modus hat. Eine Menge heißt *unimodal zerlegt,* wenn jede Menge der Partition unimodal ist. Die unimodale Zerlegung ist feiner als die angestrebte Partition, weil zu jedem sich auch noch so gering von den umliegenden Punkten unterscheidenden lokalen Maximum von $F$ eine

Menge der Zerlegung gehört. Ein jeder Cluster der finalen Partition setzt sich aus nebeneinanderliegenden Mengen der unimodalen Zerlegung zusammen. Vgl. hierzu auch GITMAN und LEVINE (1970).

Das hier vorgestellte agglomerative Verfahren entspringt der Idee, die durch unimodale Zerlegung gewonnenen Gruppen zu Clustern zusammenzufassen, die nicht nur Teilen von Bergen, sondern ganzen Bergen im Sinne des Mountain-Clusterings entsprechen. Das Verfahren wird mit einer unimodalen Partition initialisiert, die auch die feinste Partition sein kann. Der die Mountain-Function bestimmende Radius bestimmt auch die Granularität dieses Clusterverfahrens, hat also nicht unerheblichen Einfluß auf die Zahl der Cluster in der finalen Partition.

Der Algorithmus ermittelt die Anzahl der Cluster und die Prototypen nach einem lokalen Stabilitätskriterium für Cluster. Ein Cluster heißt *stabil,* wenn es nicht zu anderen Gruppen hin tendiert. Nicht stabile Gruppen zieht es zu anderen hin, denen sie sich hinzufügen. Die Vereinigungstendenz eines Clusters wird aus dessen Nachbarschaft abgeleitet. Mit Worten des Mountain-Clusterings läßt sich sagen, daß Teile eines Berges solchen Teilen hinzugefügt werden, die dem Gipfel näher sind.

### 3.14.1 Nachbarschaften

Das Verfahren basiert nicht auf Distanzen, sondern auf $k$-Nachbarschaften (Abschnitt 3.12.3). Die nächsten Nachbarn können von jedem Element eines Clusters oder nur vom Prototyp ermittelt werden. Die Nachbarschaft eines Clusters wird beschrieben durch die Anzahl der nächsten Nachbarn des Prototyps oder der Elemente in allen fremden Clustern und im Cluster selbst.

#### 3.14.1.1 Prototypen

Der Prototyp eines Clusters ist der Punkt des Clusters mit dem größten Gewicht. In jedem Cluster wird nur die $k$-Nachbarschaft des Prototyps bestimmt. Die $k$-Nachbarschaft des Clusters ist die des Prototyps. Der Prototyp gilt nicht als Nachbar von sich selbst. Ein Cluster muß also mindestens zwei Elemente enthalten, damit wenigstens ein nächster Nachbar des Prototyps im Cluster selbst liegen kann. Die minimale Anzahl von nächsten Nachbarn sei $k_{\min} := 1$. Zu einem Prototyp können bei einer Gesamtheit von $n$ Elementen nicht mehr als $k_{\max} := n-1$ Nachbarn gehören. Die Anzahl $k$ der nächsten Nachbarn liegt folglich

immer zwischen einschließlich 1 und $n-1$. Je größer $k$ ist, desto mehr nächste Nachbarn können in einem anderen Cluster liegen. Übersteigt $k$ die Clustergröße, zählen zwangsweise Elemente anderer Cluster zu den nächsten Nachbarn des Prototyps.

Ein Prototyp ist von vielen verhältnismäßig nahe bei ihm liegenden Punkten umgeben. Diese Konstellation ist die Ursache für sein hohes Gewicht. Der Prototyp eines stabilen Clusters liegt etwa in dessen Mitte. Liegt ein Prototyp am Rand eines Clusters, liegen auch viele seiner nächsten Nachbarn außerhalb. Die Anzahl der externen nächsten Nachbarn gibt den Grad der Vereinigungstendenz mit den jeweiligen anderen Clustern an. Liegt ein Prototyp am Rand eines Clusters, so ist er im übertragenen Sinne möglicherweise kein Gipfel eines Berges, sondern lediglich der höchste Punkt an einem Abschnitt des Bergrückens. Nach einer Agglomeration zweier Cluster ist immer der höher bewertete Prototyp der Prototyp des neuen Clusters.

#### 3.14.1.2 Elemente

Die andere Möglichkeit ist, daß die $k$-Nachbarschaft eines Clusters aus den $k$-Nachbarschaften der Elemente gebildet wird. Für jedes Element eines Clusters wird die $k$-Nachbarschaft bestimmt und gezählt, wie viele Elemente daraus in den anderen und dem eigenen Cluster liegen. Elemente im Inneren eines Clusters haben fast ausschließlich Elemente des Clusters selbst als Nachbarn. Nur bei Randpunkten liegt eine größere Zahl nächster Nachbarn in anderen Clustern. Die Nachbarschaft eines Clusters besteht nun aus allen Nachbarn seiner Elemente zusammen. Dabei treten einzelne Elemente mehrfach auf, weil sie gemeinsame Nachbarn verschiedener Clusterelemente sind. Die Nachbarn können somit nach der Häufigkeit ihres Auftretens gewichtet werden.

### 3.14.2 Agglomerationskriterien

Die Cluster werden nach einem lokalen Stabilitätskriterium vereinigt. Dieses ermittelt als interne Stoppbedingung die Cluster und damit auch die Clusteranzahl der finalen Partition. Das Kriterium verhindert, daß nicht alle möglichen Agglomerationen durchgeführt werden. Ein Cluster gilt als stabil, wenn mehr nächste Nachbarn in ihm selbst liegen als außerhalb. Ihm wird keine Vereinigungstendenz zugesprochen. Von den instabilen Clustern werden in jedem

Schritt nur diejenigen agglomeriert, die das Kriterium lokal optimieren. Einmal vorgenommene Vereinigungen werden nicht mehr getrennt.

Die Vereinigungstendenz ist eine Funktion der Anzahlen der nächsten Nachbarn im Cluster selbst und in fremden Clustern. Anstelle der Anzahlen können auch die Gewichte der Elemente verwendet werden. Jedoch wurden mit den Anzahlen experimentell die besten Ergebnisse erzielt. Auch die Verknüpfung mehrerer der obigen Kriterien zu einem komplexen Kriterium führt nur bei einigen konkreten Datensätzen zu besseren Partitionen. Das Verfahren wird im allgemeinen aber instabiler und kann bei nur geringen Veränderungen in den Daten zu wesentlich anderen Clustern führen.

#### 3.14.2.1 Asymmetrisch

Aufgrund der Nachbarschaft eines einzelnen Clusters wird entschieden, ob und welchem anderen Cluster er hinzugefügt wird. Ein Cluster wird einem anderen hinzugefügt, wenn in diesem mehr nächste Nachbarn liegen als im Cluster selbst. Dieses Agglomerationskriterium ist asymmetrisch, da nur die Nachbarschaft des instabilen Clusters betrachtet wird. Die Vereinigungstendenz des aufnehmenden Clusters ist dabei irrelevant. Ist das Kriterium erfüllt, werden in jedem Schritt genau zwei Cluster vereinigt. Unter den instabilen Clustern, wird derjenige gewählt, von dem am meisten nächste Nachbarn in einem einzelnen fremden Cluster liegen und am wenigsten in allen anderen. Dieser Cluster ist nicht nur ausgesprochen instabil, sondern tendiert auch besonders stark zu nur einem anderen Cluster. Eine Zuordnung ist eindeutig, wenn alle nächsten Nachbarn, die außerhalb liegen, Elemente eines einzelnen Clusters sind. Erst wenn keine eindeutigen Zuordnungen mehr möglich sind, werden die Cluster betrachtet, deren nächste Nachbarn sich auf mehrere andere Cluster aufteilen. Mit gröber werdender Partition können auch diese Cluster leichter zugeordnet werden, weil einige der Cluster, auf den die nächsten Nachbarn ursprünglich aufgeteilt waren, mittlerweile zusammengelegt worden sind.

#### 3.14.2.2 Symmetrisch

Es werden stets die beiden Cluster vereinigt, die die meisten nächsten Nachbarn gemeinsam haben.

### 3.14.3 Größe der Nachbarschaften

Es werden Varianten unterschieden, die für alle Cluster dieselbe Zahl nächster Nachbarn extern bestimmen und solche, die für jeden Cluster eine eigene Zahl nächster Nachbarn intern aus der Clustergröße ableiten.

#### 3.14.3.1 Extern

Die Vorgabe einer Zahl $k_t$ nächster Nachbarn im $t$-ten Schritt für alle Prototypen führt zu etwa gleich großen Clustern. Für die Beziehung von Clusteranzahl $c$ und Anzahl nächster Nachbarn gilt $k_t \to n-1 \Leftrightarrow c \to 1$ und $k_t \to 1 \Leftrightarrow c \to n/2$. Die Partition und somit die Anzahl der Cluster wird bestimmt durch zwei Schleifen. Die äußere Schleife bestimmt die Anzahl der nächsten Nachbarn und die innere Schleife stabilisiert die Partition bezüglich dieser. Erst wenn in der inneren Schleife keine Agglomeration mehr durchgeführt werden kann, wird in der äußeren Schleife eine neue Zahl nächster Nachbarn bestimmt bezüglich derer die Partition wiederum innerhalb der inneren Schleife gefestigt wird. Die innere Schleife führt sukzessive alle Vereinigungen durch, die das Kriterium für die Nachbarschaftsgröße erfüllen. Erlaubt auch die neue Nachbaranzahl keine Vereinigung mehr, gilt jeder Cluster als in sich stabil und eine finale Partition ist bestimmt.

Die äußere Schleife bestimmt die Anzahl der nächsten Nachbarn aus einer vorgegebenen globalen unteren Schranke $k_{\min}$ und einer globalen oberen Schranke $k_{\max}$, einer die Nachbarschaft vergrößernden globalen Konstanten $k_{\text{extra}}$, dem $k_{t-1}$ aus dem vorigen Schleifendurchlauf $t-1$ und der Größe der Partition $c$. Experimentell haben sich die Werte $k_{\min} := 1$, $k_{\max} := n-1$ und $k_{\text{extra}} := 1$ als geeignet erwiesen. Die Größe der Nachbarschaft $k_t$ muß echt größer sein als im vorigen Schritt $t-1$, $k_t := \max(k_{t-1}+1, k_t)$. Die äußere Schleife vergrößert $k_t$ solange, bis die innere Schleife keine Agglomeration mehr vornimmt.

(NN1) Setze die Anzahl der nächsten Nachbarn im $t$-ten Schritt auf $k_t := n/c-1$. Dabei muß $k_t$ größer sein als $k_{\min}$, $k_t := \max(k_{\min}, k_t)$. $k_t$ muß größer sein als im vorigen Schritt $t-1$, $k_t := \max(k_{t-1}+1, k_t)$. $k_t$ darf nicht größer sein als $n-1$, $k_t := \min(k_t, n-1)$.

(NN2) Berechne die Werte, die die Nachbarschaft eines jeden Prototyps beschreiben.

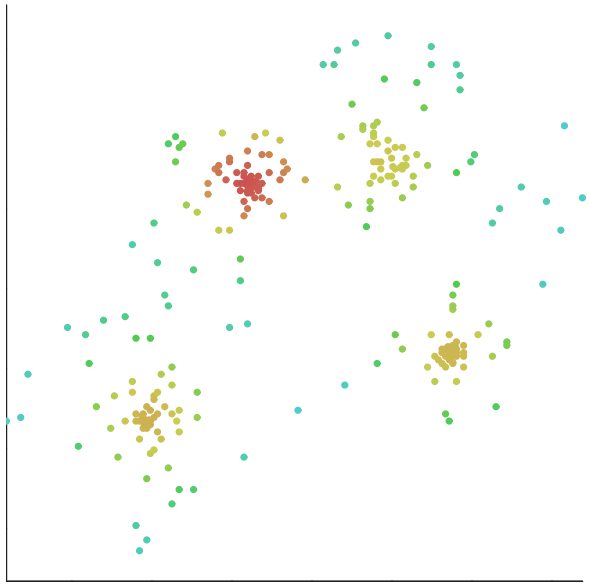


**Abbildung 3.16:** Dichten (als Farbtöne) nach der Z-Funktion mit $a = 0$ und Radius $c = 0{,}2$ von der größten Distanz

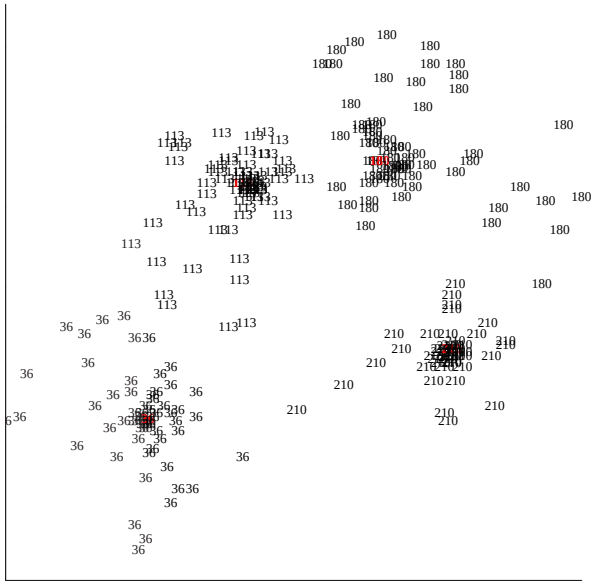


**Abbildung 3.17:** Partition der $n = 254$ Punkte in $c = 4$ Cluster in 250 Schritten mit den Dichten aus Abbildung 3.16

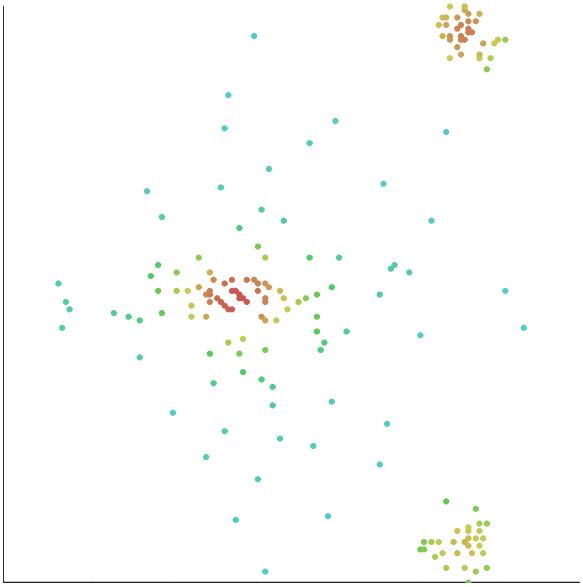


**Abbildung 3.18:** Dichten nach der Z-Funktion mit $a = 0$ und Radius $c = 0{,}15$ von der größten Distanz

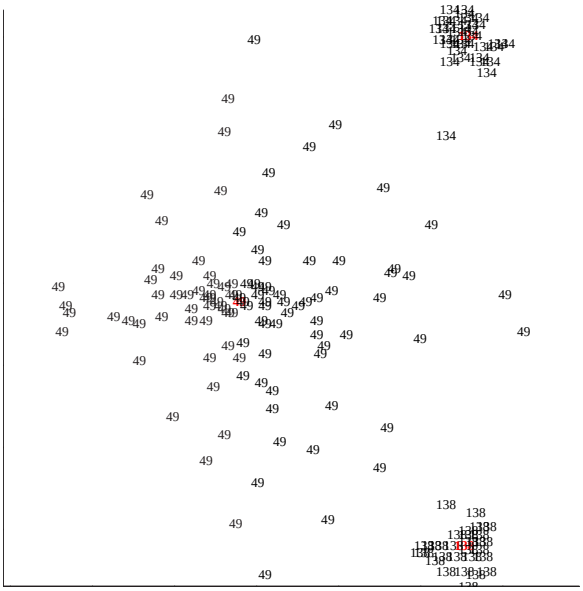


**Abbildung 3.19:** Partition der $n = 174$ Punkte in $c = 3$ Cluster in 171 Schritten mit den Dichten aus Abbildung 3.18

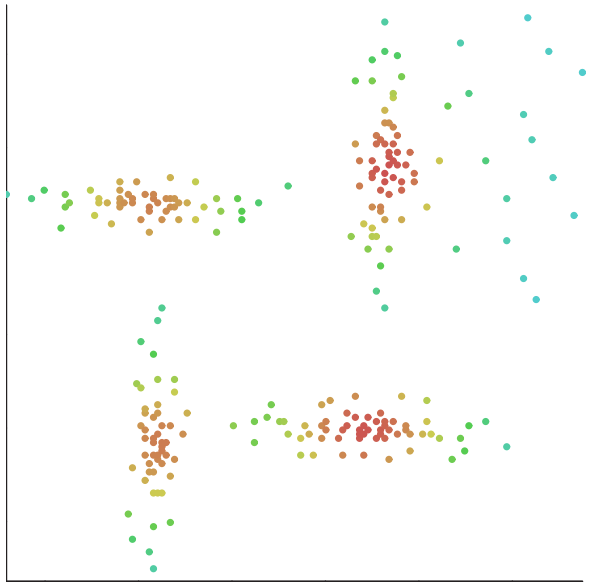


**Abbildung 3.20:** Dichten nach der Z-Funktion mit $a = 0$ und Radius $c = 0{,}2$ von der größten Distanz

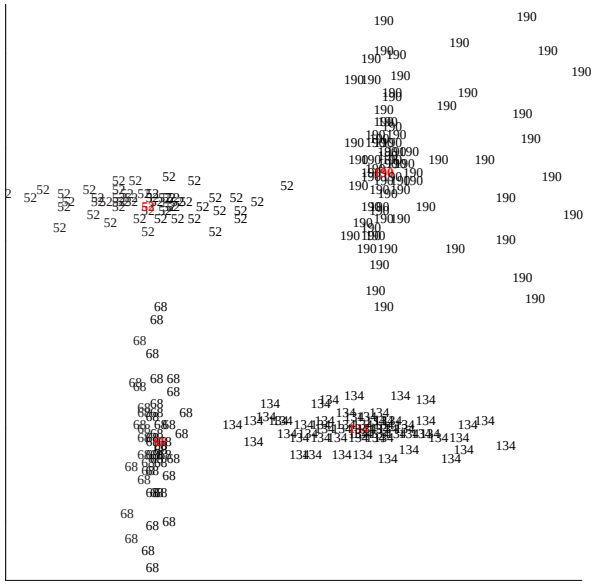


**Abbildung 3.21:** Partition der $n = 241$ Punkte in $c = 4$ Cluster in 237 Schritten mit den Dichten aus Abbildung 3.20

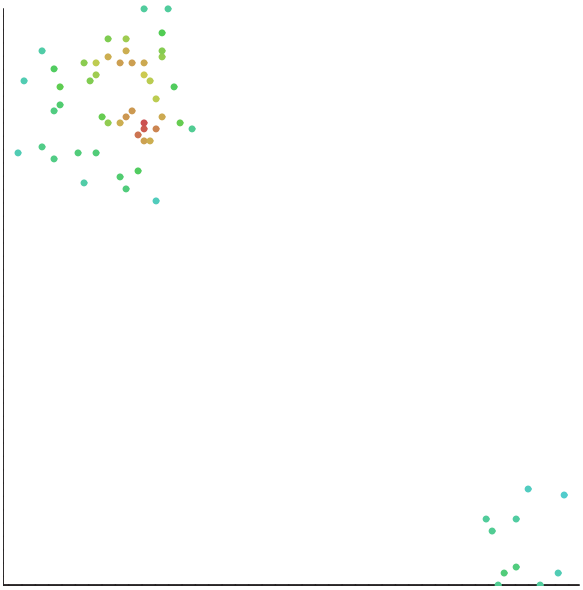


**Abbildung 3.22:** Dichten nach der Z-Funktion mit $a = 0$ und Radius $c = 0{,}07$ von der größten Distanz

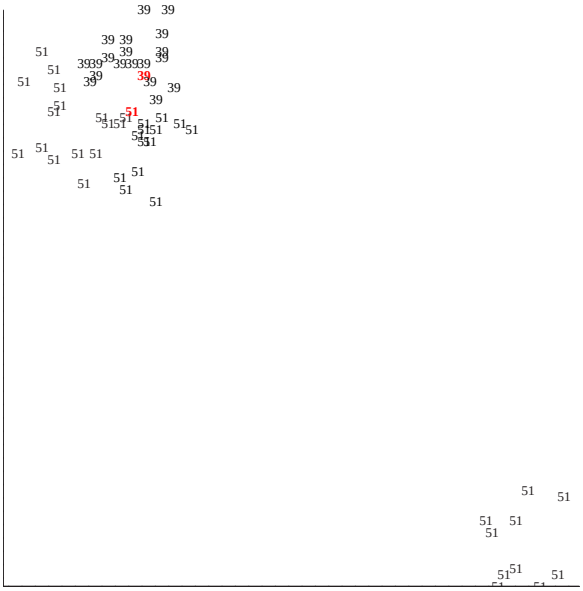


**Abbildung 3.23:** Fehlklassifikation wegen zu großem Unterschied der Clustergrößen. Partition der $n = 60$ Punkte in $c = 2$ Cluster in 58 Schritten mit den Dichten aus Abbildung 3.22

(NN3) Vereinige die Cluster, die das Kriterium am besten erfüllen und fahre mit $t + 1$ im Schritt (NN2) fort. Springe zu (NN1), wenn kein Cluster dem Kriterium mehr genügt. Beende, wenn für dieses $k_t$ kein Cluster das Kriterium je erfüllt hat.

Die Mindestgröße eines Clusters von zwei führt dazu, daß auch vereinzelte, weit außerhalb liegende Punkte einem Cluster zugewiesen werden. Cluster, die nicht groß genug sind, um die extern vorgegebene Anzahl nächster Nachbarn selbst stellen zu können, werden zuerst vereinigt. Ist $k_t$ zu groß, können weiter entfernt liegende, dichte Cluster für eine Vereinigung eher erwogen werden als direkt benachbarte kleine oder lose Gruppen, allein weil sie mehr nächste Nachbarn enthalten. Die Anzahl $k_t$ der nächsten Nachbarn darf deshalb nicht viel größer sein als der Cluster.

Abbildung 3.22 zeigt ein Beispiel, bei dem die kleinere Gruppe (unten rechtes) zu klein ist, um mit dem gegebenen Radius deutliche Dichten ausbilden zu können. Hingegen ist die größere Gruppe (oben links) so groß, daß mit dem Radius zwei Schwerpunkte erzeugt werden. Auch durch die Wahl eines anderen Radius können keine geeigneten Dichten errechnet werden. Durch eine Vergrößerung des Radius kann zwar erreicht werden, daß die größere Gruppe nur noch einen Schwerpunkt besitzt, die Dichten der kleinen Gruppe hingegen gleichen sich dadurch untereinander stärker an. Ein kleinerer Radius bewirkt, daß die kleine Gruppe einen deutlichen Schwerpunkt erhält. Die große Gruppe wird dadurch aber in noch mehr Gruppen unterteilt. Aus diesem Unterschied der Gruppengrößen resultiert letztlich die Fehlklassifikation aus Abbildung 3.23.

Zur Berechnung der Dichten (Abbildungen 3.16, 3.18, 3.20, 3.22) werden die Distanzen in das Einheitsintervall skaliert. Die Partitionen in den Abbildungen 3.17, 3.19, 3.21 und 3.23 sind Ergebnisse von asymmetrischen Agglomerationen von Clustern nach der extern bestimmten Anzahl der nächsten Nachbarn der Prototypen. Die initiale Partition ist jeweils die feinste Zerlegung. Die Cluster bestehen aus einem dichten Kern. Alle um diesen Kern verstreuten Punkte werden diesem zugewiesen. Auch weit entfernte Punkte (Ausreißer) werden einem Cluster zugewiesen. So werden die in Abbildung 3.21 oben rechts liegenden Punkte noch Cluster 190 zugeordnet, weil sie dessen Kern am nächsten sind.

Das Verfahren ermittelt die Clusterschwerpunkte zuverlässig. Die Zuordnung von Punkten, die zwischen zwei Clustern liegen, ist allerdings nicht verläßlich. In Abbildung 3.19 wird ein einzelner Punkt unterhalb von Cluster 134 noch

diesem zugeordnet, obwohl er wie alle anderen weit um den Kern von Cluster 49 verstreuten Punkte noch zu diesem hätte gezählt werden können. In Abbildungen 3.17 werden die Punkte, die rechts der Mitte zwischen Cluster 180 und 210 liegen, Cluster 180 zugeordnet, obwohl einige davon dem Kern von Cluster 210 näher sind.

#### 3.14.3.2 Intern

Die äußere Schleife entfällt, weil die Anzahl der nächsten Nachbarn für jeden einzelnen Cluster gesondert bestimmt wird. Bei dieser Variante neigt der größte Cluster dazu, nach und nach alle kleineren zu assimilieren.

## 3.15 Bewertung von Clusterlösungen

Alle Clusterverfahren nehmen Einteilungen aufgrund eines vorgegebenen, mehr oder weniger vagen Modells vor, ohne dabei die tatsächliche Struktur der Daten zu berücksichtigen. Stimmt die durch das Modell definierte Struktur mit der den Daten eigenen Struktur nicht überein, gibt die gefundene Lösung die Daten verzerrt wieder. Deshalb darf das Ergebnis eines einzelnen Clusterprozesses nie unkritisch akzeptiert werden. Es sollten stets nicht nur eine, sondern mehrere clusteranalytische Verfahren auf denselben Datensatz angewendet werden. Unter allen so erhaltenen Lösungen ist die zu ermitteln, die der Struktur der Daten im Hinblick auf das Untersuchungsziel am nächsten ist.

