Datamining:

Methoden integrativer Datenpräsentation

Jörg Andreas Walter

Habilitationsschrift
eingereicht an der Technischen Fakultät
der Universität Bielefeld
Juni 2003

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

“We are drowning in information but starving for knowledge.” (John Naisbett)

“Who[ever] has information fastest and uses it wins.” (Don Keough, früherer Präsident von Coca-Cola)

“Datamining will become more important, and companies will throw away nothing about their custumers because it will so be so valuable. If you’re not doing this you’re out of business.”

(Arno Penzias, Nobelpreisträger 1999)

Wir leben heute in einer Zeit des ungeheuren Wachsums digitaler Information. Schätzungen besagen, dass sich die verfügbare digitale Information alle 20 Monate und die schiere Speicherkapazität alle 9 Monate verdoppeln (Fayyad und Uthurusamy 2002). Einer der Gründe ist die exponentiell wachsende Verfügbarkeit von Rechen- und Speicherkapazität zu dramatisch fallenden Preisen – eine stetige Entwicklung, die schon von Moore (1965) quantifiziert wurde.

Die Verfügbarkeit von Information und die schnelle Umsetzung durch Extraktion von Wissen sind Schlüsselfaktoren in Entscheidungsprozessen – dies wurde von vielen auch als wettbewerbsrelevant erkannt.

Der Begriff *Datamining* ist in Analogie zum Bergbau, engl. *mining*, eingeführt worden. Die Kunst des Bergbaus besteht darin, wertvolle Erze und Mineralien in einer großen Menge von so genanntem „taubem“ Gestein aufzuspüren, zu identifizieren, herauszuarbeiten und zu bergen. Dazu ge-

hört, im Gestein zu navigieren, den reichen Flözen und Gängen zu folgen und das Gestein aufzuschießen, zu separieren, um letztlich die „Goldklumpen“ (*nuggets*) zu heben. In verschiedenen Bereichen treffen wir in unserer Zeit auf eine solche Situation an: riesige Datenhalden, die heute, dank winziger Speicherkosten, teilweise nie mehr gelöscht werden.