Die Überprüfung der Stabilität erfolgt meist informell und qualitativ. Unter allen berechneten Ergebnissen wird diejenige Lösung Gegenstand der Interpretation, die am häufigsten aufgetreten ist oder mit allen anderen Lösungen die größte Ähnlichkeit besitzt. Die so gefundene Lösung ist relativ stabil, nicht aber unbedingt immer auch für das Untersuchungsziel am brauchbarsten. Bei einer großen Anzahl von Klassifikationsobjekten kann die Stabilität dadurch überprüft werden, daß der Datensatz zufällig in einige wenige Teile zerlegt wird und diese getrennt voneinander untersucht werden. Stimmen die Ergebnisse weitgehend überein, kann die Lösung als stabil angesehen werden, andernfalls sind die gefundenen Strukturen nur schwach ausgeprägt. Eine weitere Möglichkeit der Stabilitätsprüfung ist die Wiederholung der Analyse an einer um einige Objekte verminderten oder erweiterten Population. Hierbei ist vor allem die Änderung in der Clusteranzahl interessant. Analog ist es möglich, Klassifikationsvariablen zu entfernen oder hinzuzufügen. Handelt es sich dabei um relativ unbedeutende

Variablen, sollte die Klassifikation weitgehend unverändert bleiben. Bei relevanten Variablen ist eine wesentliche Veränderung in der Clusterzusammensetzung zu erwarten.

Zur Bewertung von Clusterlösungen (*cluster evaluation, cluster assessment, cluster validity*) sind objektive und quantitative Methoden wünschenswert, die angeben wie gut die durch einen Algorithmus gebildeten Cluster den Datensatz unterteilen. Die Kriterien dafür müssen unabhängig von den Algorithmen sein. Es werden drei Bewertungsarten unterschieden. *Externe* Kriterien vergleichen eine Clusterstruktur mit bereits vorhandenen Informationen. *Interne* Kriterien bewerten die Übereinstimmung der gefundenen Strukturen mit den Daten. *Relative* Kriterien vergleichen mehrere Ergebnisse miteinander. Ferner unterteilt man die Kriterien in *globale* und *lokale* Gütemaße. Globale Gütemaße geben die Güte einer ganzen Clustereinteilung durch einen einzigen reellen Wert an. Lokale Gütemaße hingegen bewerten einzelne Cluster. Im folgenden werden einige Gütekriterien kurz vorgestellt.

## 3.15.1 Hierarchien

### 3.15.1.1 Kophänetischer Korrelationskoeffizient

Der kophänetische Korrelationskoeffizient (*cophenetic correlation coefficient*) mißt den Grad der Verzerrung der zugrundeliegenden Daten durch deren hierarchische Klassifikation. Dazu wird die Matrix der empirischen Distanzen $D := (d_{ij})_{1\leq i,j\leq n}$ mit der Matrix der transformierten kophänetischen oder ultrametrischen Distanzen $C := (c_{ij})_{1\leq i,j\leq n}$ verglichen. Da beide Matrizen symmetrisch sind, genügt es, nur die $m := (n^2 - n)/2$ Elemente oberhalb oder unterhalb der Hauptdiagonalen zu betrachten. Der empirische Korrelationskoeffizient

$$r(C,D) := \frac{\left(\frac{1}{m}\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}c_{ij}\right) - \bar{d}\,\bar{c}}{\sqrt{\left(\frac{1}{m}\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij}^2\right) - \bar{d}^2}\,\sqrt{\left(\frac{1}{m}\sum_{i=2}^{n}\sum_{j=1}^{i-1} c_{ij}^2\right) - \bar{c}^2}} \in [-1,1]$$

mit

$$\bar{d} = \frac{1}{m}\sum_{i=2}^{n}\sum_{j=1}^{i-1} d_{ij} \quad \text{und} \quad \bar{c} = \frac{1}{m}\sum_{i=2}^{n}\sum_{j=1}^{i-1} c_{ij}$$

heißt kophänetischer Korrelationskoeffizient. Hierbei ist die Kovarianz rechnerisch wie folgt vereinfacht:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\
&=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-\bar{x}y_i+\bar{x}\,\bar{y})\\
&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{y}\,\frac{1}{n}\sum_{i=1}^{n}x_i-\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}y_i+\frac{1}{n}\sum_{i=1}^{n}\bar{x}\,\bar{y}\\
&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{y}\,\bar{x}-\bar{x}\,\bar{y}+\frac{1}{n}\,n\,\bar{x}\,\bar{y}\\
&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\,\bar{y}.
\end{aligned}$$

Gilt $x_i = y_i$ für alle $i = 1, \ldots, n$, ergibt sich die Formel für die Varianz. Der Nenner ist also die Kovarianz der Matrizen $D$ und $C$. Im Zähler steht das Produkt der Standardabweichungen dieser beiden Matrizen. Es werden jeweils die

$$\sum_{1\le j<i\le n} 1 = \sum_{i=2}^{n}\sum_{j=1}^{i-1} 1 = \sum_{i=2}^{n}(i-1) = \sum_{i=1}^{n-1} i = \frac{(n-1)n}{2} = m$$

Komponenten der Matrizen aufaddiert, für deren Indizes $1 \le j < i \le n$ gilt.

Je geringer das Ausmaß der Verzerrung zwischen den ursprünglichen, empirischen und den transformierten, theoretischen Ähnlichkeitsbeziehungen ist, desto näher liegt der kophänetische Korrelationskoeffizient bei 1. Für verschiedene Transformationen durch hierarchische Clusterungen ist diejenige mit dem größten Wert am wenigsten verfälscht. Es wird also die Clusterung gesucht, für die der kophänetische Korrelationskoeffizient das Maximum annimmt,

$$\max_{C,D} r(C, D).$$

Da bei monotonen Clusterverfahren die kophänetische Matrix eine Ultrametrikmatrix ist, setzt die vollständige Übereinstimmung voraus, daß alle Tripel aus der Ähnlichkeitsmatrix die Ultrametrikungleichung erfüllen. Das ist selten der Fall, da die Ähnlichkeitsmatrix eine größere Anzahl gleicher Werte enthalten müßte. Der kophänetische Korrelationskoeffizient ist ein internes und globales

Bewertungskriterium. Vgl. BACHER (1994, S. 246), ECKES und ROSSBACH (1980, S. 97), GORDON (1981, S. 124 f.), JAIN und DUBES (1988, S. 166 f.).

#### 3.15.1.2 Stoppregel

Alle hierarchisch-agglomerativen Verfahren reduzieren letztlich die Daten zu einem einzigen Cluster, welches alle Objekte enthält. Um eine Partition der Daten zu erhalten, muß der Prozeß sukzessiver Agglomerationen vorher gestoppt werden. Die Stoppregel von MOJENA (1977) ermittelt die entsprechende Clusteranzahl und somit die gesuchte Partition aus der Hierarchie eines monotonen Clusterverfahrens. In der monoton wachsenden Folge der Agglomerationsdistanzen wird nach einem signifikanten Anstieg gesucht. Dieser entspricht einem überproportionalen Heterogenitätszuwachs und zeigt die beste Partition an. Sei hierzu $\mathcal{U} := (\mathcal{U}_0, \mathcal{U}_1, \ldots, \mathcal{U}_{n-1})$ die Hierarchie einer monotonen hierarchisch-agglomerativen Clusterung von $n$ Objekten. Ferner sei $\alpha_j$, $j = 1, 2, \ldots, n-1$, die Stufe, an der durch Agglomeration zweier Gruppen der Partition $\mathcal{U}_{j-1}$ aus dieser die Partition $\mathcal{U}_j$ entsteht. Es wird die Gruppenanzahl gewählt, die zur ersten Stufe im Dendrogramm gehört, für die

$$\alpha_{j+1} > \bar{\alpha} + k\, s_\alpha \tag{3.17}$$

gilt, wobei $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ die Fusionslevel der Stufen mit $n-1, n-2, \ldots, 1$ Cluster sind,

$$\bar{\alpha} = \frac{1}{n-1} \sum_{j=1}^{n-1} \alpha_j$$

das arithmetische Mittel der Fusionsniveaus,

$$s_\alpha = \sqrt{\frac{1}{n-2} \sum_{j=1}^{n-1} (\alpha_j - \bar{\alpha})^2}$$

deren Standardabweichung und $k$ eine Konstante, die das Heterogenitätsniveau steuert, ist. Es wird der kleinste Index $j \in \{1, 2, \ldots, n-2\}$ gesucht, für den die Bedingung (3.17) gilt. Die Gruppenanzahl ist dann $n - j$, also die Gruppenanzahl der vorigen Partition. Die feinste und die gröbste Partition werden nicht betrachtet. Ist die Bedingung für keinen Index erfüllt, kann die Stoppregel nicht angewendet werden. Je größer $k$ ist, desto heterogener sind die Cluster der Partition. Für steigendes $k$ nimmt also die Anzahl der Cluster

ab. Mojena (1977, S. 361) erzielt die besten Ergebnisse mit $2{,}75 \leq k \leq 3{,}50$. In Everitt (1993, S. 74) wird $k := 1{,}25$ vorschlagen. Diese Werte mögen dort geeignet gewesen sein, allgemein gültig sind sie sicher nicht. Sie müssen für jeden Datensatz einzeln bestimmt werden. Lediglich, wenn ein Datensatz einem anderen Datensatz, für den dieser Parameter bereits bestimmt wurde, strukturell gleicht, kann dieser Wert wiederverwendet werden. Dieses Kriterium ist intern und lokal. Vgl. Everitt (1993, S. 73 f.) und Deimer (1986, S. 81).

### 3.15.2 Partitionen

#### 3.15.2.1 Calinski und Harabasz

Calinski und Harabsz (1974) geben eine auf den Summen der quadrierten Fehlern basierenden Güte

$$CH(U) := \frac{\mathrm{Spur}(B)}{\mathrm{Spur}(W)} \cdot \frac{n-c}{c-1}$$

für scharfe Partitionen an, wobei $B$ die Interclustervarianzmatrix, $W$ die Intraclustervarianzmatrix und $c$ die Anzahl der Gruppen der Partition $U$ ist. Hierbei ist

$$\begin{aligned}
\mathrm{Spur}(W) &= \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}\, d^2(x_k, v_i) \qquad \text{nach (3.10)} \\
&= \sum_{i=1}^{c} \sum_{j=1}^{n_i} d^2(x_{i_j}, v_i)
\end{aligned}$$

und

$$\begin{aligned}
\mathrm{Spur}(B) &= \mathrm{Spur}\left( \sum_{i=1}^{c} n_i (v_i - v)(v_i - v)^T \right) \\
&= \sum_{i=1}^{c} n_i \ \mathrm{Spur}\left( (v_i - v)(v_i - v)^T \right) \\
&= \sum_{i=1}^{c} n_i \sum_{r=1}^{p} (v_{ir} - v_r)^2 \\
&= \sum_{i=1}^{c} n_i\, d^2(v_i, v)
\end{aligned}$$

mit den Mittelwertvektoren aus (3.4) und (3.3) und $n_i := \sum_{i=1}^{k} u_{ik}$. Werden statt dieser die Clusterprototypen verwendet, ist der Güteindex auch auf unscharfe

Partitionen anwendbar. Für steigende Clusteranzahl $c$ nimmt die Spur von $B$ zu und die Spur von $W$ ab.

Für die Folge der Partitionen einer Hierarchie spiegeln sich gleichmäßig im Raum verteilte Punkte in einem gleichmäßigen, glatten Verlauf des Graphen der Gütewerte wider. Sind die Punkte hingegen in $c$ natürlichen Clustern gruppiert, geht der Schritt von $c-1$ zu $c$ Clustern mit einem beträchtlichen Zuwachs einher. Aus einer Hierarchie wird die Partition mit derjenigen Clusteranzahl $c$ gewählt, für die der Index ein absolutes oder lokales Maximum annimmt, oder zumindest bei der Erhöhung auf $c$ Cluster einen vergleichsweise großen Zuwachs hat. Der Index ist global und relativ.

#### 3.15.2.2 Separationsindex

Der Separationsindex von Dunn (1974, S. 34 f.) bewertet nur scharfe Zerlegungen. Er identifiziert kompakte, gut getrennte Cluster und ist definiert durch

$$D(U) := \frac{\min\limits_{1\leq i\leq c} \min\limits_{\substack{1\leq j\leq c \\ j\neq i}} \delta(u_i, u_j)}{\max\limits_{1\leq k\leq c} \Delta(u_k)}$$

wobei

$$\delta(u_i, u_j) := \min_{\substack{x\in u_i \\ y\in u_j}} d(x, y)$$

die Distanz der Cluster (*set distance*) $i$ und $j$, $1 \leq i, j \leq c$, und

$$\Delta(u_i) := \max_{x,y\in u_i} d(x, y)$$

der Durchmesser (*diameter*) des $i$-ten Clusters und $d$ die euklidische Metrik ist. Der Durchmesser eines Clusters ist der Abstand der beiden entferntesten Punkte des Clusters und damit ein Maß für dessen Ausdehnung. Der Clusterabstand ist der Abstand derjenigen beiden Punkte aus unterschiedlichen Clustern, die am dichtesten beieinander liegen. Im Zähler von $D$ steht die kleinste aller Distanzen zwischen verschiedenen Clustern und im Nenner der größte Durchmesser aller Cluster der Partition. Der Zähler mißt die externe Heterogenität und der Nenner die interne Homogenität über alle Cluster der Partition. Bei steigender Clusteranzahl fällt der Zähler monoton, da die Anzahl derjenigen Distanzen, von denen die kleinste bestimmt wird, größer wird. Auch der Nenner fällt monoton für steigende Clusteranzahl, da die Cluster und somit ihre Durchmesser kleiner

werden. Große Werte von $D$ zeigen isolierte und kompakte Cluster an. Die beste Partition der Datenmenge löst

$$\max_{2 \leq c \leq n-1} \left\{ \max_{U \in \Omega_c} D(U) \right\},$$

wobei $\Omega_c$ die Menge der $c$-Partitionen ist. Für eine Zerlegung in kompakte und klar getrennte Cluster muß $D > 1$ gelten. Die zu bewertende Partition muß aus mindestens zwei und darf höchstens aus $n-1$ Clustern bestehen. Dunns Separationsindex bezieht die Daten in die Berechnung mit ein. Er ist global und relativ. Vgl. Bezdek (1998, S. F6.3:5) und Xie und Beni (1991, S. 842).

#### 3.15.2.3 Partitionskoeffizient

Der Partitionskoeffizient

$$PC(U) := \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^2$$

von Bezdek (1981, S. 100) mißt die Unschärfe einer unscharfen Partition. Es gilt

$$\frac{1}{c} \leq PC(U) \leq 1,$$

$$PC(U) = 1 \quad \Leftrightarrow \quad U \in \{0,1\}^{c \times n}, \tag{3.18}$$

$$PC(U) = \frac{1}{c} \quad \Leftrightarrow \quad \underset{1 \leq i \leq c}{\forall} \; \underset{1 \leq k \leq n}{\forall} \; u_{ik} = \frac{1}{c}. \tag{3.19}$$

Eine scharfe Einteilung hat nach (3.18) den Maximalwert 1. Bestehen die Daten aus disjunkten Strukturen, das heißt aus scharfen Clustern, sollten unscharfe Clusteralgorithmen an $PC$ gemessen relativ scharfe Partitionen erzeugen. Ist die Einteilung maximal unscharf, haben also alle Elemente den Zugehörigkeitswert $1/c$ zu allen Clustern, ergibt sich der Minimalwert $1/c$ (3.19). Die schärfste Partition wird durch

$$\max_{2 \leq c \leq n-1} \left\{ \max_{U \in \Omega_c} PC(U) \right\}$$

bestimmt. Der Partitionskoeffizient ist ein globaler und relativer Güteindex. Seine möglichen Ausprägungen hängen von der Anzahl $c$ der Cluster ab. Die Unschärfen von Partitionen unterschiedlicher Clusteranzahlen können nur verglichen werden,

wenn die Werteintervalle $[1/c, 1]$ jeweils in das Einheitsintervall skaliert werden. Siehe auch BEZDEK (1998, S. F6.3:5).

#### 3.15.2.4 Klassifikationsentropie

Die Klassifikationsentropie (*classification entropy, partition entropy*) von BEZDEK ist an SHANNONs Informationstheorie angelehnt. Sie ist dem Partitionskoeffizienten sehr ähnlich und definiert durch

$$PE(U) := -\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\log_a(u_{ik})$$

mit $u_{ik}\log_a(u_{ik}) := 0 :\Leftrightarrow u_{ik} = 0$ und einer logarithmischen Basis $a \in (1, \infty)$. Gewöhnlich wird der natürliche Logarithmus mit der EULERschen Zahl $e$ als Basis verwendet. Es ist

$$0 \leq PE(U) \leq \log_a(c),$$

$$PE(U) = 0 \quad \Leftrightarrow \quad U \text{ scharf}, \tag{3.20}$$

$$PE(U) = \log_a(c) \quad \Leftrightarrow \quad \underset{1\leq i\leq c}{\forall}\ \underset{1\leq k\leq n}{\forall}\ u_{ik} = \frac{1}{c}. \tag{3.21}$$

Für eine scharfe Einteilung, in der alle Informationen bekannt sind, ist nach (3.20) die Entropie 0. Bei gleichen Zugehörigkeiten – die Einteilung enthält keine Information – ist nach (3.21) die Entropie maximal. Die optimale Zerlegung löst

$$\min_{2\leq c\leq n-1}\left\{\min_{U\in\Omega_c} PE(U)\right\}.$$

Die Klassifikationsentropie ist ein globaler und relativer Index. Partitionskoeffizient und Klassifikationsentropie nutzen weder die Daten selbst, noch die Prototypen. Siehe BEZDEK (1981, S. 111 ff.), BEZDEK (1998, S. F6.3:5) und HÖPPNER et al. (1997, S. 159).

#### 3.15.2.5 Xie und Beni

Der unscharfe Fehler eines Datenpunktes $x_k$ bezüglich eines Clusterprototyps $v_i$ ist $\delta_{ik} := u_{ik}\|x_k - v_i\|$ mit der euklidischen Norm $\|\cdot\|$. Die Summe der quadrierten unscharfen Fehler aller Objekte der $i$-ten Klasse,

$$\sigma_i := \sum_{1\leq k\leq n} \delta_{ik}^2,$$

ist die Variation der $i$-ten Klasse. Die totale Variation ist die Summe der Variationen aller Klassen,

$$\sigma := \sum_{1\leq i\leq c} \sigma_i = \sum_{1\leq i\leq c} \sum_{1\leq k\leq n} \delta_{ik}^2.$$

Die Variationen hängen von den Daten $X$ selbst, der unscharfen Partition und den Prototypen ab. Xie und Beni (1991) definieren die Kompaktheit einer unscharfen $c$-Zerlegung als $\pi := \sigma/n$. Je kompakter die Klassen sind, desto kleiner ist $\pi$. Die Trennung einer unscharfen $c$-Partition definieren sie als das Quadrat der kleinsten Distanz zwischen den Clustermittelpunkten, also als $s := \min_{1\leq i,j\leq c} \|v_i - v_j\|^2$. Je stärker die Cluster getrennt sind, desto größer ist $s$. Der Güteindex mißt das Verhältnis von Kompaktheit und Trennung,

$$XB(U,V) := \frac{\pi}{s} = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2\|x_k - v_i\|^2}{n \min_{1\leq i,j\leq c} \|v_i - v_j\|^2}.$$

Gute Cluster minimieren diesen Index durch Kompaktheit (kleine Zähler) und weite Trennung (große Nenner)

$$\min_{2\leq c\leq n-1} \left\{ \min_{U,V} XB(U,V) \right\}.$$

Er ist ein globales und relatives Kriterium.

#### 3.15.2.6 Fuzzy-Hypervolumen

Gath und Geva (1989) geben zu ihrem Clusteralgorithmus (Abschnitt 3.11.2.3) gleich drei Kriterien zur Bewertung von Partitionen, deren Clusterformen durch Kovarianzmatrizen beschrieben werden, an. Mit diesen können deshalb auch die Ergebnisse der Gustafson-Kessel-Clusterungen bewertet werden. Alle drei

Kriterien sind nur anwendbar, wenn die unscharfen Kovarianzmatrizen regulär sind, da die Determinanten singulärer Matrizen 0 sind. Das Fuzzy-Hypervolumen

$$FHV(U,V) := \sum_{i=1}^{c} \sqrt{\det(C_i)}$$

erfaßt die Summe aller Clustergrößen über die unscharfen Kovarianzmatrizen $C_i$, welche die Formen und Größen der Cluster beschreiben. Das Minimum

$$\min_{2 \le c \le n-1} \left\{ \min_{U,V} FHV(U,V) \right\}$$

wird für kleine Cluster erreicht. Dieses Kriterium ist global und relativ. Vgl. hierzu im Ganzen Höppner et al. (1997, S. 163).

#### 3.15.2.7 Mittlere Partitionsdichte

Die mittlere Partitionsdichte (*average partition density*) nach Gath und Geva (1989) ist

$$APD(U,V) := \frac{1}{c} \sum_{i=1}^{c} \frac{s_i}{\sqrt{\det(C_{fi})}},$$

wobei

$$\underset{1 \le i \le c}{\forall} \quad s_i := \sum_{k=1}^{n} u_{ik} \, \mathbf{1}_{\left\{ j \in \{1,\ldots,n\} \mid (x_j - v_i)^T C_{fi}^{-1} (x_j - v_i) < 1 \right\}}(k) \tag{3.22}$$

die unscharfe Anzahl derjenigen Datenpunkte im $i$-ten Hyperellipsoid ist, deren Abstand zum Prototypen kleiner als die Standardabweichung ist. $\mathbf{1}$ bezeichne hierbei die Indikatorfunktion. In die Berechnung gehen also nur die Elemente ein, die dicht am Clusterzentrum liegen. Die hier betrachtete physikalische Dichte ist definiert als Anzahl der Punkte im Cluster pro Clustervolumen. Die mittlere Partitionsdichte spiegelt die Existenz einzelner dichter Cluster wider. Partitionen mit dichten und losen Clustern werden wegen der dichten Teilstrukturen als gut bewertet. Solche Cluster zeichnen sich durch deutliche Punkthäufungen aus. Dieses Kriterium ist zu maximieren, also

$$\max_{2 \le c \le n-1} \left\{ \max_{U,V} APD(U,V) \right\}.$$

Es ist global und relativ.

deterministisch:

| | | |
|---|---|---|
| Calinski und Harabasz | $CH(U) = \frac{\mathrm{Spur}(B)}{\mathrm{Spur}(W)} \frac{n-c}{c-1}$ | lokales max |
| Dunn | $D(U) = \frac{\min\limits_{1\le i\le c} \min\limits_{\substack{1\le j\le c \\ j\ne i}} \min\limits_{\substack{x\in u_i \\ y\in u_j}} d(x,y)}{\max\limits_{1\le k\le c} \max\limits_{x,y\in u_i} d(x,y)}$ | max |

possibilistisch:

| | | |
|---|---|---|
| Partitionskoeffizient | $PC(U) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^2$ | max |
| Klassifikationsentropie | $PE(U) = -\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log_a(u_{ik})$ | min |
| Xie und Beni | $XB(U,V) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^2 \lVert x_k - v_i \rVert^2}{n \min\limits_{1\le i,j\le c} \lVert v_i - v_j \rVert^2}$ | min |
| Fuzzy-Hypervolumen | $FHV(U,V) = \sum_{i=1}^{c} \sqrt{\det(C_{fi})}$ | min |
| Mittlere Partitionsdichte | $APD(U,V) = \frac{1}{c} \sum_{i=1}^{c} \frac{\sum_{k=1}^{n} u_{ik} \, \mathbf{1}_{\{j\in\{1,\ldots,n\} \mid (x_j-v_i)^T C_{fi}^{-1} (x_j-v_i) < 1\}}(k)}{\sqrt{\det(C_{fi})}}$ | max |
| Partitionsdichte | $PD(U,V) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \, \mathbf{1}_{\{j\in\{1,\ldots,n\} \mid (x_j-v_i)^T C_{fi}^{-1} (x_j-v_i) < 1\}}(k)}{\sum_{i=1}^{c} \sqrt{\det(C_{fi})}}$ | max |

**Tabelle 3.6:** Globale und relative Klassifikationsindizes

#### 3.15.2.8 Partitionsdichte

Die Partitionsdichte nach GATH und GEVA (1989)

$$PD(U,V) := \frac{\sum_{i=1}^{c} s_i}{FHV(U,V)}$$

drückt die generelle Dichte der Partition aus. $s_i$ berechnet sich nach (3.22). Die Partitionsdichte ist zu maximieren

$$\max_{2 \leq c \leq n-1} \left\{ \max_{U,V} PD(U,V) \right\}.$$

Dieser Index ist global und relativ.

Für die letzten drei Indizes gilt, daß eine gute Clustereinteilung im Sinne dieser Gütemaße aus klar getrennten Clustern mit minimalem Gesamtvolumen und einer Häufung der Daten nahe der Clusterzentren besteht.

# Kapitel 4

# Termassoziationen

## 4.1 Erhebung der Daten

Die Datengrundlage bildet eine Sammlung von unkorrigierten Artikeln mit allen orthographischen und grammatikalischen Fehlern, Setzfehlern und nicht zum Inhalt gehörenden zusätzlichen Angaben von Ort, Agentur, Autor, etc. aus dem Ressort „Wirtschaft und Umwelt“ der Tageszeitung „taz, die tageszeitung“. Das Korpus ist um ein Vielfaches größer als die Korpora bei früheren Untersuchungen. Wurden vorher nur einige hundert Texte verwendet, so liegt nun eine Sammlung von mehreren zigtausend Texten zugrunde. Betrachtet werden nur die längsten Texte, weil nur in ihnen ausreichend viele Beziehungen zwischen den Wörtern bestehen können (Abschnitt 5.1.2). Die flektierten Wortformen werden getaggt und lemmatisiert, Phrasen und Sätze dazu geparst. Die zu untersuchenden sprachlichen Ausdrücke werden kurz *Terme* genannt. Die Terme der Untersuchung sind die Lemmata der Wortformen in den Texten.

In dem Korpus aus $n$ pragmatisch homogenen Texten werden die Häufigkeiten von $m$ Termen in allen Artikeln gezählt. Dies ist eine Vollerhebung vom Umfang $m$. Die Terme sind die Merkmalsträger, die Texte die Merkmale und die Häufigkeiten der Terme in den Texten die Merkmalsausprägungen. Die Texte seien beliebig, aber fest durch die ersten $n$ natürlichen Zahlen benannt. Besteht der $j$-te Text, $j = 1, 2, \ldots, n$, aus $l_j$ Termen, so kann ein Term darin wenigstens 0 und höchstens $l_j$ mal vorkommen. Es sei $\Omega$ die Grundgesamtheit der Terme. Die $j$-te statistische Variable $H_j$ nimmt Ausprägungen auf der Ordinalskala $S_j := \{0, 1, \ldots, l_j\} \subset \mathbb{N} \cup \{0\}$ an, $H_j : \Omega \to S_j$, $\omega \mapsto H_j(\omega)$. Die Häufigkeiten eines Terms werden in jedem Text gezählt. Die einzelnen Variablen faßt man zu einer vektorwertigen Variablen $H := (H_1, H_2, \ldots, H_m) \colon \Omega \to S_1 \times S_2 \times \cdots \times S_m$

zusammen. Die Terme seien beliebig, aber fest durch die ersten $m$ natürlichen Zahlen indiziert. Für jeden Term $\omega_i$ sei $h_{ij} := H_j(\omega_i)$ die $j$-te Beobachtung und $h_i := (h_{i1}, h_{i2}, \ldots, h_{im})$ der Vektor aller Beobachtungen. Im Ganzen erhält man die Beobachtungsmatrix $(h_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$ der Häufigkeiten aller Terme in allen Texten. Sie wird Term-Dokument-Matrix genannt, weil jede Zeile für einen Term und jede Spalte für einen Text steht. Die meisten Komponenten der Term-Dokument-Matrix sind 0, weil ein Term gewöhnlich in nur wenigen Texten vorkommt. Die von 0 verschiedenen Werte treten so vereinzelt auf, daß das Datenbild der Matrix beinahe einfarbig in dem 0 zugeordneten Farbton erscheint. Der Datensatz ist multivariat. Dies ist eine Querschnittserhebung, da die Betrachtung synchron ist.

## 4.2 Stoppwörter

Eine Stoppwortliste ist eine Liste von Termen, die von der Untersuchung ausgeschlossen werden. Das sind solche sprachliche Ausdrücke, die keine lexikalische Bedeutung tragen oder nicht häufig genug vorkommen, um deren Verwendungsregularitäten sinnvoll untersuchen zu können.

### 4.2.1 Wortklassen

Sprachliche Elemente, die primär grammatische anstelle von lexikalischer Bedeutung tragen und vor allem syntaktische Funktionen erfüllen, heißen Funktionswörter (Bussmann, 1990, S. 260). Dazu gehören die Wörter der geschlossenen Klassen Adposition, Artikel, Interjektion, Interpunktion, Konjunktion, Partikel und Pronomen. Hingegen heißen sprachliche Ausdrücke, die eine kontextunabhängige, selbständig lexikalische Bedeutung haben, Bedeutungswörter (Bussmann, 1990, S. 118). Sie werden den offenen Klassen Adjektiv, Adverb, Substantiv und Verb zugeordnet. Nur Bedeutungswörter werden untersucht. Zu den Stoppwörtern zählen alle Funktionswörter, Hilfsverben, Modalverben, Fremdwörter, Nichtwörter, Kompositionserstglieder und Zahlen.

### 4.2.2 Termhäufigkeiten

Die Bedeutung eines sprachlichen Ausdrucks, der nur an wenigen Stellen belegt ist, ist gebrauchssemantisch/quantitativ-linguistisch oft schwer bestimmbar. Deshalb werden lexikalische Einheiten, die sehr selten auftreten, zu den

Stoppwörtern gezählt. Ein Term muß in ausreichend vielen Texten vorkommen, damit eine brauchbare Aussage über dessen Verwendungsregularität getroffen werden kann. Hochfrequente Terme sind für die Untersuchung der Bedeutungen deshalb von besonderem Interesse. Zur Bestimmung der Ähnlichkeit zweier Terme werden deren Häufigkeiten in den Texten, in denen beide oder zumindest einer von beiden vorkommt, verglichen. Die Terme müssen so gewählt werden, daß dieser Vergleich auf einer ausreichend großen Datengrundlage durchgeführt werden kann. Für die Untersuchung sind solche Terme besonders gut geeignet, die im ganzen sehr häufig sind, aber nicht in allen Texten vorkommen. Jene Wörter, deren Auftretenshäufigkeit einen vorgegebenen Schwellenwert über- oder unterschreiten, werden entfernt (Abschnitt 5.1.3).

## 4.3 Textlängennormierung

Da die Texte unterschiedlich lang sind, müssen die Skalen der Vorkommenshäufigkeiten normiert werden. Nur so können über unterschiedlich lange Texte hinweg Ähnlichkeitswerte sinnvoll verglichen und aggregiert werden. Die Häufigkeiten der Terme werden hierzu durch die Länge des jeweiligen Textes dividiert. Jeder Text hat nun die Länge 1. Dadurch erhält ein Term in einem langen Text für dieselbe Termhäufigkeit ein kleineres Gewicht als in einem kürzeren Text. Durch diese komponentenweise Skalierung werden die Termhäufigkeiten in Prozentsätze überführt. Die Anpassung der Textlängen in so früher Verarbeitungsphase erleichtert weitere Berechnungen, weil aus diesen die Problematik unterschiedlicher Textlängen ausgeklammert werden kann. Dies erschwert aber ihre nun etwas abstraktere Interpretation, da nicht mehr mit den Merkmalsausprägungen selbst, sondern nur noch mit aus ihnen abgeleiteten Gewichten operiert wird. Davon unberührt bleibt, daß lange Texte viele verschiedene Terme enthalten und deshalb mehr Terme gemeinsam haben als kurze.

## 4.4 Gewichtung der Häufigkeiten

Durch die Gewichtung werden diejenigen Häufigkeiten betont, die wesentlich erscheinen und solche abgeschwächt, die als unwichtig erachtet werden. Dies geschieht mit der Absicht, deutlicher ausgeprägte Ergebnisse zu erzielen. Das ist zwar durchaus wünschenswert, dennoch werden die Häufigkeiten dadurch

verändert, schlimmsten Falls sogar verfälscht. Deshalb dürfen – wenn überhaupt – nur plausible, dem konkreten Untersuchungsgegenstand und -ziel entsprechende Gewichtungen vorgenommen werden.

Es werden nicht die Häufigkeiten selbst, sondern die durch die Anpassung der Textlängen transformierten Werte gewichtet. Der besseren Verständlichkeit wegen wird im folgenden aber mit den Häufigkeiten argumentiert. Formal gibt es kaum einen Unterschied. Es sei $m$ die Anzahl der Terme und $n$ die Anzahl der Texte. Das Gewicht des $k$-ten Terms, $k = 1, 2, \ldots, m$, im $i$-ten Text, $i = 1, 2, \ldots, n$, sei mit $w_{ki}$ bezeichnet. Der Gewichtsvektor des $k$-ten Terms ist $w_k := (w_{k1}, w_{k2}, \ldots, w_{kn})$.

### 4.4.1 Binäre Gewichte

Im einfachsten Fall wird ein Merkmal einem Term zugeordnet oder nicht. Ein zugehöriges Attribut wird mit 1 gewichtet, ein nicht zugewiesenes mit 0. Die Verwendung binär gewichteter Terme hat den Vorteil, daß die Zuordnung zu den Texten sehr einfach ist. Es ist nicht notwendig über den Grad der Zuordnung zu entscheiden. Jedes Merkmal, das nur marginal relevant erscheint, wird zugeordnet. Es wird nur dann nicht zugewiesen, wenn es eindeutig keine Eigenschaft des Terms ist. Nachteilig ist, daß durch diese einfache binäre Gewichtung viele Terme ununterscheidbar werden, da die ihnen zugewiesenen Gewichtsvektoren gleich sind.

### 4.4.2 Allgemeine Gewichte

In die Gewichtung der Terme gehen meist drei Faktoren ein: die Häufigkeit eines Terms in einem einzelnen Text, die Häufigkeit eines Terms im ganzen Korpus und die Verteilung der Einzelhäufigkeiten über das ganze Korpus. Erstere ist eine lokale Gewichtung, die die relative Bedeutung eines Terms innerhalb eines einzelnen Textes beschreibt und die beiden letzteren globale Gewichtungen zur Betonung der relativen Bedeutung eines Terms im ganzen Korpus. Allgemeiner formuliert ist eine lokale Gewichtung eine Funktion der Einzelhäufigkeiten und somit abhängig sowohl vom Term als auch vom Text. Eine globale Gewichtung ist eine Funktion abhängig entweder vom Term oder vom Text. Die meisten globalen Gewichtungen beziehen sich auf Terme. Das Gewicht eines Terms ergibt sich meist aus einer Verknüpfung lokaler und globaler Gewichte.

#### 4.4.2.1 Gewichtung der Terme

Die Gewichte der Terme werden aus deren Häufigkeiten in den Texten abgeleitet. Es sei $m$ die Anzahl der Terme und $n$ die Anzahl der Texte. Es sei $s_i$, $i = 1, 2, \ldots, n$, die Länge (*size*) des $i$-ten Textes gemessen in Token aller verwendeten $m$ Terme. Die Anzahl der Vorkommen $tf_{ki} \in \{0, 1, 2, \ldots, s_i\}$ des $k$-ten Terms, $k = 1, 2, \ldots, m$, im $i$-ten Text nennt man dessen Termhäufigkeit[1] (*term frequency*). Die Korpushäufigkeit (*collection frequency*) des $k$-ten Terms

$$cf_k := \sum_{i=1}^{n} tf_{ki}$$

gibt an, wie oft er in allen Texten zusammen vorkommt. Das Korpus besteht insgesamt aus

$$\sum_{k=1}^{m} cf_k$$

Termtoken. Die Anzahl der Texte

$$df_k := \sum_{i=1}^{n} \mathbf{1}_{\mathbb{N}}(tf_{ki}) \ \in \{1, 2, \ldots, n\}\,,$$

in denen der Term $k$ vorkommt, nennt man dessen Texthäufigkeit (*document frequency*). Funktionen der Termhäufigkeit $tf_{ki}$ gewichten lokal. Funktionen der Texthäufigkeit $df_k$ oder der Korpushäufigkeit $tf_k$ sind globale Gewichtungen.

#### 4.4.2.2 Inverse Texthäufigkeit

Die inverse Texthäufigkeit (*inverse document frequency*) ist umgekehrt proportional zur Texthäufigkeit. Sie ist umso kleiner, in je mehr Texten der Term vorkommt. In je weniger Texten nämlich ein Term vorkommt, desto stärker ist seine trennende Wirkung für diejenigen Texte, in denen er vorkommt. Die

[1] Die hier verwendete Terminologie und auch einige Formeln sind aus dem Information-Retrieval (Salton et al., 1964; Salton, 1968; Salton et al., 1975, 1976; Salton und Wu, 1980; Salton und McGill, 1983; Salton, 1986b; Salton et al., 1986; Salton, 1986a; Salton und Buckley, 1988; Salton, 1989; Salton et al., 1994) entlehnt. Ziel des Information-Retrievals ist es, möglichst schnell und genau Texte, deren Indexterme einem vorgegebenen Suchvektor von Termen entsprechen, in einer Sammlung von Texten zu identifizieren. Die Bedeutung sprachlicher Ausdrücke wird im Information-Retrieval nicht betrachtet. Zudem sind Texte die zu vergleichenden Untersuchungseinheiten und Terme ihre Merkmale. Im Gegensatz dazu sind hier die Terme die Merkmalsträger, deren semantische Ähnlichkeiten aufgrund von Texten als Merkmale ermittelt werden. Rein formal entspricht dies dem Transponieren der Term-Dokument-Matrix.

inverse Texthäufigkeit ist ein globales Gewicht. Sie wird mit der gewöhnlichen Termhäufigkeit multipliziert. Dem $k$-ten Term im $i$-ten Text wird das aus einem lokalen und einem globalen Gewicht zusammengesetzte Gewicht $w_{ki} := tf_{ki}\, idf_k$ zugewiesen. Terme, die in wenigen Texten häufig vorkommen, aber gleichzeitig im ganzen selten sind, werden dadurch stark gewichtet.

**Beispiele**

Eine einfache inverse Texthäufigkeit für den $k$-ten Term ist der Reziprokwert seiner Texthäufigkeit,

$$idf_k := \frac{1}{df_k} \in \left\{\frac{1}{n}, \frac{1}{n-1}, \ldots, 1\right\} \subset (0, 1]. \tag{4.1}$$

Die klassische inverse Texthäufigkeit bei SALTON und BUCKLEY (1988) ist

$$idf_k := \ln \frac{n}{df_k} \in \left\{0, \ln \frac{n}{n-1}, \ldots, \ln n\right\} \subset [0, \infty).$$

Sie ist im Gegensatz zu (4.1) kein stauchender Faktor für die Termhäufigkeit. Für festes $tf_{ki}$ verändert sich $w_{ki}$ umgekehrt proportional zu $df_k$. Allgemeiner formuliert ist die inverse Texthäufigkeit eine streng monoton fallende Funktion der Texthäufigkeit.

#### 4.4.2.3 Gewichtung der Texte

Der Textdiskriminierungswert gibt den Grad an, zu dem ein Text zur Unterscheidung von Termen beiträgt. Für das gesamte Korpus wird eine positive Dichte $s$ aus den paarweisen Ähnlichkeiten $s_{kl}$ aller Terme aggregiert,

$$s := \sum_{k=1}^{m} \sum_{l=1}^{k-1} s_{kl}.$$

Je größer die Dichte ist, desto näher liegen die Terme beieinander. Es wird angenommen, daß die Terme sich besser trennen lassen, je geringer die Dichte ist. Ein Text trägt somit zur besseren Unterscheidung der Terme bei, wenn sein Weglassen die Dichte vergrößert. Die Dichte wird vor und nach der Projektion auf den um den Text verringerten Raum berechnet. Unterscheiden sich beide Dichten kaum, trennt der Text nur unwesentlich. Es sei $s_{kl}^{(i)}$ die Ähnlichkeit der Terme $k$

und $l$, die sich ergibt, wenn der $i$-te Text ignoriert wird. Es sei $s^{(i)}$ die Dichte über alle $s_{kl}^{(i)}$. Der Diskriminierungswert eines Textes $i$ ist der Quotient

$$d_i := \frac{s^{(i)}}{s}.$$

Ist er größer als 1, trägt der $i$-te Text zur deutlicheren Trennung der Terme bei. Ist er kleiner als 1, hat er geringe Unterscheidungskraft. Er gewichtet den Text global. Der $k$-te Term im $i$-ten Text wird gewichtet mit

$$w_{ki} := tf_{ki}\, d_i.$$

Texte, die nur aus sehr wenigen Termen bestehen, tragen nur wenig zur Unterscheidbarkeit der Terme bei. Sie sind zu selten oder zu spezifisch, um einen angemessenen Anteil an den relevanten Termen zu beschreiben. Texte, die nahezu alle Terme enthalten, sind als Diskriminatoren ebenso wenig geeignet, weil sie zu allgemein sind. Die besten Diskriminatoren sind solche, deren Termhäufigkeiten weder zu niedrig noch zu hoch sind. Sie erlauben es, einzelne bestimmte Terme von allen anderen der Sammlung zu unterscheiden (Salton et al., 1975, S. 619). Für eine größere Anzahl von Texten ist deren Diskriminierung praktisch zu rechenintensiv, weil für jeden Text alle paarweisen Ähnlichkeiten der Terme neu berechnet werden müssen.

## 4.5 Äquivalenz von Termen

Zwei Terme werden äquivalent genannt, wenn sie sich in allen Texten gegenseitig ersetzen lassen, ohne die Bedeutung der Texte zu verändern (Oakes, 1998, S. 121). In den folgenden Analyseschritten werden die Untersuchungseinheiten einzig durch ihre Gewichtsvektoren identifiziert. Verschiedene Untersuchungseinheiten können dieselben Merkmalsausprägungen aufweisen. Daraus darf nicht geschlossen werden, daß diese Untersuchungseinheiten identisch sind. Sie sind lediglich bezüglich der ausgewählten Merkmale ununterscheidbar und werden sich in weiteren Berechnungen stets gleich verhalten. Untersuchungseinheiten mit gleichen Beobachtungsvektoren können zumindest formal problemlos zu einer abstrakten Untersuchungseinheit zusammengefaßt werden.

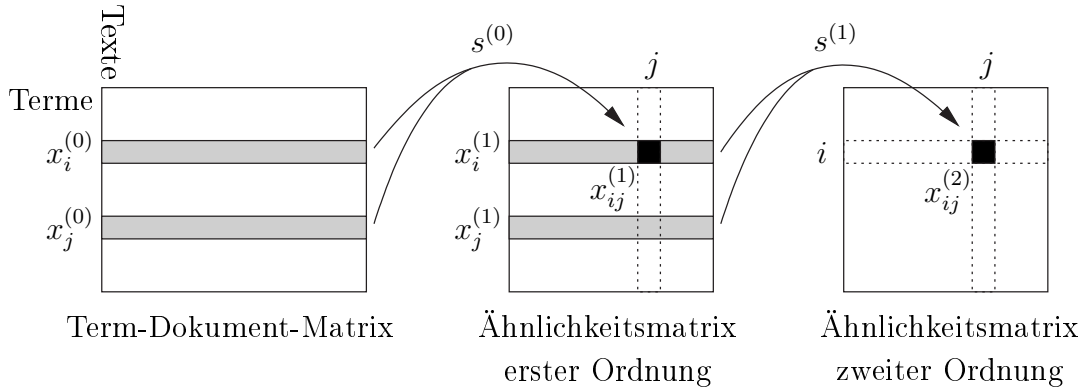



**Abbildung 4.1:** Vektorraummodell

## 4.6 Vektorraummodelle

Es sei $m$ die Anzahl der Terme und $n$ die Anzahl der Texte. Ein Term wird durch einen $n$-elementigen Vektor in einem durch die Texte aufgespannten Vektorraum $V^{(0)}$ beschrieben. Die Texte werden als paarweise orthogonal und somit linear unabhängig modelliert. Sie bilden die $n$-elementige kanonische Basis des Vektorraums. Das Vektorraummodell beschreibt die Ähnlichkeiten (Abschnitt 4.9) erster und höherer Ordnungen der endlich vielen Termvektoren $x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}$ aus $V^{(0)}$. Die unmittelbar aus den Termvektoren berechneten Ähnlichkeiten heißen Ähnlichkeiten erster Ordnung. Sie werden im ersten Schritt paarweise mit einer Ähnlichkeitsfunktion $s^{(0)} : V^{(0)} \times V^{(0)} \to \mathbb{R}$ berechnet. Die Ähnlichkeit erster Ordnung der Vektoren $x_i^{(0)}$ und $x_j^{(0)}$ ist $x_{ij}^{(1)} := s^{(0)}\big(x_i^{(0)}, x_j^{(0)}\big)$. Die aus den Ähnlichkeiten erster Ordnung gebildete $m \times m$-Matrix heißt Ähnlichkeitsmatrix erster Ordnung. Die Ähnlichkeiten erster Ordnung des $i$-ten Vektors zu allen anderen Vektoren bilden den Vektor $x_i^{(1)} := \big(x_{i1}^{(1)}, x_{i2}^{(1)}, \ldots, x_{im}^{(1)}\big)$. Ähnlichkeiten höherer Ordnung sind Ähnlichkeiten von Ähnlichkeitsvektoren (Abbildung 4.1). Die Ähnlichkeiten der $t$-ten Ordnung werden aus denen der nächstniedrigeren $(t-1)$-ten Ordnung berechnet. Die Ähnlichkeit $t$-ter Ordnung der Vektoren $x_i^{(t-1)}$ und $x_j^{(t-1)}$ ist $x_{ij}^{(t)} := s^{(t-1)}\big(x_i^{(t-1)}, x_j^{(t-1)}\big)$. Vgl. Ruge (1995, S. 87 ff.).

Das Vektorraummodell hat den Nachteil, daß die $2n + 1$ hinzutretenden Komponenten bei der Vergrößerung von $n$ auf $n + 1$ Terme im allgemeinen nicht einfach ergänzt werden können. Beim Ergänzen oder Streichen von Termen oder Texten müssen die Matrizen vollständig neu berechnet werden.

## 4.7 Hauptkomponentenanalyse

Ein Ziel der Hauptkomponentenanalyse (*principal component analysis*) ist die Identifizierung neuer, bedeutungsvoller, den Daten inhärenter Merkmale, den sogenannten Hauptkomponenten. Damit verbunden ist die Bestimmung der den Daten eigenen Dimension abweichend von der des Modells. Die Hauptkomponentenanalyse ist eine mathematische Methode zur Bestimmung derjenigen linearen Transformation, die die Streuung der Daten am deutlichsten wiedergibt. Die Daten werden so gedreht, daß die maximalen Variabilitäten auf die Achsen projiziert werden. Die erste Hauptkomponente zeigt in die Richtung, in der die Daten am stärksten streuen. Alle weiteren abstrakten Faktoren werden so gebildet, daß sie möglichst viel der Streuung beschreiben und orthogonal zu allen anderen Achsen sind. Die zweite Hauptkomponente zeigt somit in diejenige Richtung, die orthogonal zur ersten ist und in der die verbleibende Streuung am stärksten ist. Die dritte Hauptkomponente enthält die größte Streuung orthogonal zur ersten und zweiten Achse. Die letzte Achse steht senkrecht zu allen anderen und weist die geringste Streuung auf. Auf diese Weise wird eine orthogonale Basis im Sinne der Methode der kleinsten Quadrate (*least squares method*) des Vektorraums der Daten erzeugt.

Für quadratische Datenmatrizen können die abstrakten Faktoren durch eine Eigenzerlegung (Abschnitt 2.5.6) bestimmt werden. Die bevorzugte Methode zur Faktorisierung einer Datenmatrix ist jedoch die Singulärwertzerlegung (Abschnitt 2.10.5). Sie zerlegt beliebige rechteckige Matrizen und ist zudem numerisch stabiler. Die Hauptachsen sind dann durch die Singulärvektoren und die Streuungen durch die Quadrate der zugehörigen Singulärwerte gegeben. Die Singulärvektoren werden nach der Größe der zugehörigen Singulärwerte fallend geordnet. Mit abnehmenden Singulärwerten nehmen die Streuungen in Richtung der Singulärvektoren ab.

Der vom Modell vorgegebene Vektorraum wird von den Merkmalen aufgespannt. Seine Dimension entspricht folglich der Anzahl der Merkmale. Tatsächlich liegen die Daten in einem Untervektorraum mit einer Dimension, die gleich der Anzahl der echt positiven Singulärwerte ist. Die Anzahl[2] der positiven

[2] Praktisch führen Fehler durch Rundungen und begrenzte Rechengenauigkeit in den numerischen Operationen zu kleinen Singulärwerten nahe 0, die theoretisch gleich 0 sein sollten. Dadurch spannt nahezu jeder Datensatz mit $m$ Merkmalsträgern und $n$ Merkmalen mit $m \leq n$ einen $m$-dimensionalen Raum auf, obwohl theoretisch die Daten in einem kleineren Untervektorraum liegen.

Singulärwerte ist somit auch gleich der Anzahl der linear unabhängigen Vektoren in den gegebenen Daten.

Ein weiteres Ziel der Hauptkomponentenanalyse ist die Verringerung der Dimension der Beobachtungsvektoren dadurch, daß die Daten nur über die Hauptkomponenten aufgespannt werden, bezüglich derer sie die größte Varianz haben. Einerseits soll das Problem durch die Reduzierung der Dimension vereinfacht werden. Andererseits sollen dabei möglichst wenig Informationen verloren gehen. Die Singulärwertzerlegung liefert eine Faktorisierung, die eine Verkleinerung der Dimension mit minimalem Fehler zur Ausgangsmatrix im Sinne der Matrix-2-Norm ermöglicht (Abschnitt 2.10.5). Ist die Streuung eines Datensatzes in eine Richtung gering, bedeutet dies, daß alle Daten in diese Richtung eine vergleichsweise ähnliche Ausprägung haben und sich bezüglich des durch diese Richtung repräsentierten abstrakten Merkmals kaum unterscheiden. Eine solche Richtung trägt dementsprechend wenig zur Diskriminierung der Daten bei. Sie trägt gar nicht zur Unterscheidung der Daten bei, wenn alle Merkmalsträger in dieser Richtung denselben Wert haben. Die Singulärvektoren mit den kleinsten Singulärwerten geben Richtungen an, die nur wenig über die Daten aussagen. Deshalb ist es möglich, für die weitere Analyse nur die Singulärvektoren mit den größten Singulärwerten auszuwählen und die Singulärvektoren mit den kleinsten Singulärwerten wegzulassen. Die Untersuchungseinheiten werden dazu in den Untervektorraum projiziert, der die Streuung entlang der Singulärvektoren – wie oben beschrieben – bestmöglich wiedergibt. Die Singulärrichtungen, die nur wenig über die Daten aussagen, werden gestrichen und die dimensionsverminderten Daten in den ursprünglichen Raum zurücktransformiert.

Die Bestimmung der Anzahl der wesentlichen Singulärvektoren ist eine offene Frage. Es gibt einige – darunter auch schlicht unsinnige – Kriterien, aber keines, daß die Zahl notwendiger und hinreichender Komponenten objektiv ermittelt. Ihre Anzahl kann lediglich eingegrenzt werden (Diehl und Kohr, 1983, S. 362 f.). Sie hängt letztlich nur davon ab, welche Größe des Unterschieds zwischen der Matrix und ihrer Reduktion noch hinnehmbar ist. Je kleiner die Singulärwerte der gestrichenen Komponenten sind, desto kleiner ist die Matrix-2-Norm der Differenz der Matrix und ihrer Reduktion. Der Fehler ist vernachlässigbar gering, wenn nur Komponenten mit Singulärwerten nahe 0 getilgt werden. Sind die Singulärwerte entfernter Komponenten größer, werden die Daten verändert. Indirekt wird der Fehler dadurch gering gehalten, daß von den größten Singulärvektoren nur

diejenigen beibehalten werden, die einen vorgegebenen hohen[3] Prozentsatz der Gesamtmasse der Streuung auf sich konzentrieren (Abschnitt 2.15).

In vielen praktischen Anwendungen hat sich die Hauptkomponentenanalyse zur Datenkomprimierung zumindest nicht als Verschlechterung, häufig sogar als Verbesserung herausgestellt. Das liegt wohl im wesentlichen daran, daß nach dem Löschen feiner Unterschiede nur die Grobstruktur übrig bleibt.

## 4.8 Latent-Semantic-Indexing

Es sei $m$ die Anzahl der Terme und $n$ die Anzahl der Texte. Ein Term wird repräsentiert durch einen Vektor in $\mathbb{R}^n$, wobei jede der $n$ Achsen einen Text repräsentiert. Die Texte sind die Merkmale. Das Latent-Semantic-Indexing ersetzt die Texte durch neue abstrakte Merkmale (Hauptkomponenten). Die Terme werden so nicht mehr durch Texte, sondern durch die nicht beobachtbaren, eigentlichen (*underlying, latent, hidden*) Konzepte beschrieben. Diese latenten Strukturen sind keine festen Abbildungen zwischen Termen und Konzepten, sondern hängen von den Verwendungsregularitäten der Terme im konkret verwendeten Korpus ab.

Die Term-Dokument-Matrix $A \in \mathbb{R}^{m \times n}$ wird mittels Singulärwertzerlegung in ein Produkt aus drei Matrizen zerlegt, $A = U\Sigma V^T$. Die Zeilen der Matrix $A$ stehen für $m$ Terme in einem $n$-dimensionalen Vektorraum. Die Elemente der Standardbasis entsprechen den $n$ Texten. Es sei $\mathcal{E}_n$ die kanonische Basis in $\mathbb{R}^n$. Es sei $\mathcal{V}$ die Basis aus den rechten Singulärvektoren. Diese Basis ist das Ergebnis der Transformation der Standardbasis, so daß die Streuung entlang der transformierten Achsen maximal ist. Deren Elemente bildeten die Spalten von $V$. Zum Übergang von $\mathcal{E}_n$ nach $\mathcal{V}$ gehört die Basiswechselmatrix $V = VE_n^{-1}$ und die Koordinatentransformationsmatrix $V^{-1}$ (Abschnitt 2.4.4). Es gilt $V^{-1} = V^T$, weil $V$ orthogonal ist (Abschnitt 2.10.2). Die Koordinaten der Termvektoren bezüglich der kanonischen Basis sind die Termvektoren selbst. Die Koordinaten der Termvektoren bezüglich der transformierten Basis $\mathcal{V}$ sind die Spalten von $V^T A^T = (AV)^T$ und somit die Zeilen von $AV = U\Sigma$. Alle Spalten von $U\Sigma$, deren Index größer als der Rang $r$ von $A$ ist, sind 0, weil die Termvektoren zwar Elemente eines $n$-dimensionalen Raumes sind, zur ihrer Beschreibung aber ein $r$-dimensionaler Untervektorraum genügt. Die Werte der ersten Spalten

[3] meist über 90%, aber natürlich stets auch abhängig von den konkreten Daten und Untersuchungszielen

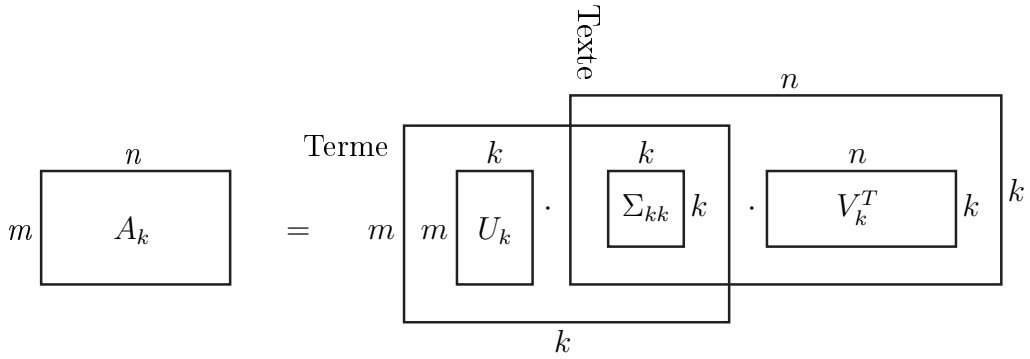



**Abbildung 4.2:** Singulärwertzerlegung im Latent-Semantic-Indexing

von $U\Sigma$ unterscheiden sich am stärksten. Mit wachsendem Spaltenindex nimmt die Variation innerhalb einer Spalte ab. Die Komponenten der letzten Zeilen unterscheiden sich nur noch relativ wenig.

Betrachtet man anstelle der Terme die Texte, sind also die Texte die Untersuchungseinheiten und die Terme die Merkmale, ist die Singulärwertzerlegung wie folgt zu interpretieren. Die Spalten der $m \times n$-Matrix $A$ repräsentieren $n$ Texte in einem $m$-dimensionalen Vektorraum. Jede Dimension entspricht einem der $m$ Terme. Der Vektorraum wird deshalb von der $m$-elementigen Standardbasis erzeugt. Es sei $\mathcal{E}_m$ die $m$-elementige kanonische Basis. Die Spalten der $m \times m$-Einheitsmatrix $E_m$ sind die Einheitsvektoren dieser Basis. Es sei $\mathcal{U}$ die Basis aus den linken Singulärvektoren und $U$ eine Matrix gebildet aus den linken Singulärvektoren als Spalten. Zum Wechsel von $\mathcal{E}_m$ nach $\mathcal{U}$ gehört die Basiswechselmatrix $U = UE_m^{-1}$ und die Koordinatentransformationsmatrix $U^{-1}$. Es gilt $U^{-1} = U^T$, weil $U$ orthogonal ist. Die Koordinatenvektoren der Textvektoren bezüglich der kanonischen Basis sind die Textvektoren selbst. Die Koordinatenvektoren der Textvektoren bezüglich der Basis $\mathcal{U}$ sind die Spalten[4] von $U^TA = \Sigma V^T$. Alle Zeilen von $U^TA$, deren Index größer als der Rang von $A$ ist, sind 0.

Dimensionsreduktion wird dadurch erreicht, daß nur die ersten $k$ Singulärwerte $\sigma_1, \sigma_2, \ldots, \sigma_k$ beibehalten und alle anderen auf 0 gesetzt werden. Es sei $\Sigma_k$ die so gewonnene Diagonalmatrix. Dann ist $A_k = U\Sigma_k V^T$ die reduzierte Matrix. Die wesentlichen Eigenschaften des Datensatzes bleiben erhalten, weil die Elemente der Hauptdiagonalen der Größe nach absteigend geordnet sind. Die reduzierte

[4]Mitunter werden auch die Zeilen von $U$ als Terme und die Spalten von $V^T$ als Texte bezeichnet. Diese nicht ganz richtige Setzung ist motiviert dadurch, daß die Diagonalmatrix $\Sigma$ lediglich die Richtungen streckt oder kürzt.

Matrix $A_k$ hat dieselbe Anzahl an Zeilen und Spalten wie $A$. Im allgemeinen hat sie auch keine Nullzeilen oder Nullspalten mehr als $A$. Lediglich ihr Rang $k$ ist kleiner als der Rang $r$ von $A$. Die Zeilen liegen in dem $k$-dimensionalen Untervektorraum, der von den ersten $k$ rechten Singulärwerten aufgespannt wird. Die Spalten liegen in dem $k$-dimensionalen Untervektorraum, der von den ersten $k$ linken Singulärwerten erzeugt wird. Die reduzierte Matrix läßt sich auch entsprechend der reduzierten Singulärwertzerlegung durch $A_k = U_k \Sigma_{kk} V_k^T$ bestimmen, wobei $\Sigma_{kk} = \Sigma(1{:}k, 1{:}k)$ nur aus den ersten $k$ Spalten und $k$ Zeilen von $\Sigma$, $U_k = U(:, 1{:}k)$ und $V_k = V(:, 1{:}k)$ nur aus den ersten $k$ Spalten von $U$ bzw. $V$ bestehen (Abbildung 4.2).

Eine Datenreduktion, das heißt eine Verkleinerung der Datenmatrix durch Streichung von Zeilen oder Spalten, ist nur möglich, wenn innerhalb der Untervektorräume mit den transformierten Daten gerechnet wird. Alle Spalten von $A_k V = U \Sigma_k$ mit einem Index größer als $k$ sind 0. Alle Zeilen von $U^T A_k = \Sigma_k V^T$ mit einem Index größer als $k$ sind 0.

## 4.9 Assoziationsmaße

Rieger modelliert Bedeutung sprachlicher Zeichen als einen zweistufigen Prozeß zunehmender kombinatorischer Einschränkung jeweils noch vorhandener Wahlmöglichkeiten in einem Vektorraummodell zweiter Ordnung (Rieger, 1977; Rieger, 1989, S. 194 ff.; Rieger, 1990). Diesen Einschränkungen entsprechen die linguistischen Grundrelationen der syntagmatischen und paradigmatischen Beziehungen von sprachlichen Ausdrücken. In der erste Stufe werden die syntagmatischen Assoziationen der Terme aus den Häufigkeiten der Kookurrenzen der Terme in den Texten eines Korpus berechnet. In der zweiten Stufe werden die paradigmatischen Assoziationen aus den syntagmatischen Assoziationen der ersten Stufe durch Vergleich der Kotexte, in denen die Terme vorkommen, ermittelt. Die gezählten Vorkommenshäufigkeiten werden so auf Vektoren, den sogenannten Bedeutungspunkten, eines hochdimensionalen Vektorraums, der semantischer Raum heißt, abgebildet. Die Lage der Vektoren zueinander gibt Aufschluß über die Bedeutungsähnlichkeiten der durch sie repräsentierten Terme. Die paarweisen Ähnlichkeiten der Terme werden aus deren Gewichtsvektoren berechnet. Je größer die Anzahl der Vektorkomponenten ist, desto mehr Möglichkeiten gibt es für Vektoren, verschieden zu sein. Es wird vereinfachend angenommen, daß alle Terme eines Textes syntagmatisch miteinander verbunden

sind. Terme, die häufig gemeinsam auftreten, sind also syntagmatisch affin und heißen assoziiert in erster Ordnung. Terme, die die gleichen Assoziationen erster Ordnung eingehen, nennt man assoziiert in zweiter Ordnung. Sie sind in den gleichen Kotexten austauschbar und somit paradigmatisch ähnlich. Für Termassoziationen mit einer Ordnung größer als zwei gibt es keine sinnvolle Interpretation.

### 4.9.1 Syntagmatische Ähnlichkeiten

Für jedes Paar von Termen wird in jedem Text eine lokale syntagmatische Ähnlichkeit als Funktion der Vorkommenshäufigkeiten bestimmt. Diese Funktion bewertet ähnliche Termgewichte stark und voneinander abweichende schwach. Die lokalen Ähnlichkeiten werden so zu einer globalen syntagmatischen Ähnlichkeit zusammengefaßt, daß viele hohe textweise Ähnlichkeiten zu einer hohen globalen Ähnlichkeit beitragen, viele niedrige lokale Ähnlichkeiten ihr aber entgegenwirken. Maße, die das Nichtvorkommen zweier Terme in einem Text als Ähnlichkeit bewerten, dürfen nicht verwendet werden. Terme sind nicht miteinander assoziiert, nur weil sie in vielen Texten beide nicht vorkommen. Sie sind lediglich unzureichend miteinander vergleichbar, weil die Zahl der Texte nicht ausreicht, eine aussagestarke Ähnlichkeit bestimmen zu können. Zur Berechnung der Ähnlichkeit zweier Terme aus deren Vorkommenshäufigkeiten in Texten dürfen somit nur diejenigen Texte herangezogen werden, in denen wenigstens einer der beiden Terme vorkommt. Die Ähnlichkeit eines jeden Paares von Termen wird so aufgrund anderer und auch unterschiedlich vieler Texte berechnet. Die Güte einer solchen Ähnlichkeit ist höher, wenn viele Texte dazu beitragen. Sind die Terme hingegen nur in wenigen Texten enthalten, ist sie gering. Die Ähnlichkeit und ihre Güte werden zu einem Wert verschmolzen, so daß Termpaare, deren Assoziation auf einer kleineren Sammlung von Texten beruht, auch einen kleineren Ähnlichkeitswert erhalten. Geringe Ähnlichkeit ist von schwacher Güte nicht mehr zu unterscheiden. Bei der Auswahl der Terme und der Texte ist deshalb darauf zu achten, daß die Unterschiede in der Zahl der den Ähnlichkeitswerten zugrundeliegenden Texten nicht zu groß sind.

Zählen zwei Terme in einem Text gleich viele Vorkommen, kann daraus nicht geschlossen werden, daß sie lokal ähnlich sind. Vielmehr muß die Termhäufigkeit im Verhältnis zur Korpushäufigkeit gesehen werden. Ein Term, der im ganzen Korpus nur selten vorkommt, aber in einem einzelnen Text die gleiche Häufigkeit wie ein Term mit hoher Korpushäufigkeit aufweist, ist ein in diesem Maße

höheres lokales Gewicht zuzuordnen. Ein Term kommt häufig in einem Text vor, wenn seine prozentuale Häufigkeit in diesem größer ist als sein Anteil an der Gesamthäufigkeit (Abschnitt 4.9.1.5).

Die syntagmatische Ähnlichkeit zweier Terme wird aufgrund gemeinsamen Vorkommens in Texten als Merkmal bestimmt. Sie steigt mit wachsender Anzahl gemeinsamer Merkmalsausprägungen. Es sei $x := (x_1, \ldots, x_n)$ der Merkmalsvektor eines Termes mit Komponenten aus $\{1, 0\}$, wobei $x_i = 1$ angibt, daß der Term im $i$-ten Text vorkommt und $x_i = 0$ besagt, daß er dort nicht vorkommt. Der binäre Vektor $x$ gibt an, in welchen Texten ein Term auftritt. Das kanonische Skalarprodukt $\sum_{i=1}^{n} x_i y_i \in \{0, 1, \ldots, n\}$ zählt, in wie vielen Texten die Terme $x$ und $y$ zusammen vorkommen. Es wird nur gemessen, ob Terme gemeinsam auftreten, weswegen es sich hier um Kookkurrenzen von Types handelt. Werden zusätzlich die Häufigkeiten beachtet, sind die Komponenten also nicht mehr 0 oder 1, sondern nichtnegativ ganzzahlig, werden die Kookkurrenzen von Token gezählt. Kommen im $i$-ten Text $x_i \in \mathbb{N} \cup \{0\}$ Token vom Type $x$ und $y_i$ Token vom Type $y$ vor, zählt man $x_i y_i$ mögliche Token-Token-Paare. Die Häufigkeit kann als Gewicht verstanden werden, welches schwerer wiegt, wenn der Term häufiger vorkommt.

Dem Termgewicht 0, welches das Nichtvorkommen eines Termes in einem Text angibt, kommt eine besondere Bedeutung zu. Es ist nicht ein beliebiger Wert auf der Skala der Merkmalsausprägungen, sondern muß bei der Konstruktion eines Maßes für die syntagmatische Ähnlichkeit gesondert behandelt werden. Metriken können deshalb generell nicht verwendet werden. Wird nur gemessen, ob ein Term in einem Text vorkommt oder nicht, ist ein binäres Ähnlichkeitsmaß (Abschnitt 2.17.9) nur aus den Anzahlen $n_{11}$, $n_{01}$ und $n_{10}$ zu bilden. Es setzt die Affinitäten $n_{11}$ mit den Repugnanzen $n_{01}$ und $n_{10}$ in Beziehung. Alle binären Ähnlichkeitsfunktionen, die die Anzahl der Nichtübereinstimmungen $n_{00}$ betrachten, können nicht sinnvoll verwendet werden. Dazu gehören das kanonische Skalarprodukt für binäre Vektoren, der Simple-Matching-Coefficient und das Maß von HAMANN. Die bei diesen Koeffizienten deutlich ausgeprägten Ähnlichkeiten der Terme sind einzig auf die Anzahl $n_{00}$ der Abwesendheiten eines Termpaares in beiden Texten zurückzuführen. Ebenso verhält es sich mit Metriken für allgemeine Gewichte. Im folgenden werden einige Funktionen vorgestellten, die den Anforderungen an ein Maß für syntagmatische Ähnlichkeiten gerecht werden.

#### 4.9.1.1 Ein Assoziationsmaß für binäre Termgewichte

Es sei $n_{1\cdot} \geq 1$ oder $n_{\cdot 1} \geq 1$. Damit kommt wenigstens einer der beiden Terme wenigstens einmal in einem Text vor. In

$$n_{11} - u \;=\; n_{11} - (n_{01} + n_{10}) \;\in \{-n, -n+1, \ldots, n-1, n\}$$

ist $n_{11}$ der Wert für die syntagmatische Ähnlichkeit und $n_{01} + n_{10}$ der für die syntagmatische Unähnlichkeit der beiden betrachteten Terme. Ist der Wert der Ähnlichkeit größer als der der Unähnlichkeit, $n_{11} > n_{01} + n_{10}$, ziehen sich die Terme an. Überwiegt die Unähnlichkeit, $n_{11} < n_{01} + n_{10}$, stoßen sie sich ab. Sind beide Werte gleich, so sind sie indifferent. Geteilt durch die Anzahl $n_{11}+n_{01}+n_{10}$ der Texte, in denen wenigstens einer der Terme vorkommt, ergibt sich

$$r := \frac{n_{11} - u}{n_{11} + u} \;=\; \frac{n_{11} - (n_{01} + n_{10})}{n_{11} + n_{01} + n_{10}} \;=\; \frac{n_{11} - n_{01} - n_{10}}{n_{11} + n_{01} + n_{10}} \;\in [-1, +1].$$

Die einleitende Setzung verhindert eine Division durch 0. Der Koeffizient ist $-1$ genau dann, wenn $n_{11} = 0$. Dann sind nur Repugnanzen, aber keine Affinitäten aufgetreten. Er ist 0 genau dann, wenn $n_{11} = n_{01}+n_{10}$. Er ist 1 genau dann, wenn $n_{01} + n_{10} = 0$. Dann wurden nur Affinitäten, aber keine Repugnanzen gezählt.

Das Maß läßt sich auch textweise herleiten. Die Texte werden einzeln bewertet und die Einzelwertung aufsummiert. Die Summe wird durch den Betrag der größtmöglichen Gesamtbewertung dividiert, um einen Wert aus dem Intervall $[-1, 1]$ zu erhalten. Ein Text erhält eine $+1$, wenn beide Terme darin vorkommen. Ihm wird eine $-1$ zugeordnet, wenn nur einer der Terme darin vorkommt. Kommen beide Terme in einem Text nicht vor, wird dieser rein technisch mit 0 bewertet, damit er zur Gesamtsumme nicht beiträgt. Letztere Texte werden auch nicht zur Berechnung der betraglich maximal möglichen Gesamtnote herangezogen. Dann ist

$$\begin{aligned} t : \{0,1\} \times \{0,1\} \;&\to\; \{-1, 0, 1\}, \\ (x, y) \;&\mapsto \begin{cases} +1, & \text{falls } \; x = y = 1, \\ \;\;0, & \text{falls } \; x = y = 0, \\ -1, & \text{sonst}, \end{cases} \end{aligned}$$

die Funktion zur Bewertung der einzelnen Texte. Der Koeffizient ist dann

$$\begin{aligned} r : \{0,1\}^n \times \{0,1\}^n &\to [-1,1], \\ (x,y) &\mapsto \frac{\sum_{i=1}^{n} t(x_i,y_i)}{\sum_{i=1}^{n} |t(x_i,y_i)|}. \end{aligned} \tag{4.2}$$

Dieses Maß ergibt sich aus dem von HAMANN durch Streichen von $n_{00}$.

#### 4.9.1.2 Ein Assoziationsmaß für Termgewichte aus dem Einheitsintervall

Es seien $x, y \in [0,1]^n$, $n \in \mathbb{N}$, mit $x \neq 0$ oder $y \neq 0$ die Gewichtsvektoren zweier Terme. Einem Text wird durch

$$\begin{aligned} t : [0,1] \times [0,1] &\to [-1,1], \\ (x,y) &\mapsto \begin{cases} 0, & \text{falls } x = y = 0, \\ 1 - |x-y|, & \text{falls } x, y > 0, \\ -x, & \text{falls } x > 0,\ y = 0, \\ -y, & \text{falls } x = 0,\ y > 0, \end{cases} \end{aligned}$$

0 zugeordnet wenn er beide Terme nicht enthält. Kommen beide Terme in ihm vor, wird der Betrag der Differenz der zugehörigen Termgewichte vom größtmöglichen Affinitätswert 1 abgezogen. Je kleiner die Differenz ist, desto größer ist die Affinität. Enthält ein Text nur einen von beiden Termen, wird dies als Abstoßung gewertet. Je größer das Gewicht dieses einen Termes ist, desto größer ist die Repugnanz. Die Funktion $t$ ist unstetig auf den Achsen. Durch Aggregation der textweisen Assoziationen ergibt sich der Koeffizient

$$\begin{aligned} r : [0,1]^n \times [0,1]^n &\to [-1,1], \\ (x,y) &\mapsto \frac{\sum_{i=1}^{n} t(x_i,y_i)}{\sum_{i=1}^{n} |t(x_i,y_i)|}. \end{aligned} \tag{4.3}$$

Für Komponenten aus $\{0,1\}$ berechnet er dieselben Werte wie (4.2).

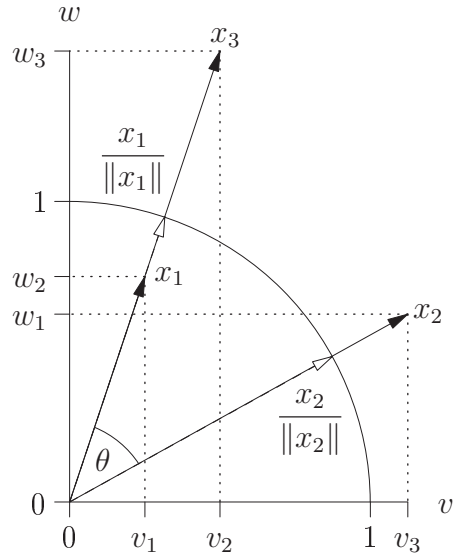



**Abbildung 4.3:** Die Richtung eines Termvektors $(x_1, x_2, x_3)$ ist bestimmt durch seine Gewichte $(v_1, v_2, v_3, w_1, w_2, w_3)$ für die Texte $(v, w)$: $x_1 = (v_1, w_1)$, $x_2 = (v_2, w_2)$, $x_3 = (v_3, w_3)$. Die Längen $(\|\cdot\|)$ der Vektoren werden zu 1 normiert $(\cdot/\|\cdot\|)$ und somit nicht betrachtet. Dadurch ist $x_1/\|x_1\| = x_3/\|x_3\|$. Die Ähnlichkeit zwischen zwei Termvektoren ist der Cosinus des von ihnen eingeschlossenen Winkels $(\cos\theta = \measuredangle(x_1, x_2) = \measuredangle(x_2, x_3) = \measuredangle(x_2/\|x_2\|, x_3/\|x_3\|))$.

#### 4.9.1.3 Kanonisches Skalarprodukt

Das kanonische Skalarprodukt $\sum_{i=1}^{n} x_i y_i$ direkt angewendet auf die Gewichtsvektoren $x$ und $y$ zweier Terme ist als Maß für die Ähnlichkeit erster Ordnung nicht geeignet, weil es die komplementäre Verteilung der Terme im Korpus nicht beschreibt. Der Beitrag zur Summe ist 0 nicht nur von Texten, in denen beide Terme nicht vorkommen, sondern auch von Texten, in denen nur einer von beiden fehlt. Letzteres beschreibt eine Abstoßung, die als solche hätte bewertet werden müssen. Hingegen bewertet ersteres einen für die Bestimmung der Ähnlichkeit gänzlich irrelevanten Text. Statt der Gewichtsvektoren selbst wird deshalb mit normierten Vektoren und somit mit dem Cosinus des Winkels zwischen den Vektoren gerechnet.

#### 4.9.1.4 Cosinus des Winkels

Die Richtung eines Termvektors ergibt sich aus dessen Gewichten. Je größer ein Gewicht ist, desto stärker wird der Vektor in die Richtung des zugehörigen Textes abgelenkt. Werden zwei Termvektoren stets ähnlich stark von den Textgewichten

beeinflußt, zeigen sie auch ungefähr in die gleiche Richtung. Zeigen zwei Vektoren in genau dieselbe Richtung, ist der zwischen ihnen liegende Winkel 0. Der Cosinus des Winkels ist 1. Sie sind maximal ähnlich, wenn sie in dieselbe Richtung zeigen, auch dann, wenn die Vektoren unterschiedlich lang sind. Stehen die Vektoren senkrecht aufeinander, ist der Winkel $\pi/2$ und dessen Cosinus 0. Zwei Termvektoren stehen senkrecht aufeinander, wenn sie in keinem Text gemeinsam vorkommen. Wird als Ähnlichkeit zweier Termvektoren der Cosinus des von ihnen eingeschlossenen Winkels gesetzt, ist die Orthogonalität die größte Unvereinbarkeit. Zeigen sie in entgegengesetzte Richtung, ist der Winkel $\pi$ und dessen Cosinus $-1$. Da die Termvektoren aber alle im ersten Quadranten liegen, kann dieser Fall nicht auftreten. Die Winkel liegen alle zwischen 0 und $\pi/2$. Die Ähnlichkeiten der Termvektoren liegen somit alle im Einheitsintervall $[0, 1]$, wobei 0 die kleinstmögliche und 1 die größtmögliche Ähnlichkeit angibt. Die Längen der Vektoren werden nicht beachtet.

#### 4.9.1.5 Riegers Korrelationskoeffizient

Es sei $x_{it} \in \mathbb{R}$, $i = 1, \ldots, m$, $t = 1, \ldots, n$, das Gewicht des $i$-ten Types im $t$-ten Text. Die Gewichte bilden eine $m \times n$-Matrix $(x_{it})_{1 \leq i \leq m,\, 1 \leq t \leq n}$ mit Komponenten aus $\mathbb{R}$. Eine Zeile dieser Matrix besteht aus den Gewichten eines Termes in allen Texten. Eine Spalte enthält die Gewichte aller Terme eines Textes. Die Summe der $i$-ten Zeile

$$x_{i\cdot} := \sum_{t=1}^{n} x_{it}$$

ist dann das Gewicht des $i$-ten Termes im ganzen Korpus. Die Summe der $t$-ten Spalte

$$x_{\cdot t} := \sum_{i=1}^{m} x_{it}$$

ist das Gewicht des $t$-ten Textes. Das Gesamtgewicht des Korpus ist dann

$$x_{\cdot\cdot} := \sum_{t=1}^{n} \sum_{i=1}^{m} x_{it}.$$

Der $i$-te Term hat den Anteil $x_{i\cdot}/x_{\cdot\cdot}$ am Gesamtgewicht. In einem Text vom Gewicht $x_{\cdot t}$ müßte der $i$-te Term demnach $x^*_{it} := (x_{i\cdot}\, x_{\cdot t})/x_{\cdot\cdot}$ wiegen. Somit ist $x^*_{it}$ ein Schätzwert für das Gewicht des $i$-ten Terms im $t$-ten Text dafür, daß „ausschließlich der Zufall – nicht aber sprachliche Beziehungen – für [des Terms]

Vorkommen im betreffenden Text verantwortlich [ist]. Erst vor dem Hintergrund dieser in den jeweiligen Schätzwerten quantifizierten Zufälligkeit der Beziehungen von Elementen zueinander werden deren jeweiligen Abweichungen davon als eine möglicherweise systematische, auf strukturelle Zusammenhänge hinweisende Regularität der sprachlichen Beziehungen faßbar“ (RIEGER, 1989, S. 199). Die Abweichung des gemessenen vom geschätzten Wert ist $\delta_{it} := x_{it} - x^*_{it}$. Liegt der gemessene Wert $x_{it}$ oberhalb des Equilibriums $x^*_{it}$, wiegt der $i$-te Term im $t$-ten Text bezüglich seines Gesamtgewichts im ganzen Korpus schwer. Die Differenz $\delta_{it}$ ist positiv. Je größer sie ist, desto wesentlicher ist der Text als Merkmal für den Term. Liegt $x_{it}$ unterhalb, ist das Gewicht des $i$-ten Terms im $t$-ten Text gering. Die Abweichung ist negativ. Der RIEGERsche Korrelationskoeffizient

$$r_{ij} := \left\langle \frac{\delta_i}{\|\delta_i\|}, \frac{\delta_j}{\|\delta_j\|} \right\rangle = \frac{\sum_{t=1}^{n} \delta_{it}\,\delta_{jt}}{\sqrt{\sum_{t=1}^{n} \delta_{it}^2}\,\sqrt{\sum_{t=1}^{n} \delta_{jt}^2}}$$

ist das kanonische Skalarprodukt der normierten Abweichungsvektoren und somit nach Definition des Winkels (Abschnitt 2.9) der Cosinus des von den Abweichungsvektoren eingeschlossenen Winkels. Für alle $1 \leq i, j \leq m$ ist $r_{ij} \in [-1, 1]$, denn für $n \in \mathbb{N}$, $x_1, x_2, \ldots x_n$, $y_1, y_2, \ldots, y_n \in \mathbb{R}$ gilt $|\sum_{i=1}^{n} x_i y_i| \leq \sum_{i=1}^{n} |x_i y_i| \leq \sqrt{\sum_{i=1}^{n} |x_i|^2}\sqrt{\sum_{i=1}^{n} |y_i|^2}$ nach der CAUCHY-SCHWARZschen Ungleichung. Im Zähler werden die Produkte der Abweichungen der zu Paaren zusammengefaßten Meßwerte von ihren jeweiligen Schätzwerten aufsummiert. Der Nenner hat lediglich eine normierende Funktion. Er sorgt dafür, daß $|r_{ij}| \leq 1$ gilt und der Korrelationskoeffizient maßstabsunabhängig ist. Wegen der Normierung gehen nur die Richtungen der Abweichungsvektoren, jedoch nicht deren Länge in die Berechnung ein. Der Korrelationskoeffizient wird um so größer, je stärker diejenigen Wertepaare $(x_{it}, x_{jt})$ überwiegen, bei denen $x_{it}$ und $x_{jt}$ etwa gleich groß sind. Für $r_{ij} = 0$ heißen $x_i$ und $x_j$ unkorreliert, für $r_{ij} < 0$ negativ und für $r_{ij} > 0$ positiv korreliert. Die Matrix der Korrelationskoeffizienten $(r_{ij})_{1 \leq i,j \leq m} \in [-1, 1]^{m \times m}$ ist symmetrisch. Die Diagonalelemente sind alle 1.

RIEGER verwendet anstelle der allgemeineren Gewichte die Vorkommenshäufigkeiten selbst. Dann gibt $x_{ij} \in \mathbb{N} \cup \{0\}$ die Anzahl der Termtoken des $i$-ten Types im $t$-ten Text an. Die Matrix $(x_{it})_{1 \leq i \leq m, 1 \leq t \leq n}$ ist die Term-Dokument-Matrix. Die Summe $x_{i\cdot}$ der $i$-ten Zeile ist die Häufigkeit des $i$-ten Types im ganzen

Korpus. Die Summe $x_{\cdot t}$ der $t$-ten Spalte ist die Länge des $t$-ten Textes gemessen in Termtoken. Das Korpus besteht aus $x_{\cdot\cdot}$ Token. Jeder $(x_{\cdot\cdot}/x_{i\cdot})$-te Token im Korpus ist statistisch gesehen vom Type $i$. In einem Text aus $x_{\cdot t}$ Token dürfte demnach der $i$-te Type

$$x^*_{it} = \frac{x_{\cdot t}}{\frac{x_{\cdot\cdot}}{x_{i\cdot}}} = \frac{x_{i\cdot}\, x_{\cdot t}}{x_{\cdot\cdot}}$$

mal erwartet werden.

### 4.9.2 Paradigmatische Ähnlichkeiten

Die syntagmatische Ähnlichkeit der Terme $i$ und $j$ im Vektorraummodell ist $x^{(1)}_{ij} := s^{(0)}\big(x^{(0)}_i, x^{(0)}_j\big)$. Der Vektor $x^{(1)}_i := \big(x^{(1)}_{i1}, x^{(1)}_{i2}, \ldots, x^{(1)}_{im}\big)$ faßt alle numerisch spezifizierten syntagmatischen Verwendungsregularitäten des $i$-ten Terms zu allen Termen zusammen. Die paradigmatische Ähnlichkeit der Terme $i$ und $j$ ist $x^{(2)}_{ij} := s^{(1)}\big(x^{(1)}_i, x^{(1)}_j\big)$. Hohe paradigmatische Ähnlichkeit liegt vor, wenn die Vektoren syntagmatischer Ähnlichkeiten komponentenweise geringe Unterschiede aufweisen. Die paradigmatische Ähnlichkeit ist also die Ähnlichkeit der syntagmatischen Ähnlichkeiten. Sie gibt den Grad der Austauschbarkeit zweier Terme an. Zur Bestimmung der paradigmatischen Ähnlichkeit sind folglich die gewöhnlichen Ähnlichkeitsmaße für Vektoren geeignet. Je gröber das Maß ist, desto deutlicher unterscheiden sich die Ähnlichkeiten.

### 4.9.3 Semantische Räume

Zwei Terme haben nun eine ähnliche Bedeutung, wenn sie nahezu gleiche paradigmatische Ähnlichkeiten aufweisen. Die zu Vektoren zusammengefaßten numerisch spezifizierten paradigmatischen Verwendungsregularitäten $x^{(2)}_i := \big(x^{(2)}_{i1}, x^{(2)}_{i2}, \ldots, x^{(2)}_{im}\big)$ sind die Bedeutungspunkte. Die Bedeutungspunkte werden zur Menge $\mathcal{X}^{(2)}$ zusammengefaßt. $\mathcal{X}^{(2)}$ zusammen mit einem Ähnlichkeitsmaß $s^{(2)}$ bildet den semantischen Raum $(\mathcal{X}^{(2)}, s^{(2)})$. Ein semantischer Raum stellt die Bedeutungsähnlichkeiten von Termen räumlich dar. Die topologischen Nachbarschaften der Bedeutungspunkte gibt Aufschluß über die Bedeutungsähnlichkeiten der Terme. Die Ähnlichkeiten $x^{(3)}_{ij} := s^{(2)}\big(x^{(2)}_i, x^{(2)}_j\big)$ der Bedeutungspunkte dürfen nicht als absolute Werte verstanden werden. Sie ermöglichen den Vergleich von Bedeutungspunkten stets nur in Bezug auf die Population als ganzes. Terme, deren Bedeutungspunkte im semantischen Raum näher zusammen liegen, sind in

ihren Bedeutungen ähnlicher, als solche, deren Punkte weit voneinander entfernt sind. Der semantische Raum setzt jeden Term zu jedem anderen in Beziehung. Aufgrund seiner Konstruktion hat er daher genauso viele Dimensionen wie Elemente und ist deswegen sehr dünn besetzt. Er ist so konstruiert, daß alle Terme paarweise nichts miteinander zu tun haben können, das heißt auf gleiche Weise maximal verschieden sein können.

Zu jeder Textsammlung und jeder Auswahl von Termen daraus gehört ein semantischer Raum. Wird ein Text oder auch nur ein Term zur Untersuchung hinzugenommen oder entfernt, muß der gesamte semantische Raum über beide Stufen neu berechnet werden. Eine inkrementelle Berechnung bei Erweiterung der Datengrundlage um einen Term ist nicht möglich.

# Kapitel 5

# Ergebnisse

Die im folgenden beschriebenen Beobachtungen gehen meist auf die Untersuchung der Hierarchien agglomerativer Verfahren zurück. Die Betrachtung einer ganzen Agglomerationshierarchie ermöglicht nämlich einen bei weitem aufschlußreicheren Überblick über die Daten als der vergleichsweise kleine Einblick durch einzelne von partitionierenden Verfahren berechnete Zerlegungen. Hierarchisch-agglomerative Verfahren sind immer dann vorzuziehen, wenn über die Struktur der Daten im voraus nichts bekannt ist (Abschnitt 3.2). Dendrogramme sind die einzige Möglichkeit zur Visualisierung der Clusterergebnisse, weil die Punkte des semantischen Raumes wegen dessen großer Dimension nicht selbst dargestellt werden können.

Nur die wesentlichen Erkenntnisse werden vorgestellt. Dazu genügt eine rein qualitative Beschreibung der Ergebnisse. Die Untersuchung beschränkt sich auf ein Korpus und ist auch nur relativ zu diesem Korpus zu beurteilen. Das Korpus ist *keine* Stichprobe, von der auf die Bedeutung der Terme im allgemeinen geschlossen werden darf. Ein Analyseverfahren gilt als geeignet, wenn die zu Clustern zusammengefaßten Elemente auch intuitiv syntagmatisch und paradigmatisch ähnliche sprachliche Ausdrücke repräsentieren.

## 5.1 Terme und Texte

Nicht alle Terme und Texte sind für die Untersuchung gleich gut geeignet. Sie müssen nicht nur theoretisch begründbaren Bedingungen genügen (Kapitel 4), sondern auch entsprechend einiger Größen des konkreten Korpusses ausgewählt werden. Alle Terme, die nicht ausgewählt wurden, werden gelöscht, so daß die Texte nur noch aus den zu untersuchenden Termen bestehen.

### 5.1.1 Ähnlichkeiten

Ziel ist es, Ähnlichkeiten von Termen aus Texten eines Korpus zu untersuchen. Terme und Texte müssen so ausgewählt werden, daß dies auch möglich ist. Es ist darauf zu achten, daß nicht nur Terme untersucht werden, für die von vornherein nur geringe paarweise Ähnlichkeiten bestimmt werden können. Dazu müssen ausreichend viele Terme in ausreichend vielen Texten zusammen vorkommen. Kommen hingegen Terme paarweise nur selten in denselben Texten vor, kann höchstens eine mehr oder weniger starke Unähnlichkeit berechnet werden. Im allgemeinen sind Terme zu betrachten, die in vielen, aber nicht in allen Texten vorkommen. Terme, die fast überall auftreten, haben zwar eine große Ähnlichkeit zu allen anderen, sind von ihnen aber auch zu wenig unterschieden, um sinnvoll Bedeutungen differenzieren zu können. Eine Obergrenze für die Texthäufigkeit eines Terms ist somit ebenfalls erforderlich.

### 5.1.2 Auswahl der Texte

Es werden nur Texte mit mittlerer[1] Zahl an Wörtern verwendet. Dadurch werden kurze Texte, zum Beispiel Agenturmeldungen, von der Untersuchung ausgeschlossen. Diese sind zu zahlreich und zu kurz, um als Kotexte zur Betrachtung von Verwendungsregularitäten dienen zu können. Andererseits werden lange Texte, zum Beispiel mehrseitige Berichte, nicht verwendet. Sie verbinden zu viele Terme miteinander, wodurch die Terme in ihrer Verwendung beliebig werden können.

Je weniger Texte verwendet werden, desto weniger intuitiv sind die berechneten Ähnlichkeiten. Der Datensatz darf nicht so klein sein, daß Regularitäten der Termverwendung überhaupt nicht mehr festgestellt werden können. Auch sind nicht alle Texte gleich gut geeignet. Texte können entweder nach der Anzahl ihrer Termtoken oder der Anzahl ihrer Termtypes (Abbildungen 5.1 und 5.2) ausgewählt werden. Die plausibel erscheinende Annahme, daß die in Token gemessenen längsten Texte maßgeblich für die

[1] zwischen 50 und 300 aller Wörter; erst später werden die Texte auf die ausgewählten Terme der Untersuchung reduziert.

Die hier und im folgenden angegebenen konstanten Werte und Wertebereiche haben sich im Laufe der Untersuchung ergeben. Sie resultieren zum einen aus der Anwendung der obigen theoretischen Überlegungen auf die konkret vorliegenden Daten. Zum anderen haben sie sich in vielen Testreihen auf diese Größenordnungen eingependelt. Die Parameter der Analyse sind keineswegs auf genau diese Werte festgelegt. Die Analyse ist im ganzen so stabil, daß kleinere bis mittlere Veränderungen die Ergebnisse qualitativ vernachlässigbar gering beeinflussen.

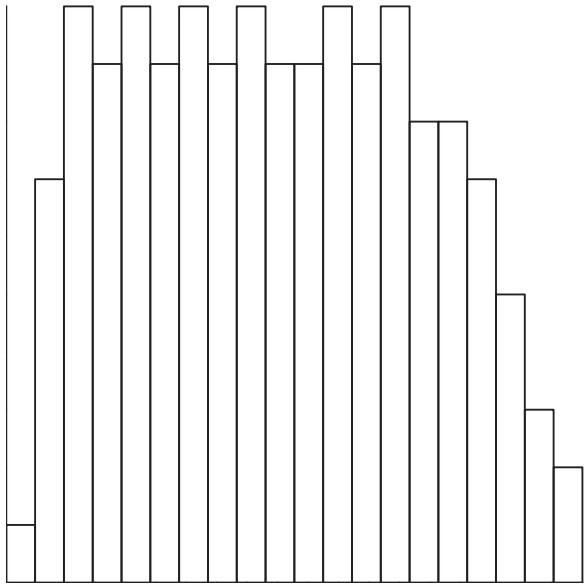


**Abbildung 5.1:** Klassierte Häufigkeiten der voneinander verschiedenen Längen der Texte in Tokens der Substantive

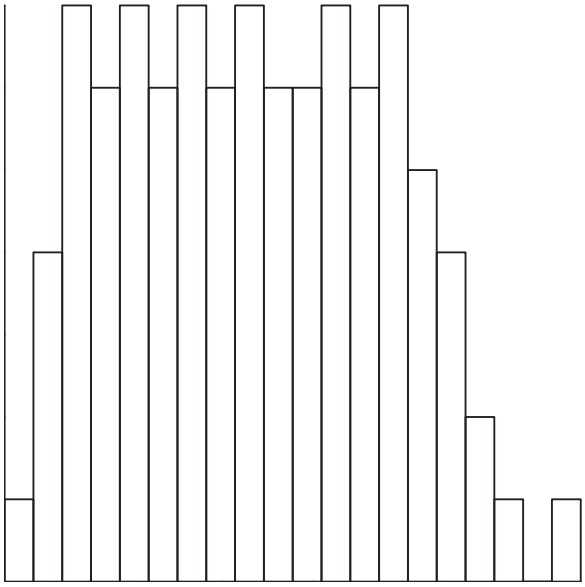


**Abbildung 5.2:** Klassierte Häufigkeiten der voneinander verschiedenen Anzahlen von Types der Substantive in den Texten

Ähnlichkeit der Terme sind und in Token gezählte kurze Texte nur einen geringen Beitrag dazu leisten, ist nicht richtig. Je weniger Texte mit geringer Zahl an Token verwendet werden, desto schlechter werden die Ähnlichkeiten. Werden zu viele dieser kurzen Texte aus dem Korpus gestrichen, wird er unbrauchbar. Das zeigt sich darin daß Terme, die bei hoher Textanzahl zutreffend in gleiche Gruppen geclustert werden, bei niedriger Textanzahl getrennt werden.

Ein Text muß, um überhaupt zur Ähnlichkeit zweier Terme beitragen zu können, Tokens von mindestens zwei unterschiedlichen Types aufweisen. Ansonsten würde der Text den einzigen Termtype nur von allen anderen abgrenzen. Ähnlich wie bei den Token nimmt auch hier die Granularität der Ähnlichkeit ab, wenn immer mehr Texte mit nur wenigen verschiedenen Types gestrichen werden. Jedoch ist auch bei geringer Textanzahl die niedrigere Auflösung noch aussagekräftig. Texte mit vielen verschiedenen Types sind für die Untersuchung wesentlich. Texte mit nur wenigen verschiedenen Types können weggelassen werden, ohne daß sich die Ergebnisse grundlegend verändern.

Bei der Auswahl der Texte wird nur ein Minimum[2] an Types vorgeschrieben. Die Textlänge in Token muß folglich mindestens genau so groß sein, da es zu jedem Type mindestens ein Token gibt. Umgekehrt hingegen kann nicht von der Tokenanzahl auf die Typeanzahl eines Textes geschlossen werden. Bei einem in Token gemessenen langen Text können die Token nur zu wenigen Types gehören.

### 5.1.3 Auswahl der Terme

Die Terme, die Gegenstand der Untersuchung sein sollen, werden aufgrund ihrer Texthäufigkeiten (Abbildung 5.3) ausgesucht. Ein Term, der in jedem zweiten oder dritten Text des Korpusses enthalten ist, kommt zu häufig vor, um seine Bedeutung durch seine Verwendungsregularität bestimmen zu können. Es gibt allerdings nur sehr wenige Terme mit einer so hohen Texthäufigkeit. Diese korrelieren mit nahezu allen Termen positiv und tragen somit zur paarweisen Ähnlichkeit nahezu aller Terme bei. Sie werden durch Vorgabe einer Obergrenze für die Texthäufigkeit von der Untersuchung ausgeschlossen. Allerdings hat sich gezeigt, daß mit abnehmender Texthäufigkeit die Ähnlichkeitsstrukturen gröber

[2] mindestens zehn Types pro Text haben sich als praktikabel erwiesen. Das absolute theoretische Minimum sind zwei Types pro Text. Ein solcher Text wäre aber einfach zu kurz, um zur Bestimmung der Affinität oder Repugnanz mehrerer Terme beitragen zu können.

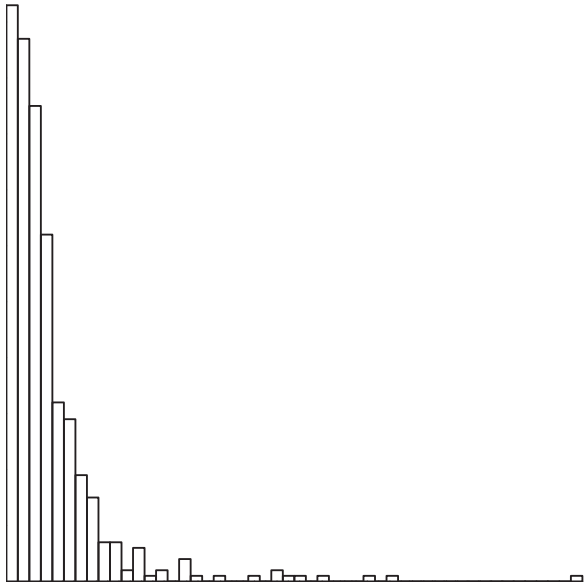


**Abbildung 5.3:** Klassierte Häufigkeiten der voneinander verschiedenen Texthäufigkeiten der Substantive aus etwa siebentausend Texten

werden. Außer den Termen, die in fast allen Texten vorkommen, sollten Terme mit möglichst großer[3] Texthäufigkeit untersucht werden.

Wegen des immensen Rechenaufwandes und Speicherbedarfs wird nur eine begrenzte[4] Anzahl an Termen betrachtet. Wird die Anzahl der Terme vergrößert, werden die Gruppen des Clusterergebnisses mit zusätzlichen Termen des Bedeutungsumfeldes aufgefüllt, im weiteren aber nicht verändert. So wird beispielsweise die Gruppe *Arbeitslosigkeit, Wirtschaft, Wachstum, wachsen, niedrig, sinken, steigen, Prozent* in dem Ausschnitt aus Abbildung 5.4 des Dendrogramms von 1000 Bedeutungswörtern bei einer Vergrößerung um weitere 1000 um *Konjunktur, Aufschwung, Rezession, Bruttoinlandsprodukt, Wirtschaftswachstum* ergänzt (Abbildung 5.5). Eine feinere Einteilung in Untergruppen ist möglich. Die Interpretation wird dadurch deutlich einfacher. Anhand der hinzugetretenen Terme ist ersichtlich, daß mancher Term, der vorher unangemessen zugeordnet erschien, sehr wohl richtig platziert ist.

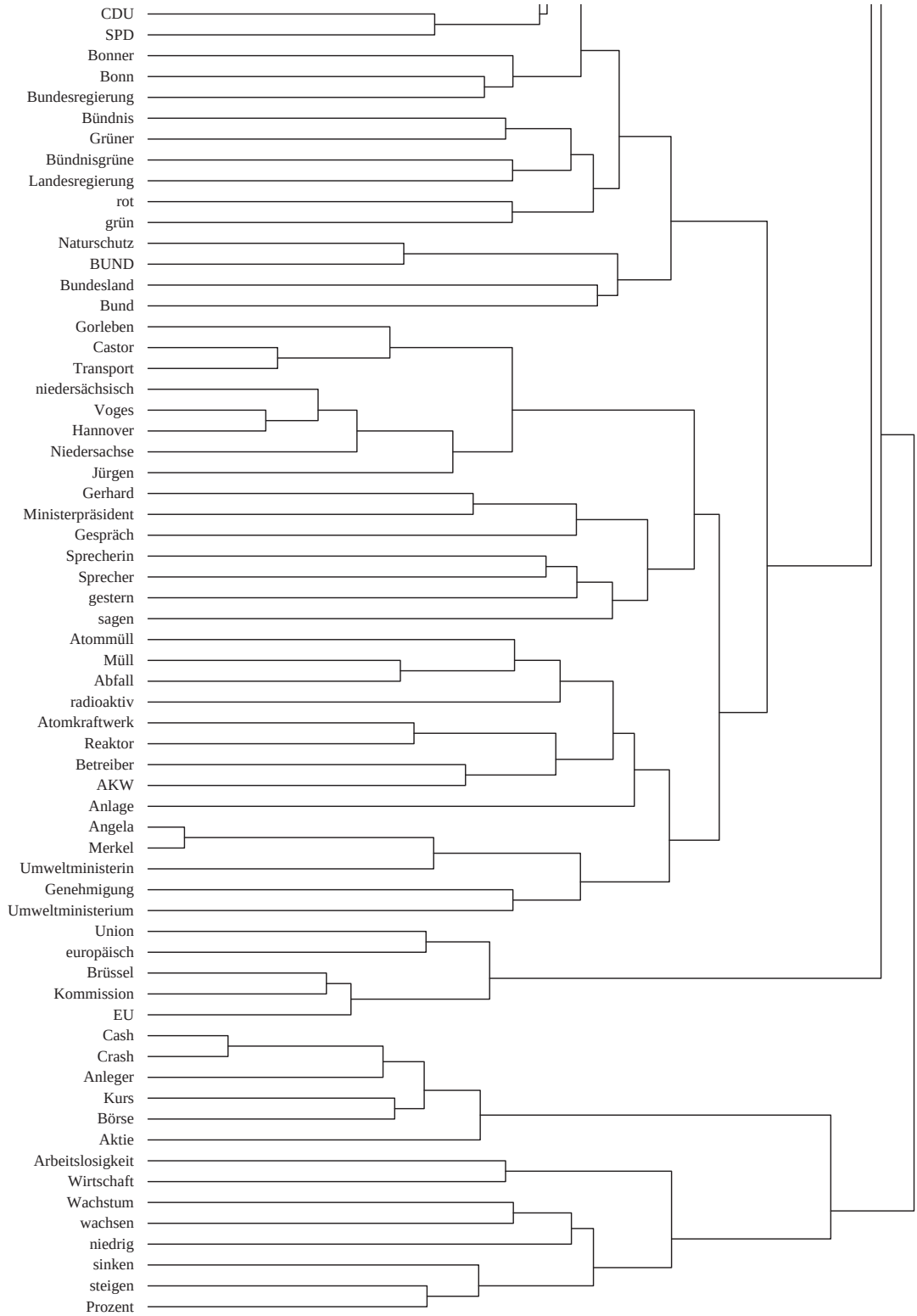



**Abbildung 5.4:** Ausschnitt des Dendrogramms von Average-Linkage mit der euklidischen Metrik der 1000 Bedeutungswörter mit dem syntagmatischen Assoziationsmaß von Rieger und dem paradigmatischen Assoziationsmaß von Tanimoto

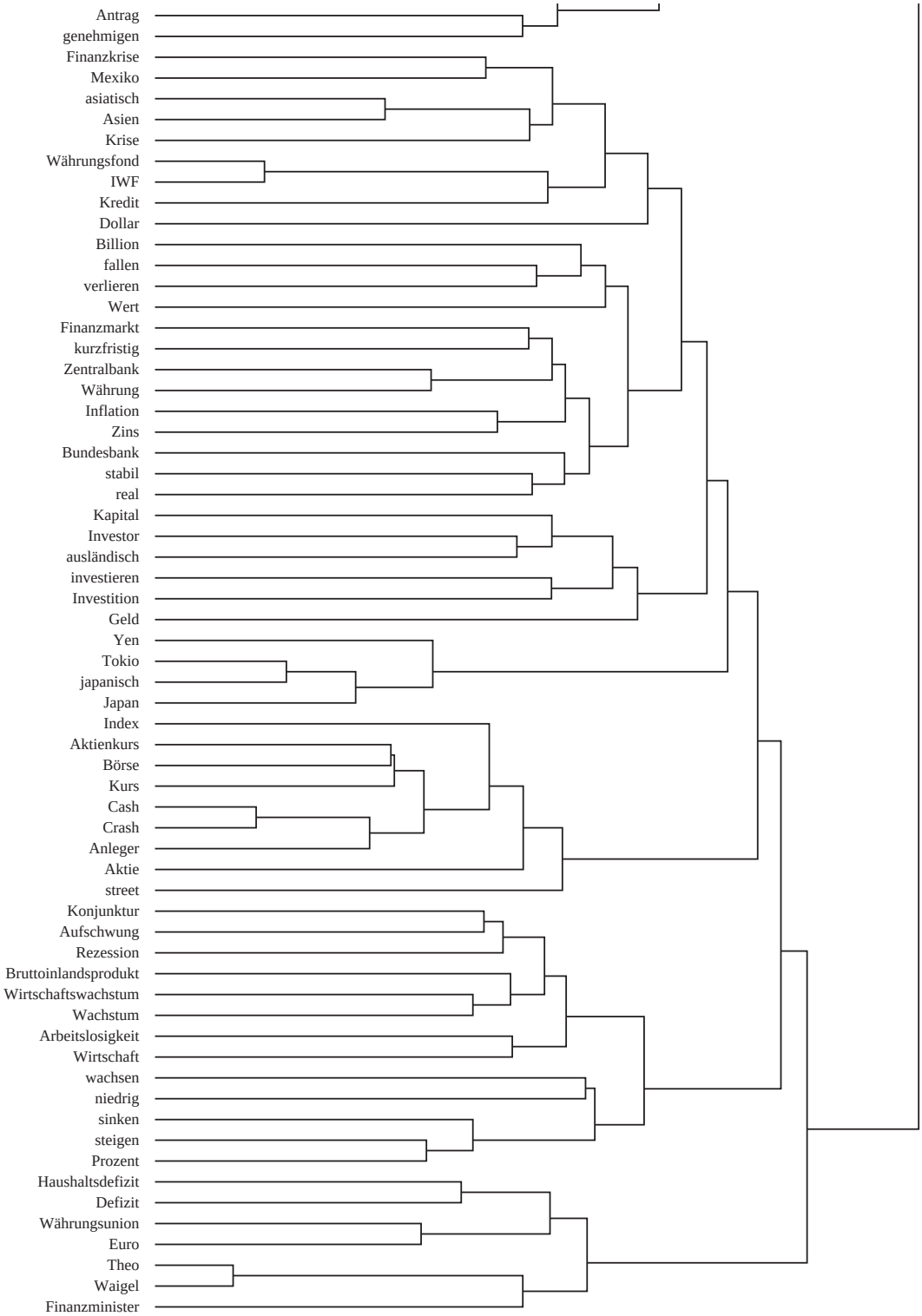



**Abbildung 5.5:** Ausschnitt des Dendrogramms von Average-Linkage mit der euklidischen Metrik der Bedeutungspunkte der auf 2000 Bedeutungswörter vergrößerten Grundgesamtheit mit dem syntagmatischen Assoziationsmaß von Rieger und dem paradigmatischen Assoziationsmaß von Tanimoto

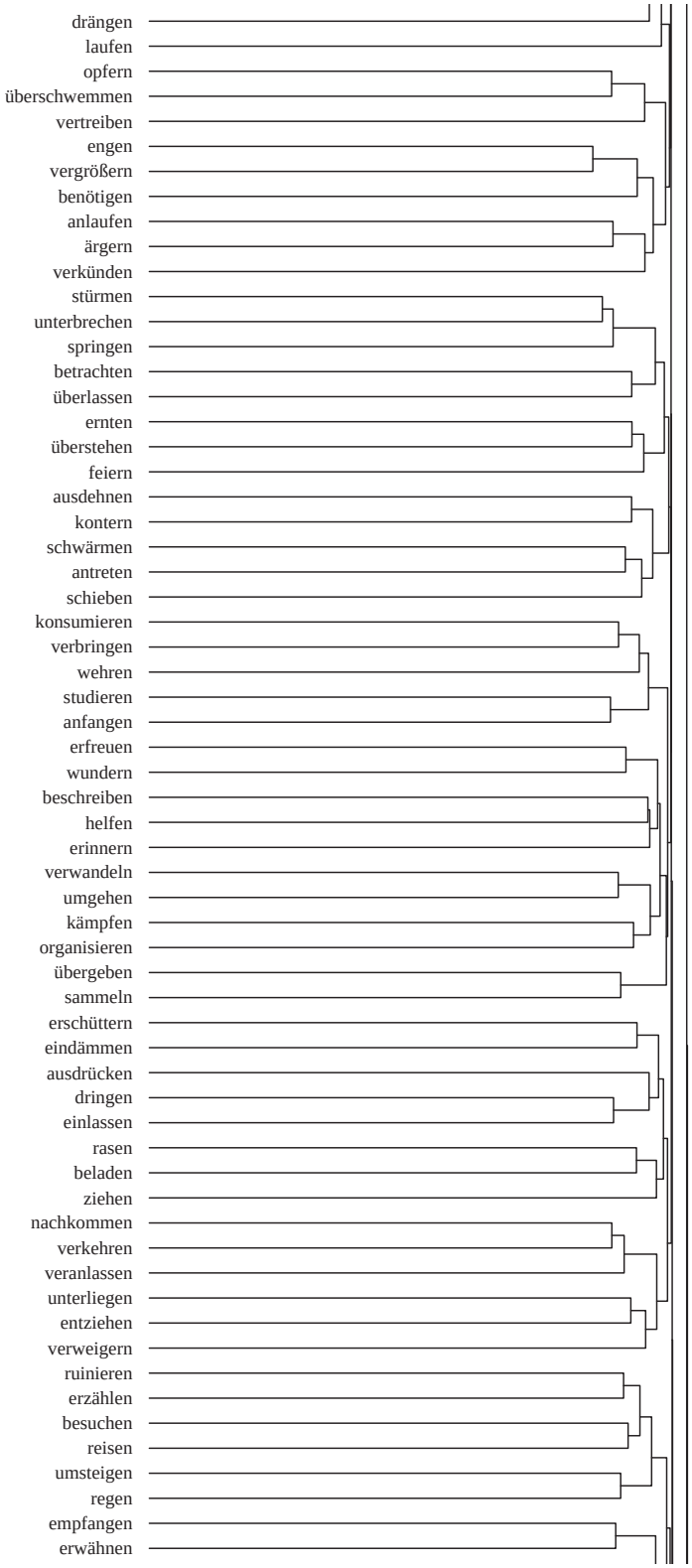



**Abbildung 5.6:** Ausschnitt des Dendrogramms von Average-Linkage mit der euklidischen Metrik der Bedeutungspunkte der Verben mit dem syntagmatischen Assoziationsmaß von Rieger und der euklidischen Metrik als paradigmatischem Assoziationsmaß

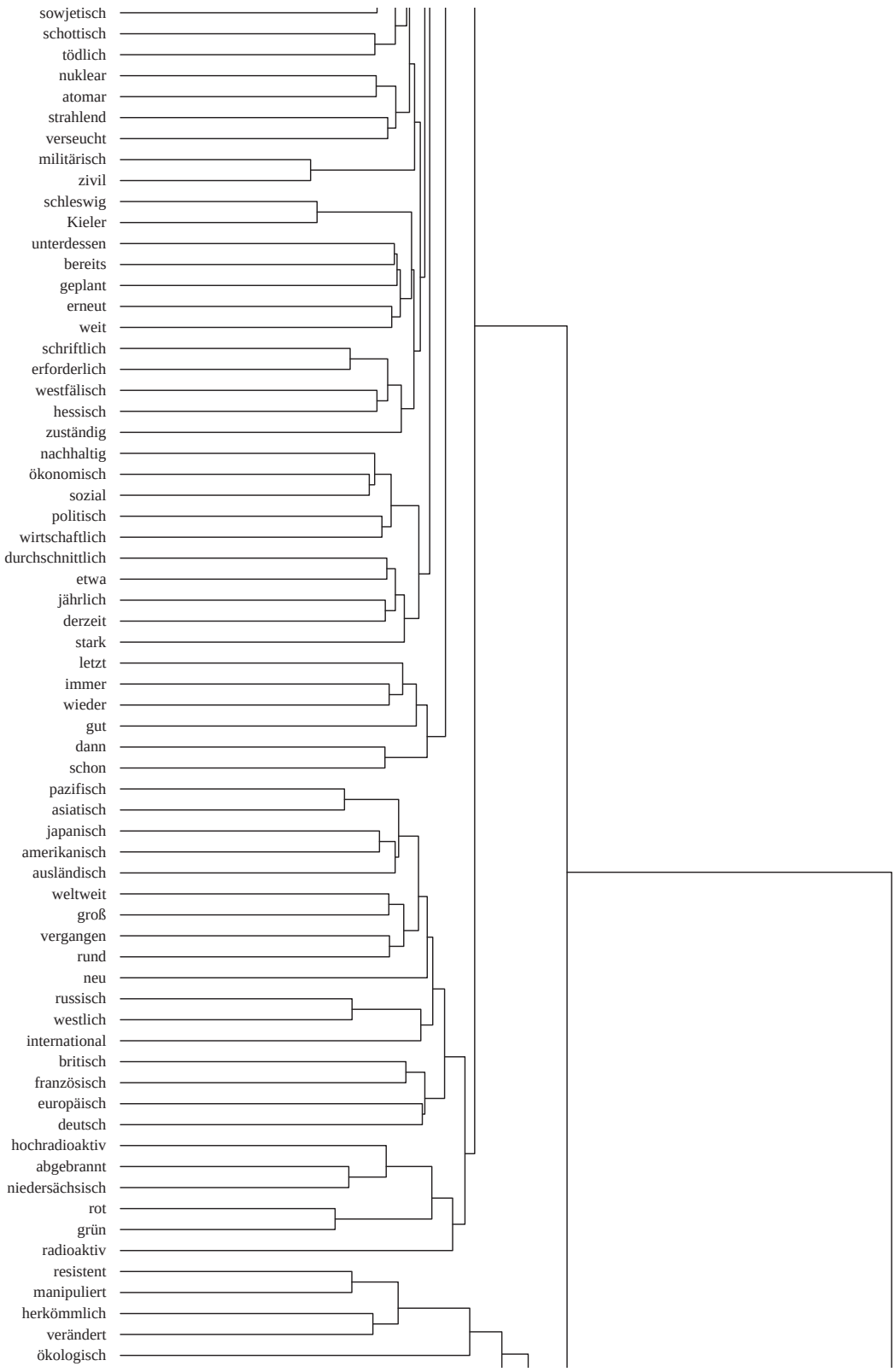



**Abbildung 5.7:** Ausschnitt des Dendrogramms von Average-Linkage mit der euklidischen Metrik der Bedeutungspunkte der Adjektive mit dem Koeffizienten von Rieger in der ersten und der euklidischen Metrik in der zweiten Assoziationsstufe

### 5.1.4 Wortarten

Es werden lemmatisierte Bedeutungswörter betrachtet, die nur aus Buchstaben bestehen, also keine Satz-, Sonderzeichen oder Ziffern enthalten, und mindestens zwei Zeichen lang sind. Sie werden nach automatisch und unüberwacht vergebenen Wortarttags selektiert. Falsch klassifizierte Lemmata sind daher zu erwarten. Solche, die in einer Stoppliste zu finden sind, werden deshalb gelöscht.

Die Untersuchungen werden jeweils nur für Substantive, Verben, Adjektive und für alle Bedeutungswörter zusammen durchgeführt. Verben gruppieren sich nicht, Adjektive fast nicht. Die Verben sind kaum unterscheidbar. Sie sind sich alle auf etwa gleichem Niveau fremd und gehören zu einem einzigen Cluster. Abbildung 5.6 zeigt, daß die niedrigsten Agglomerationsniveaus oberhalb der Hälfte des höchsten aller Niveaus liegen, dann aber alle Verben mit kleinen Niveauabständen zusammengefaßt werden. Bei den Adjektiven ist die wechselseitige Repugnanz etwas weniger stark ausgeprägt als bei den Verben. Es gibt sehr viele kleine und kleinste Gruppen zusammengehöriger Adjektive, die ineinander übergehen. Einige wenige Adjektive sind affin (Abbildung 5.7). Da Verben und Adjektive sich zum Clustern nicht eignen, sind die weiteren Untersuchungen auf Substantive beschränkt. Die Substantive bilden entsprechend ihrer Verwendungsregularitäten im zugrundeliegenden Korpus Cluster aus.

### 5.1.5 Gewichtung der Terme

Die Gewichtung der Terme mit der einfachen inversen Texthäufigkeit und der klassischen inversen Texthäufigkeit nach SALTON (Abschnitt 4.4.2.2) beeinflußt die Ergebnisse verschwindend gering. Es gibt keine wesentlichen Unterschiede zwischen den Ergebnissen mit den gewichteten und ungewichteten Häufigkeiten. Die durch die Gewichtung erhofften deutlicheren Ausprägungen von semantischer Abstoßung und Anziehung sind nicht eingetreten. Da die Gewichtung aber eine Veränderung der Daten ist, wird im weiteren auf sie verzichtet.

## 5.2 Clustertypen

Im semantischen Raum gibt es keine isolierten Gruppen. Es gibt aber auch keine zwei Bedeutungspunkte hoher Ähnlichkeit. Letzteres liegt im wesentlichen

[3] nur Terme mit einer Texthäufigkeit nicht größer als etwa zwei Drittel der Textanzahl; hier wurde die Texthäufigkeit auf 4600 bei knapp siebentausend Texten beschränkt

[4] die tausend oder zweitausend Terme mit den größten Texthäufigkeiten

daran, daß mit steigender Zahl der Texte das Niveau der ersten Agglomerationen zunimmt. Anhand der Hierarchien können zwei extreme Typen von Clustern identifiziert werden. Sie werden im folgenden beschrieben. Alle anderen Cluster setzen sich aus diesen zusammen.

### 5.2.1 Terme allgemeiner Verwendung

Terme allgemeiner Verwendung sind solche, deren Verwendungsregularitäten so unspezifisch sind, daß sie schon als beinahe beliebig bezeichnet werden können. Das sind Terme, die in so vielen verschiedenen Syntagmen stehen, daß sie sich paradigmatisch nicht mehr abgrenzen können, oder solche Terme, die so selten vorkommen, daß von einer Regularität ihrer Verwendung gar nicht gesprochen werden kann. Die Terme dieser Gruppen fallen auf etwa gleichem, aber verhältnismäßig hohem Niveau zusammen (Abbildung 5.8). Der Spann aller Vereinigungen der Gruppe beträgt nur einen Bruchteil des niedrigsten Agglomerationsniveaus der Gruppe. Die Terme haben somit alle gleich wenig miteinander zu tun. Diese Gruppen von Termen allgemeiner Verwendung heißen Cluster vom Typ A.

### 5.2.2 Terme spezieller Verwendung

Terme spezieller Verwendung sind solche, die aufgrund ihrer Verwendungsregularität Gruppen inhaltlich zusammengehöriger Terme ausbilden. Sie treten häufig in sehr speziellen Kotexten auf. Die Terme dieser Gruppen werden auf Agglomerationsstufen mit etwa gleichem Abstand zueinander zusammengefaßt (Abbildung 5.9). Diese Gruppen enthalten die ähnlichsten Termpaare. Das Niveauintervall der Vereinigungen ist ein Vielfaches des Agglomerationsniveaus der beiden ähnlichsten Terme der Gruppe. Die Gruppen von Termen spezieller Verwendung heißen Cluster vom Typ S. Cluster vom Typ S sind im allgemeinen kleiner als Cluster vom Typ A.

### 5.2.3 Andere Cluster

Es gibt Gruppen im ganzen Spektrum zwischen den beiden Extremen. Die Typ-A-Cluster schließen sich mit steigendem Agglomerationsniveau den weniger heterogenen Clustern an. Ausgeprägte Typ-S-Cluster setzen sich von diesen jedoch meist ab. Diese Gruppenarten sind in der Average-Linkage-Hierarchie mit der euklidischen Metrik besonders deutlich ausgeprägt.

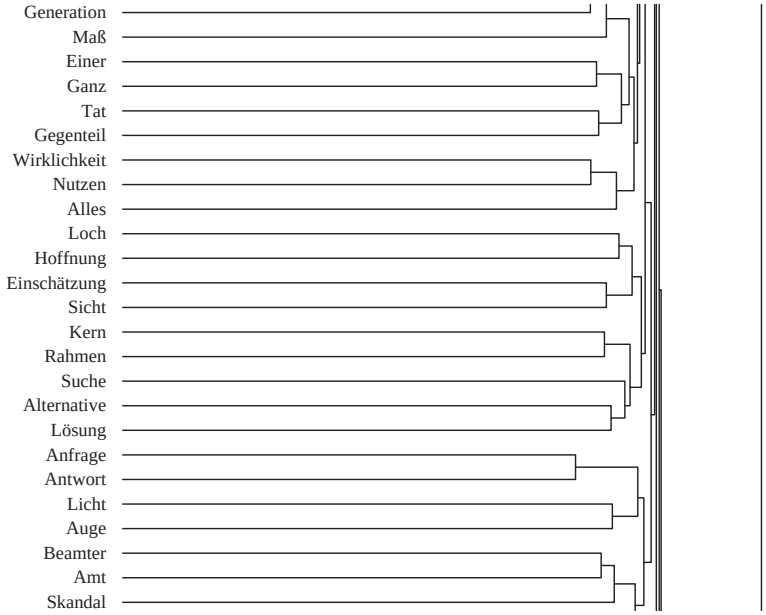



**Abbildung 5.8:** Ausschnitt einer Gruppe vom Typ A im Dendrogramm von Average-Linkage mit der euklidischen Metrik der Bedeutungspunkte der Substantive mit dem Maß von Rieger in der syntagmatischen und dem Maß von Tanimoto in der paradigmatischen Assoziationsstufe

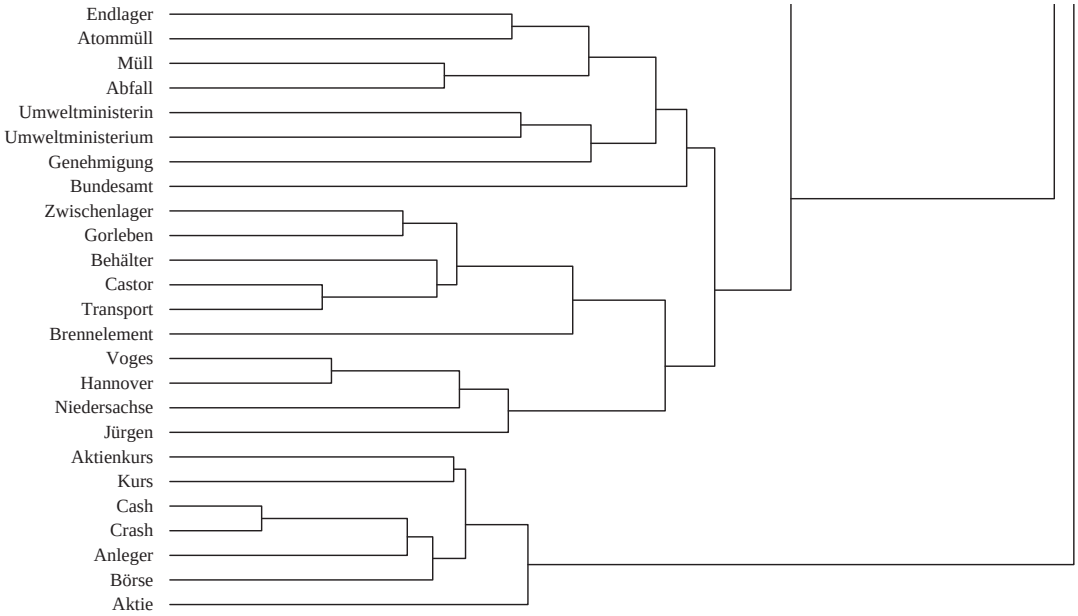



**Abbildung 5.9:** Typ-S-Gruppen im Dendrogramm von Average-Linkage mit der euklidischen Metrik der Bedeutungspunkte der Substantive mit dem Maß von Rieger in der syntagmatischen und dem Maß von Tanimoto in der paradigmatischen Assoziationsstufe

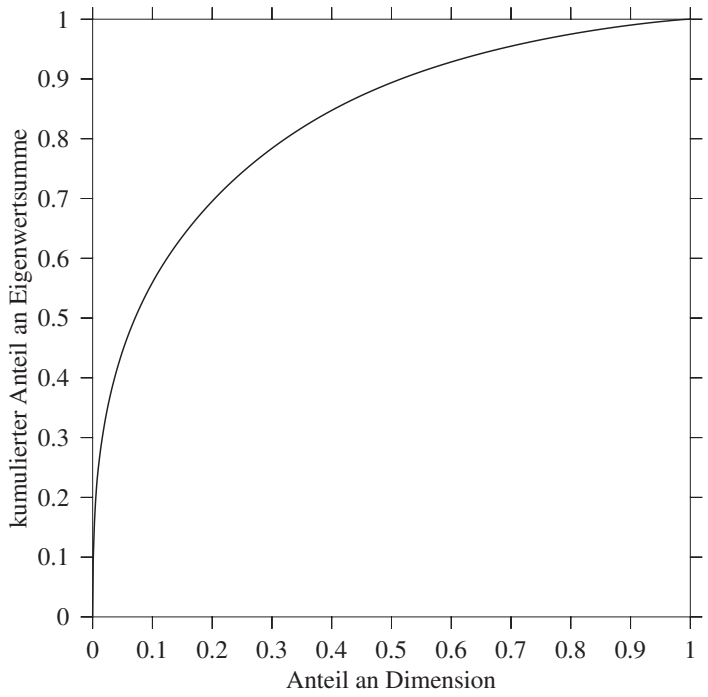



**Abbildung 5.10:** Konzentration der größten Eigenwerte der linken Singulärvektoren

## 5.3 Dimensionsreduktion

Die Zahl der Dimensionen kann verkleinert werden, um den Rechenaufwand bei der Verarbeitung der mit bis zu mehreren Millionen Komponenten umfassenden Term-Dokument-Matrizen zu verringern. Die Term-Dokument-Matrix, die zu dem durch die Singulärwertzerlegung berechneten Vektorraum mit den Singulärvektoren als Basis gehört, ist kleiner, aber nicht alle Assoziationsmaße sind für diese Termvektoren geeignet. Alle Maße, die an fixen Punkten in Koordinaten der kanonischen Basis ausgerichtet sind, können nicht mehr verwendet werden. In diese Richtung wurde die Untersuchung deshalb nicht weitergeführt.

Es wurde – wie bei der Termgewichtung – vermutet, daß eine Reduktion auf das Wesentliche zu deutlich ausgeprägteren Ergebnissen führt, ohne allzu große Fehler hinnehmen zu müssen. Tatsächlich konzentrieren sich die Streuungen der Datensätze in den Singulärrichtungen mit den größten Singulärwerten, so daß die Voraussetzung für eine zumindest rein technisch sinnvolle Dimensionsreduktion mit Hilfe der Singulärwertzerlegung gegeben ist (Abbildung 5.10). Allerdings führt die Streichung der Singulärvektoren, deren Singulärwerte am kleinsten sind, zu Informationsverlust und gröberen

Hierarchien, aber nicht zu besseren Ergebnissen. Auf je mehr Basisvektoren verzichtet wird, desto stärker nehmen die Typ-A-Strukturen zu Gunsten von Strukturen ab, die den Typ-S-Strukturen formal äußerlich ähneln, aber nicht wie diese inhaltlich interpretiert werden können. Die Gesamtstruktur nähert sich solchen Strukturen, die auch durch äußere Isolation beschrieben werden können. Da die Qualität der Ähnlichkeiten mit abnehmender Dimension ebenfalls abnimmt, sind deren Ergebnisse aber wenig aufschlußreich. Obwohl Landauer und Dumais (1997) das Latent-Semantic-Indexing allgemein als ergebnisverbessernd erachten, ist es hier inhaltlich semantisch nicht anwendbar.

## 5.4 Termassoziationsmaße

### 5.4.1 Syntagmatische Ähnlichkeiten

Die syntagmatischen Ähnlichkeiten berechnet mit dem Korrelationskoeffizient von Rieger und mit dem empirische Korrelationskoeffizient unterscheiden sich nur geringfügig. Bei der Messung der Ähnlichkeit mit dem Maß von Rieger sind die syntagmatisch homogensten Gruppen solche, die sich nach außen von allen anderen trennen, intern aber vergleichsweise heterogen sind. Diese Gruppen enthalten die ähnlichsten Termpaare, weisen aber auf den weiteren Agglomerationsstufen etwa gleich große Zuwächse auf. Diese Gruppen bestehen aus Termen, die nur in bestimmten Kotexten vorkommen. Die Gruppen, deren Terme nahezu auf gleichem Niveau zusammenfallen, bestehen aus Termen, die im Korpus in unterschiedlichsten Zusammenhängen immer wieder miteinander vorkommen. Die meisten Terme gehören zu letzteren Gruppen.

Neben dem empirischen Korrelationskoeffizienten ist die Strukturierung mit dem Maß von Tanimoto der mit dem Koeffizienten von Rieger am nächsten, gibt aber die syntaktischen Verwendungsregularitäten nicht mehr so fein aufgelöst wieder. Typ-S-Cluster enthalten vereinzelt unangemessen zugeordnete Terme. Stärker noch ist dies beim Cosinusmaß zu beobachten. Das Assoziationsmaß (4.3) für Termgewichte aus dem Einheitsintervall aus Abschnitt 4.9.1.2 muß als ungeeignet erachtet werden, weil das negative Verhalten, welches auch beim Cosinusmaß auftritt, noch deutlicher ausgeprägt ist. So gibt es nicht nur einzelne unangemessen gruppierte Terme. Zudem agglomerieren einige Terme ihrem Zusammenvorkommen nach auf zu niedrigem Niveau.

Das Cosinusmaß für binäre Textzugehörigkeiten der Terme bestimmt erwartungsgemäß gröbere Ergebnisse als das allgemeine Cosinusmaß für reelle Gewichte. Die Einteilung der Terme in Gruppen ist deutlicher, prinzipiell aber gleich. Wider Erwarten sind die mit der spezielleren binären Variante (4.2) berechneten Ergebnisse besser als die des oben als ungeeignet erkannten allgemeineren Maßes (4.3) für Termgewichte aus dem Einheitsintervall. Die Gruppierungen entsprechen den Daten eher als die des Cosinus für binär gewichtete Vorkommen. Sie sind zwar grob, aber durchaus befriedigend. Das Maß von TANIMOTO für binäre Termgewichte ist in Bezug auf die Qualität vergleichbar mit dem binären Cosinusmaß.

### 5.4.2 Paradigmatische Ähnlichkeiten

Metriken und Assoziationsmaße liefern unterschiedliche Ergebnisse. Während Metriken nur einseitig die Distanzen bestimmen, werden bei Korrelationsmaßen beidseitig sowohl Abstoßung als auch Anziehung betrachtet. Die wesentlichen Termähnlichkeiten werden aber durch alle Maße angemessen beschrieben. Die MINKOWSKI-1-Metrik und die euklidische Metrik unterscheiden sich wenig. Die Gruppeneinteilung von TANIMOTO ist noch einfacher zu interpretieren als die der euklidischen Metrik. Das Maß von TANIMOTO bildet größere zusammenhängende Gruppen eines Themengebietes während die euklidische Metrik diese auf mehrere nicht direkt benachbarte Gruppen verteilt. Das Maß von TANIMOTO und der empirische Korrelationskoeffizient ziehen die Bedeutungspunkte im ganzen weiter auseinander als die MINKOWSKI-1- und -2-Metriken. Die Typ-A-Gruppen sind beim empirischen Korrelationskoeffizient schwächer ausgeprägt als bei TANIMOTO. Am größten sind die Abstände zwischen den Niveaus beim Cosinus. Auch wenn das Cosinusmaß sich zur Bestimmung der Termassoziationen erster Ordnung als weniger gut geeignet erwiesen hat, kann es zur Berechnung der Assoziationen zweiter Ordnung sinnvoll verwendet werden. Generell eignen sich Korrelationsmaße besser zur Berechnung der Termassoziationen zweiter Ordnung als Metriken. Wird in der ersten Stufe der Cosinus verwendet, sind Metriken in der zweiten Stufe generell nicht geeignet. Die richtige Kombination aus einem Assoziationsmaß erster und zweiter Ordnung ist wesentlich. Der Korrelationskoeffizient von RIEGER in der ersten Stufe und TANIMOTO in der zweiten Stufe sind zur Beschreibung der Bedeutungspunkte am besten geeignet.

### 5.4.3 Hierarchien der Bedeutungspunkte

Das WARD-Verfahren (Abbildung 5.11) stellt die Einteilung in Gruppen viel deutlicher heraus als Average-Linkage (Abbildung 5.12). Dadurch geht aber die Unterscheidung der Typ-A- und -S-Cluster verloren. Wegen der sonst nirgends so klaren Unterteilung in Gruppen, sind WARD-Dendrogramme am besten geeignet, Partitionen durch Schnitte im Dendrogramm zu gewinnen. Bei der vergleichsweise deutlichen Gruppierung darf nicht übersehen werden, daß auch hier die niedrigsten Agglomerationsniveaus für eine WARD-Hierarchie ungewöhnlich hoch sind. Das Median-Verfahren gibt lediglich die Grundstrukturen wieder. Die Zentroid-Methode ist ungeeignet, weil sie die Beziehungen zwischen den Bedeutungspunkten zu stark verzerrt.

Wird als paradigmatisches Assoziationsmaß eine Metrik verwendet, ist nur Average-Linkage mit der euklidischen Metrik zur hierarchisch-agglomerativen Clusterung geeignet. Bei den anderen nichtmetrischen Assoziationsmaßen ist Single-Linkage weniger adäquat als Complete-Linkage. Die Complete-Linkage-Ergebnisse liegen denen von Average-Linkage zwar nahe, erreichen sie aber nicht. Das Maß von RIEGER in der ersten Stufe und das Maß von TANIMOTO in der zweiten Stufe berechnen so stabile Ergebnisse, daß alle hierarchisch-agglomerativen Verfahren die gleichen Grobstrukturen im semantischen Raum aufzeigen. Somit hat sich auch die aus Hierarchien erster Clusteranalysen abgeleitete Befürchtung, die Anzahl der Texte könne zu groß und damit die Ähnlichkeiten der Bedeutungspunkte zu gering sein, um sie unüberwacht klassifizieren zu können, als falsch erwiesen. Es muß nur die richtige Kombination gefunden werden: die syntagmatischen Ähnlichkeiten mit dem Korrelationskoeffizienten von RIEGER, die paradigmatischen Ähnlichkeiten mit dem Ähnlichkeitsmaß von TANIMOTO und die Gruppierung der Bedeutungspunkte mit dem hierarchisch-agglomerativen WARD-Verfahren mit der euklidischen Metrik. Hard-$c$-Means bestätigt unter Vorgabe der entsprechenden Clusteranzahl die aus der WARD-Hierarchie ausgewählten Partitionen.

### 5.4.4 Clusterprototypen

Aus je mehr Texten die Bedeutungspunkte berechnet werden, desto unähnlicher werden sie. Selbst die größten Ähnlichkeiten sind vergleichsweise gering, weil

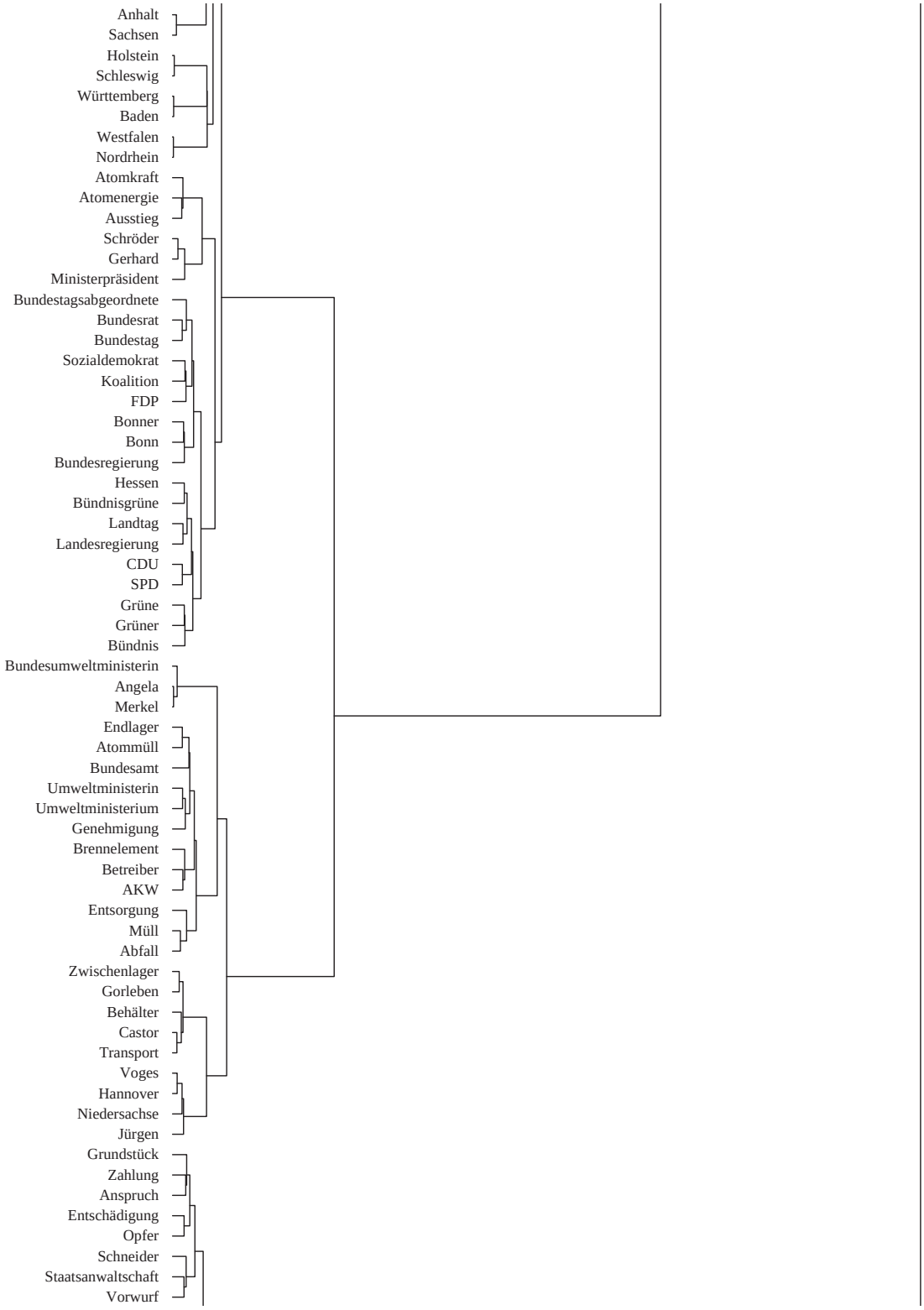



**Abbildung 5.11:** Ausschnitt des Dendrogramms von WARD mit der euklidischen Metrik der Bedeutungspunkte der Substantive mit dem Maß von RIEGER in der ersten und dem Maß von TANIMOTO in der zweiten Assoziationsstufe

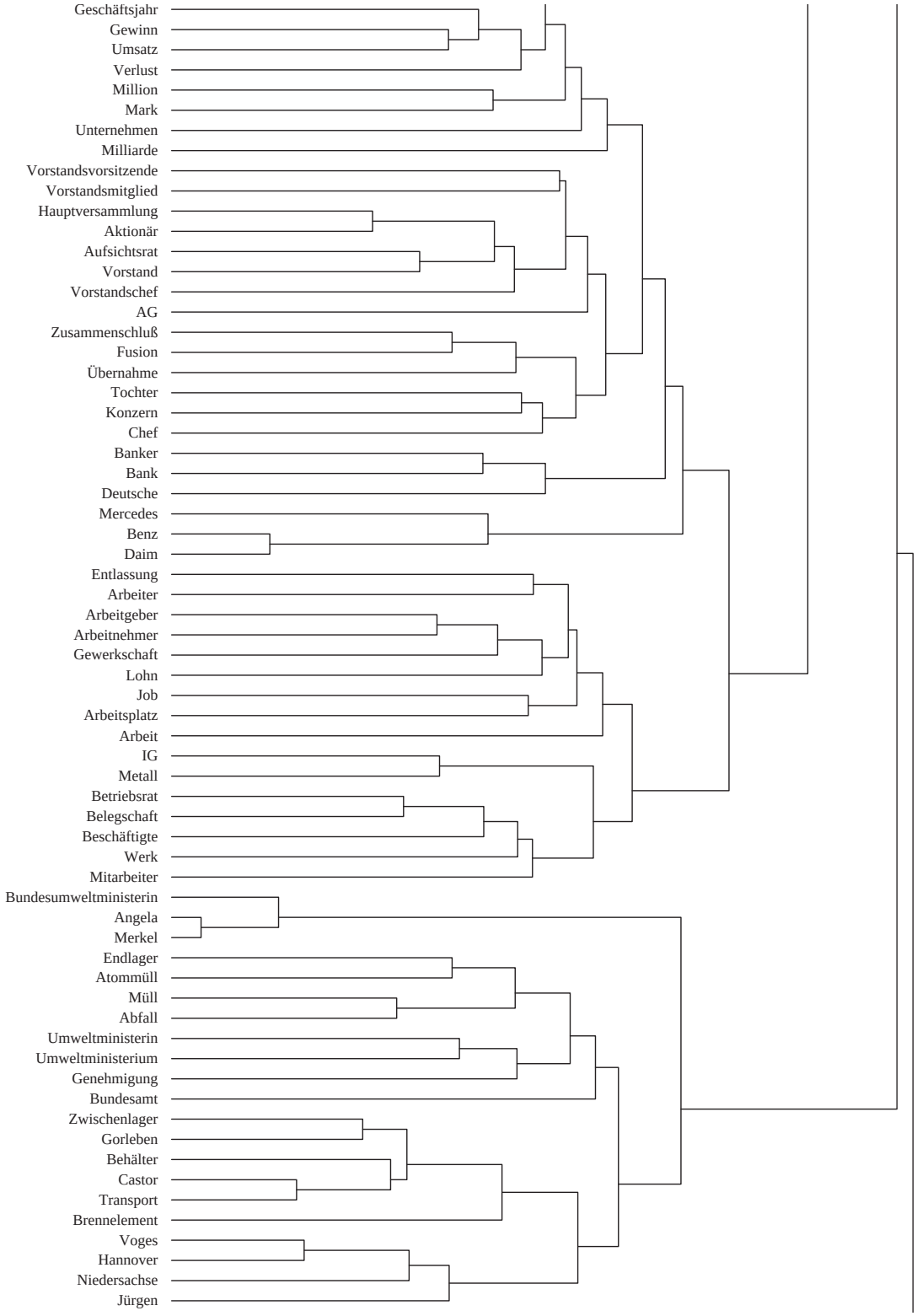



**Abbildung 5.12:** Ausschnitt des Average-Linkage-Dendrogramms mit der euklidischen Metrik der Bedeutungspunkte der Substantive mit dem Maß von RIEGER in der ersten und dem Maß von TANIMOTO in der zweiten Assoziationsstufe

die Anzahl der Texte die Anzahl der Terme um ein Vielfaches übersteigt[5]. Im semantischen Raum zeigt sich diese allgemeine Heterogenität dadurch, daß die kleinsten Abstände sehr groß im Vergleich zu den größten Abständen sind. Die Distanzen konzentrieren sich in einem Intervall, welches verglichen mit der Spannweite aller Distanzen recht kurz ist. Deshalb ist die Wahl eines geeigneten Radius für die Mountain-Function schwierig. Die in Abschnitt 3.13.1 vorgestellte Vorgehensweise zur Ermittlung des Radius berechnet einen Radius der etwa halb so lang ist wie die längste Distanz. Dieser Radius ist so groß, daß bei der Berechnung der Dichten mit einer Mountain-Function jeder Punkt zur Dichte eines jeden anderen beiträgt. Damit beim Subtractive-Clustering wenigstens lokale Unterschiede der Dichten erkannt werden können, muß deshalb ein anderer, viel kleinerer Radius manuell gewählt werden. Ist der vorgegebene Radius zu groß, werden nämlich alle Terme zu einem Cluster zusammengefaßt. Das läßt nur den Schluß zu, daß es nur eine ausgeprägte Punkthäufung gibt.

### 5.4.5 k-Nearest-Neighbors

Bei euklidischen Abständen erkennt das selbststabilisierende $k$-Nearest-Neighbors-Verfahren aus Abschnitt 3.14 bei vorgegebenen kleinen Radien für die Dichtemessung viele kleine Cluster. Die Clusteranzahl ist sehr sensibel bezüglich der Länge der Radien. Schon kleine Veränderungen des Radius verändern die Anzahl der Cluster erheblich. Werden die Distanzen mit der MINKOWSKI-1-Metrik berechnet oder die zu Distanzen überführten Ähnlichkeiten des TANIMOTO-Maßes (Abschnitt 2.17.4) verwendet, ist die Anzahl der Cluster klein, aber stabil bezüglich kleiner Veränderungen des Radius. Auch wenn diese grobe Zerlegung in nur wenige Cluster aus den Hierarchien der klassischen

[5] Wenige Terme werden in vielen Texten betrachtet. Wenige Untersuchungseinheiten werden aufgrund vieler Merkmale verglichen. Das scheint sehr ungewöhnlich. Es ist im allgemeinen nämlich üblich, viele Untersuchungseinheiten anhand von einigen wenigen Merkmalen zu vergleichen. Der Unterschied dieser beiden Ansätze, wird deutlich, wenn nur binäre Textzugehörigkeiten betrachtet werden. Dann nimmt bei fester Zahl an Termen mit steigender Anzahl der Texte die kombinatorische Möglichkeit zu, so daß die binären Termvektoren weniger übereinstimmende Komponenten haben. Ist die Zahl der Texte größer als die der Terme, ist es sogar rein theoretisch möglich, daß in einem Text kein Term mit einem anderen zusammen vorkommt. Ist die Zahl der Texte hingegen kleiner als die der Terme, muß es Übereinstimmungen geben, wodurch Ähnlichkeiten erzwungen werden. Generell – auch bei beliebigen Merkmalsskalen – nimmt die Ähnlichkeit der Terme bei Verringerung der Textanzahl zu. Dadurch werden zwar rein technisch die Clustereinteilungen deutlicher, aber auch die Datengrundlage so dünn, daß Regularitäten der Termverwendung nicht mehr festgestellt werden können und die Gruppierung der Terme deshalb nicht mehr inhaltlich als Bedeutungsähnlichkeit interpretiert werden darf.

agglomerativen Verfahren herausgelesen werden kann, so sind die Cluster doch zu groß, um die Terme angemessen zu unterteilen. Die Ursache hierfür liegt in den oben genannten Dichteverhältnissen im semantischen Raum.

### 5.4.6 Unscharfe Verfahren

Obwohl die Gruppen, wie aus den Dendrogrammen der hierarchisch-agglomerativen Verfahren ersichtlich ist, sich nur schwach voneinander abgrenzen und ineinander übergehen, strukturieren unscharfe Clusteranalyseverfahren die Menge der Bedeutungspunkte nicht. Sie bestimmen unabhängig von der Initialisierung immer die maximal unscharfe Partition. Unscharfe Verfahren erlauben zwar eine nahezu beliebig feine Zuordnung zu den Clustern, sind aber nicht anwendbar. Sie setzen nämlich nicht nur voraus, daß dicht und dünn besetzte Gebiete ineinander übergehen, sondern auch daß es mindestens zwei Kerngebiete gegeben muß, unter denen die Zugehörigkeiten aufgeteilt werden können.

Aus diesem Ergebnis und der Schwierigkeit, mit Hilfe des Subtractive-Clusterings geeignete Kandidaten für Clusterprototypen zu finden, muß geschlossen werden, daß der semantische Raum nicht wie bisher angenommen aus fließenden, dünnbesiedelten Übergängen zwischen dichten Schwerpunkten besteht. Vielmehr muß aus der maximal unscharfen Einteilung durch die unscharfen Verfahren und den inhaltlich höchst befriedigenden Hierarchien der scharfen Einteilung gefolgert werden, daß zwar Bedeutungspunkte ähnlich verwendeter Terme beieinander liegen, diese sich jedoch nach außen weder scharf noch unscharf als Gruppe identifizieren lassen. Nur der durch scharfe Verfahren vorgegebene Zwang zur Partitionierung führt zur Gruppenbildung. Hingegen ging Rieger (1983) noch von einer natürlichen scharfen Gruppeneinteilung aus. Bei Hard-$c$-Means zwingt die Minimierung der Streuung innerhalb der Cluster zur Partitionierung und bei den hierarchisch-agglomerativen Verfahren ist es der Betrachter, der den Schnitt im Dendrogramm aufgrund seines Expertenwissens macht. Semantische Räume sind eher der Art wie in Abbildung 3.2 (c) und nicht wie für unscharfe Algorithmen notwendig in (d). Die Punkte werden nur durch inneren Zusammenhalt gebunden. Getrennt werden sie dort, wo das Kriterium den schwächsten Zusammenhalt erkennt.

# Kapitel 6

# Zusammenfassung und Ausblick

## 6.1 Resümee

In dieser Arbeit wurden im Rahmen der deskriptiven und explorativen Datenanalyse Vorgehensweisen erprobt und Bedingungen ermittelt, durch die aus einem Korpus pragmatisch homogener Texte ein semantischer Raum berechnet werden kann, in dem räumlich benachbarte Bedeutungspunkte auch Terme mit intuitiv ähnlichen Bedeutungen repräsentieren. Als Korpus wurde eine Sammlung von Zeitungstexten verwendet. Zu jedem Wort wurde sein Lemma und die dem Kotext entsprechende Wortart bestimmt. Die Texte wurden auf Bedeutungswörter reduziert, weil nur diese lexikalische Bedeutungen tragen.

Ziel dieser Arbeit war es ursprünglich nur, ein numerisches Maß zur Bestimmung der Bedeutungsähnlichkeiten von Termen in natürlichsprachlichen Texten aufgrund der topologischen Nachbarschaften von Bedeutungspunkten in semantischen Räumen zu entwickeln. Ausgangspunkt einer solchen Untersuchung sind bereits vorgefertigte semantische Räume. Im Laufe der Forschung hat sich jedoch gezeigt, daß schon die Assoziationsmaße zur Berechnung semantischer Räume im Vektorraummodell substantiell die resultierenden Beziehungsstrukturen beeinflussen. Es ist zudem festgestellt worden, daß sich nicht alle Terme als Untersuchungseinheiten und auch nicht alle Texte als Merkmale auf gleiche Weise eignen. Der Problemfokus hat sich somit verschoben. Zum einen hat er sich erweitert auf die Suche nach denjenigen Voraussetzungen, die es erlauben, semantische Räume zu erstellen, die die Bedeutungsähnlichkeiten der Terme zutreffend wiedergeben. Zum anderen wurde die zweistufige Berechnung semantischer Räume selbst zu einem Forschungsschwerpunkt.

Die Vorkommenshäufigkeiten der Terme wurden in allen Texten gezählt. Da das Zusammenvorkommen der Terme in den Texten deren Bedeutungsähnlichkeiten im semantischen Raum bestimmt, wurden ausschließlich Terme untersucht, deren Vorkommenshäufigkeiten eine solche Bestimmung überhaupt sinnvoll erlauben. Dazu gehören einerseits nur solche Terme, die häufig genug vorkommen, um überhaupt von einer Regularität ihrer Verwendung sprechen zu können, und andererseits nur diejenigen, die nicht in fast allen Texten vorkommen, weil sie mit nahezu allen anderen Termen assoziiert und somit in ihrer Verwendung zu unspezifisch sind. Ebenso wurden Texte mit zu geringem oder zu großem Termumfang aus dem Korpus eliminiert, weil in ihnen zu wenige oder zu viele Terme zusammen vorkommen. Die Textlängen wurden mittels Termgewichtung normiert, damit Texte unterschiedlicher Länge miteinander verglichen werden können.

Die ausgewählten Terme wurden entsprechend ihrer Vorkommenshäufigkeiten in den Texten des Korpus gewichtet. Die Gewichtung der Terme hat jedoch nicht dazu geführt, daß die wesentlichen Strukturen im semantischen Raum deutlicher hervortreten. Es wurde außerdem untersucht, ob die Dimension der Daten mit Hilfe des Latent-Semantic-Indexing reduziert werden kann. Die grundlegenden Strukturen treten dadurch aber ebenfalls nicht klarer hervor. Beide Methoden wurden nicht verwendet, weil sie nicht nur zu keiner Verbesserung führen, sondern obendrein auch noch die Daten verändern.

Ein weiterer Untersuchungsschwerpunkt war die numerische Spezifizierung der Termassoziationen aufgrund der Verwendungsregularitäten der Terme im ganzen Korpus. Die Bedeutungskonstitution der Terme wurde durch eine zweistufige, den linguistischen Grundrelationen der syntagmatischen und paradigmatischen Beziehungen entsprechende Prozedur im Vektorraummodell zweiter Ordnung modelliert. In der ersten Stufe wird für jedes Paar von Termen in jedem Text eine lokale syntagmatische Ähnlichkeit als Funktion der Vorkommenshäufigkeiten bestimmt und diese zu einer globalen, numerisch spezifizierten Ähnlichkeit aggregiert. Dazu können nur solche Maße verwendet werden, die die lokale Abwesenheit beider Terme gar nicht, das Vorkommen beider Terme in einem Text als Affinität und das Vorhandensein nur eines der beiden Terme jeweils als Repugnanz werten. In der zweiten Stufe werden die paradigmatischen Assoziationen aus den syntagmatischen Assoziationen der ersten Stufe durch Vergleich der Kotexte der Terme ermittelt. Dabei bestehen keine generellen Einschränkungen bezüglich des Maßes. Es ist weniger wichtig, daß die Maße

mathematisch wünschenswerte Eigenschaften besitzen. Sie müssen vor allem inhaltlich interpretierbar und nachvollziehbar sein. Nur mit Assoziationsmaßen, die sowohl Affinität als auch Repugnanz zweier Terme berücksichtigen, konnten Termassoziationen angemessen modelliert werden. Die Güte eines semantischen Raumes wird bestimmt durch die Assoziationsmaße erster und zweiter Ordnung im Vektorraummodell. Es hat sich gezeigt, daß das im allgemeinen gern verwendete Cosinusmaß weniger gute Bedeutungsräume aufbaut als der RIEGERsche Korrelationskoeffizient als syntagmatisches Assoziationsmaß und das Maß von TANIMOTO als paradigmatisches Assoziationsmaß.

Die hohe Dimensionalität des semantischen Raumes verbietet die direkte Visualisierung und erlaubt nur mittelbare Schlußfolgerungen auf seine Struktur. Im semantischen Raum werden Terme ähnlicher Bedeutung durch Punkte repräsentiert, die räumlich nahe beieinander liegen. Aus früheren Untersuchungen war bekannt, daß semantische Räume nicht aus getrennten, scharfen Clustern bestehen. Deshalb wurde angenommen, daß Cluster unscharf ineinander übergehen. Unscharfe Clusteranalyseverfahren sind jedoch Verfahren zur Verbesserung einer Anfangspartition, die entweder durch die Vorgabe der Clusterprototypen oder einer Startpartition initialisiert werden müssen. Auf jeden Fall muß stets die Anzahl der Cluster vorgegeben werden. Zudem zwingen diese Verfahren die in ihren Kriterien definierten Struktureigenschaften innerer Homogenität und äußerer Heterogenität den Daten auch dann auf, wenn diese nicht den Strukturen entsprechen, die den Daten eigen sind. Deshalb wurde hier ein neues Clusterverfahren mit einem intrinsischen, sich einzig auf den inneren Zusammenhalt der Elemente beziehenden Kriterium entwickelt. Es ist ein auf $k$-Nachbarschaften und physikalischen Dichten basierendes agglomeratives Verfahren, welches eine scharfe Partition und somit auch die Anzahl der Cluster selbsttätig ermittelt. Dieses Verfahren wurde auf Datenmengen mit unscharfen Gruppierungen, die die Grenzen vorhandener unscharfer Clusteranalyseverfahren aufzeigen, erfolgreich getestet. Es ermittelt die Cluster anhand der Punktdichten in den Umgebungen der Punkte und den Abständen der Punkte des semantischen Raumes untereinander. Auch der Radius für die Dichtemessung, der die Nachbarschaften der Punkte bestimmt, wird automatisch ermittelt. Der einzige verbleibende freie Parameter ist eine unscharfe Menge, mit der die Dichten aus den Punktabständen berechnet werden. Die Z-Funktion ist ein vergleichsweise einfacher, geeigneter Wert für diesen Parameter. Bei diesem Verfahren ist das Element mit der höchsten Dichte der Prototyp eines Clusters. Ein Cluster

gilt als stabil, wenn die meisten nächsten Nachbarn seines Prototyps im Cluster selbst liegen. Ausgehend von der feinsten Partition faßt das Verfahren sukzessive benachbarte unstabile Cluster zusammen, bis nur noch stabile Cluster übrigbleiben und ist dadurch selbststabilisierend. Das Verfahren zerlegt den Datensatz in eine scharfe Partition. Es erkennt zwar die Prototypen, aber die scharfe Partitionierung eines unscharfen Datensatzes führt zu unbefriedigenden Zuordnungen der Randpunkte. Eine nachträgliche Fuzzifizierung durch ein unscharfes Analyseverfahren überführt die scharfe in eine unscharfe Partition, in der auch die Randpunkte mit entsprechenden Zugehörigkeiten versehen sind.

Allerdings hat sich dieses Verfahren für semantische Räume nicht als geeignet erwiesen. Ebenso konnten mit dem Subtractive-Clustering keine geeigneten Clusterschwerpunkte zur Initialisierung unscharfer $c$-Means-Varianten bestimmt werden. Hingegen sind Gruppierungen mit klassischen hierarchisch-agglomerativen Verfahren und $c$-Means möglich. Die Einteilungen entsprechen den Bedeutungen der Terme, so wie sie sich aus deren Verwendung im zugrundeliegenden Korpus ergeben. Dennoch optimieren unscharfe Iterationsverfahren diese scharfen Partitionen, die aus den Hierarchien der klassischen agglomerativen Verfahren durch einen Schnitt im Dendrogramm gewonnen werden, nicht zu geeigneten unscharfen Einteilungen. Folglich bestehen semantische Räume nicht aus mehreren dichten Punkthäufungen, die durch Bereiche geringer Punktdichte ineinander übergehen. Sie sind nicht durch unscharfe Partitionen beschreibbar. Ein globales, intrinsisches Maß basierend auf der Strukturvorstellung von scharfen oder unscharfen Gruppen in semantische Räume kann also gar nicht gefunden werden. Jedoch wurde gezeigt, daß mit den richtigen Assoziationsmaßen semantische Räume erzeugt werden, in denen Punkte, die Terme ähnlicher Bedeutung repräsentieren, benachbart sind. Die Punkte bilden aber keine Gruppen, weder scharf noch unscharf, aus. Der innere Zusammenhalt allein läßt sich clusteranalytisch nicht nutzen, um Terme zu semantisch homogenen Gruppen zusammenzufassen. Erst zusätzlich durch den von scharfen Verfahren vorgegebenen Zwang zu äußerer Abgrenzung werden Gruppen gebildet. Dadurch werden die Punkte künstlich dort getrennt, wo die interne Homogenität am schwächsten ist. Folglich entstehen semantisch intuitiv plausible Partitionen.

## 6.2 Ausblick

Semantische Räume können durch globale Maße strukturiert werden. Mit einer intrinsischen Nachbarschaftsdefinition allein gelingt dies nicht. Erst die Erweiterung um externe Heterogenität erzwingt die Bildung von Gruppen. Zudem können semantische Räume nicht durch unscharfe Partitionen beschrieben werden. Sie bestehen nicht aus mehreren unscharf ineinanderübergehenden Punkthäufungen. Es gibt also keine, weder scharfe noch unscharfe, natürliche Gruppen. Dennoch liegen Punkte ähnlicher Bedeutung zusammen. Die Untersuchungsergebnisse geben Anlaß zu folgender begründeten, in zukünftigen Arbeiten zu überprüfenden Vermutung: Ein semantischer Raum besteht aus einem dichten Kern von Termen allgemeiner Verwendung. Mit größer werdender Entfernung und abnehmender Dichte von diesem Kern liegen Terme mit zunehmend speziellerer Verwendung. Die Annahme, daß es nur eine ausgeprägte Punkthäufung gibt, erklärte auch das Scheitern unscharfer Clusteranalyseverfahren, weil diese mehrere solcher Punkthäufungen voraussetzen. Die verschiedenen Richtungen, in denen die Terme spezieller Verwendung liegen, entsprechen unterschiedlichen Bedeutungsrichtungen. Dabei ist zu beachten, daß es in den hochdimensionalen semantischen Räumen für jeden Term eine eigene Richtung gibt. Deshalb ist es zumindest theoretisch möglich, daß alle Terme nichts miteinander zu tun haben. Diese Eigenschaft erlaubt es allerdings auch zu modellieren, daß ein einzelner Term in unterschiedlichsten Kotexten gebraucht werden kann. Aber eben diese Eigenschaft eines Termes, Bedeutungsähnlichkeit zu anderen Termen unterschiedlichster Verwendung haben zu können, kann durch eine globale Strukturierung mit einem für alle Terme gleichen Maß nicht erfaßt werden. Damit stellt sich die Aufgabe, die Bedeutungsähnlichkeiten lokal aus der Sicht eines jeden einzelnen Termes zu beschreiben.

# Literaturverzeichnis


ANDERSON, EDGAR: The irises of the Gaspe peninsula. *Bulletin of the American Iris Society*, Bd. 59, S. 2–5, 1935

BACHER, JOHANN: *Clusteranalyse. Anwendungsorientierte Einführung.* Oldenbourg, München, 1994

BACKER, ERIC: *Computer-assisted reasoning in cluster analysis.* Prentice Hall, New York, 1995

BANDEMER, HANS und NÄHTER, WOLFGANG: *Fuzzy data analysis.* Kluwer, Dordrecht, 1992

BEZDEK, JAMES C.: *Pattern recognition with fuzzy objective function algorithms.* Plenum Press, New York, 1981

BEZDEK, JAMES C.: Partition Structures: A Tutorial. In: *Analysis of Fuzzy Information*, hrsg. von JAMES C. BEZDEK, Kap. 6. CRC Press, Boca Raton/Florida, 1987

BEZDEK, JAMES C.: Fuzzy clustering. In: *Handbook of Fuzzy Computation*, hrsg. von ERIQUE H. RUSPINI, PIERO P. BONISSONE, und WITOLD PEDRYCZ, Kap. F6. Institute of Physics Publishing, Bristol/Philadelphia, 1998

BRONSTEIN, ILJA N., SEMENDJAJEW, KONSTANTIN A., MUSIOL, GERHARD, und MÜHLIG, HEINER: *Taschenbuch der Mathematik.* Harri Deutsch, Thun, Frankfurt am Main, zweite Aufl., 1995

BUSSMANN, HADUMOD: *Lexikon der Sprachwissenschaft.* Kröner, Stuttgart, 2. Aufl., 1990

CALINSKI, T. und HARABSZ, J.: A dendrite method for cluster analysis. *Communications in Statistics*, Bd. 3, Nr. 1, S. 1–27, 1974

CHIU, S. L.: Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems*, Bd. 2, Nr. 3, S. 267–278, 1994

CLIFFORD, HAROLD TREVOR und STEPHENSON, WILLIAM: *An introduction to numerical classification*. Academic Press, New York, 1975

DEIMER, REINHARD: *Unscharfe Clusteranalysemethoden. Eine problemorientierte Darstellung zur unscharfen Klassifikation gemischter Daten.* Schulz-Kirchner, Idstein, 1986

DIEHL, JOERG M. und KOHR, HEINZ U.: *Deskriptive Statistik*. Fachbuchhandlung für Psychologie, Frankfuhrt am Main, 5. Aufl., 1983

DUDA, RICHARD O. und HART, PETER E.: *Pattern classification and scene analysis*. Wiley, New York, 1973

DUNN, J. C.: A fuzzy relative of the ISODATA process and its use in detecting compact, well-seprated clusters. *Journal of Cybernetics*, Bd. 3, Nr. 3, S. 32–57, 1974

ECKES, THOMAS und ROSSBACH, HELMUT: *Clusteranalysen*. Kohlhammer, Stuttgart, 1980

EISENREICH, GÜNTHER: *Lineare Algebra und analytische Geometrie*. Akademie-Verlag, Berlin, 1980

EVERITT, BRIAN S.: *Cluster Analysis*. Arnold, London, 3. Aufl., 1993

FERSCHL, FRANZ: *Deskriptive Statistik*. Physica, Würzburg, 1978

FERSCHL, FRANZ: *Deskriptive Statistik*. Physica, Würzburg, 3. Aufl., 1985

FISCHER, GERD: *Lineare Algebra*. Vieweg, Braunschweig, 9. Aufl., 1989

FLOREK, K., LUKASZEWICZ, J., PERKAL, J., STEINHAUS, H., und ZUBRZYCKI, S.: Sur la liaison et la division des point d'un ensemble fini. *Colloquium Mathematicum*, Bd. 2, S. 282–285, 1951

GATH, I. und GEVA, A. B.: Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 11, Nr. 7, S. 773–781, 1989

GITMAN, ISRAEL und LEVINE, MARTIN D.: An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique. *IEEE Transactions on Computers*, Bd. 19, Nr. 7, S. 583–593, 1970

GOLUB, GENE H. und VAN LOAN, CHARLES FRANCIS: *Matrix computations*. Johns Hopkins University Press, Baltimore, 2. Aufl., 1989

GORDON, ALLEN D.: *Classification. Methods for the exploratory analysis of multivariate data*. Chapman and Hall, London, 1981

GORDON, ALLEN D.: *Classification*. Chapman and Hall/CRC, Boca Raton, 2. Aufl., 1999

GOWER, J. C.: A comparison of some methods of cluster analysis. *Biometrics*, Bd. 23, S. 623–637, 1967

GOWER, J. C. und ROSS, G. J. S.: Minimum Spanning Trees and Single Linkage Cluster Analysis. *Applied Statistics*, Bd. 18, S. 54–64, 1969

GUSTAFSON, DONALD E. und KESSEL, WILLIAM C.: Fuzzy Clustering with a Fuzzy Covariance Matrix. In: *Proceedings of the IEEE Conference on Decision and Control, Conference on Decision and Control*, S. 761–766, 1978

HEILER, SIEGFRIED und MICHELS, PAUL: *Deskriptive und Explorative Datenanalyse*. Oldenbourg, München, 1994

HÖPPNER, FRANK, KLAWONN, FRANK, und KRUSE, RUDOLF: *Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifizierung und Datenanalyse*. Vieweg, Braunschweig, 1997

HORN, ROGER A. und JOHNSON, CHARLES R.: *Matrix analysis*. Cambridge University Press, Cambridge, 1985

HORN, ROGER A. und JOHNSON, CHARLES R.: *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1991

JAIN, ANIL K. und DUBES, RICHARD C.: *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs N.J., 1988

JOHNSON, LEE W., RIESS, R. DEAN, und ARNOLD, JIMMY T.: *Introduction to linear algebra*. Addison-Wesley, Reading Mass., 3. Aufl., 1993

KAUFMAN, LEONARD und ROUSSEEUW, PETER J.: *Finding groups in data. An introduction to cluster analysis*. Wiley, New York, 1990

KOECHER, MAX: *Lineare Algebra und analytische Geometrie*. Springer, Berlin, 2. Aufl., 1985

LANCE, G. N. und WILLIAMS, W. T.: A general theory of classificatory sorting strategies. 1. Hierarchical systems. *The Computer Journal*, Bd. 9, S. 373–380, 1967

LANDAUER, THOMAS K. und DUMAIS, SUSAN T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, Bd. 104, Nr. 2, S. 211–240, 1997

LAX, PETER D.: *Linear algebra*. Wiley, 1997

LAY, DAVID C.: *Linear algebra and its application*. Addison-Wesley, Reading Mass., 2. Aufl., 1997

LIPSCHUTZ, SEYMOUR: *Lineare Algebra. Theorie und Anwendung*. McGraw-Hill, London, 1977

LÜTKEPOHL, HELMUT: *Handbook of matrices*. Wiley, Chichester, 1996

MAGNUS, JAN R.: *Linear structures*. Griffin, London, 1988

MARRIOTT, F. H. C.: Optimization methods of cluster analysis. *Biometrika*, Bd. 69, Nr. 2, S. 417–421, 1982

MINNOTTE, MICHAEL C. und WEST, R. WEBSTER: The data image: a tool for exploring high dimensional data sets. In: *Proceedings of the American Statistical Association Section on Statistical Graphics*, 1998. Preprint

MIRKIN, BORIS: *Mathematical classification and clustering*. Kluwer, Dordrecht, 1996

MOJENA, R.: Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, Bd. 20, Nr. 4, S. 359–363, 1977

NERING, EVAR D.: *Linear algebra and matrix theory*. Wiley, New York, 2. Aufl., 1970

OAKES, MICHAEL P.: *Statistics for corpus linguistics*. Edinburgh University Press, Edinburgh, 1998

PRIM, R. C.: Shortest Connection Networks and Some Generalizations. *Bell System Technical Journal*, Bd. 36, S. 1389–1401, 1957

RIEGER, BURGHARD: Bedeutungskonstitution. Einige Bemerkungen zur semiotischen Problematik eines linguistischen Problems. In: *Semiotik. Anwendungen in der Literatur- und Textwissenschaft*, hrsg. von R. GUNZENHÄUSER, S. 55–68. Athenäum, Frankfurt am Main, 1977

RIEGER, BURGHARD: *Unscharfe Semantik. Die empirische Analyse, quantitive Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutung in Texten*. Lang, Frankfurt am Main, 1989

RIEGER, BURGHARD: Unscharfe Semantik. Zur numerischen Modellierung vager Bedeutungen von Wörtern als fuzzy-Mengen. In: *Forum '90 Wissenschaft und Technik. Neue Anwendungen mit Hilfe aktueller Computer-Technologien*, hrsg. von HANS-JÜRGEN FRIEMEL, G. MÜLLER-SCHÖNBERG, und A. SCHÜTT, S. 80–104. Springer, Berlin, 1990

RIEGER, BURGHARD B.: Clusters in semantic space. Analysing natural language texts to model word meaning as a procedural representation. In: *Actes du Congrès International Informatique et Science Humaines, Lièges (Laboratoire d'Analyse Statistique des Langues Anciennes)*, hrsg. von L. DELATTE, S. 805–814, 1983

RUGE, GERDA: *Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation*. Sprache und Computer. Georg Olms, Hildesheim, 1995

SALTON, G., WONG, A., und YU, C. T.: Automatic indexing using term discrimination and term precision measurements. *Information Processing And Management*, Bd. 12, S. 43–51, 1976

SALTON, GERARD: *Automatic Information Organization and Retrieval*. Computer Science Series. McGraw-Hill, New York, 1968

SALTON, GERARD: Another look at automatic text-retrieval systems. *Communications of the Association for Computing Machinery*, Bd. 29, Nr. 7, S. 648–656, 1986a

SALTON, GERARD: On the use of term associations in automatic information retrieval. In: *Proceedings of COLING '86*, S. 380–386. Institut für angewandte Kommunikations- und Sprachforschung, Bonn, 1986b

SALTON, GERARD: *Automatic text processing. The transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading Mass., 1989

SALTON, GERARD, ALLAN, JAMES, und BUCKLEY, CHRIS: Automatic structuring and retrieval of large text files. *Communications of the Association for Computing Machinery*, Bd. 37, Nr. 2, S. 97–108, 1994

SALTON, GERARD und BUCKLEY, CHRISTOPHER: Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, Bd. 24, Nr. 5, S. 513–523, 1988

SALTON, GERARD, FOX, EDWARD A., und WU, HARRY: Extended boolean information retrieval. *Communications of the Association for Computing Machinery*, Bd. 26, Nr. 12, S. 1022–1036, 1986

SALTON, GERARD und MCGILL, MICHAEL J.: *Introduction to modern information retrieval*. McGraw-Hill, 1983

SALTON, GERARD, WONG, A., und YANG, C. S.: A vector space model for automatic indexing. *Communications of the Association for Computing Machinery*, Bd. 18, Nr. 11, S. 613–620, 1975

SALTON, GERARD und WU, HARRY: A term weighting model based on utility theory. In: *Proceedings of the 3rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, S. 9–22. Cambridge, 1980

SALTON, GERARD et al.: Information storage and retrieval. Scientific report no. ISR-7 to National Science Foundation, Computation Laboratory of Harvard University, Cambridge Mass., 1964

SCHADER, MARTIN: *Scharfe und unscharfe Klassifikation qualitativer Daten*. Athenäum/Hain/Scriptor/Hanstein, Köngistein/Ts., 1981

SNEATH, P. H. A.: The application of computers to taxonomy. *Journal of General Microbiology*, Bd. 17, S. 201–226, 1957

SOKAL, ROBERT R. und SNEATH, PETER H. A.: *Principles of numerical taxonomy*. Freeman, San Francisco, 1963

SPÄTH, HELMUTH: *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion*. Oldenbourg, München, 1975

STEINHAUSEN, DETLEF und LANGER, KLAUS: *Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation*. de Gruyter, Berlin, 1977

TANIMOTO, T. T.: An elementary mathematical theory of classification and prediction. Techn. Ber., International Business Machines, New York, 1958

WARD, JOE H.: Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, Bd. 58, S. 236–244, 1963

XIE, XUANLI LISA und BENI, GERARDO: A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 13, Nr. 8, S. 841–874, 1991

YAGER, RONALD R. und FILEV, DIMITAR P.: Approximate clustering via the mountain method. *IEEE Transactions on Systems, Man, and Cybernetics*, Bd. 24, Nr. 8, S. 1279–1284, 1994a

YAGER, RONALD R. und FILEV, DIMITAR P.: Generation of fuzzy rules by mountain clustering. *Journal of Intelligent and Fuzzy Systems*, Bd. 2, S. 209–219, 1994b

ZADEH, LOTFI A.: Fuzzy sets. *Information and Control*, Bd. 8, Nr. 3, S. 338–353, 1965

# Index

# Erklärung

Hiermit versichere ich, daß ich diese Arbeit „Nachbarschaften im semantischen Raum“ selbständig und nur unter Benutzung der angegebenen Hilfsmittel angefertigt habe.

Trier, den 25. Juli 2006 Armin Wegner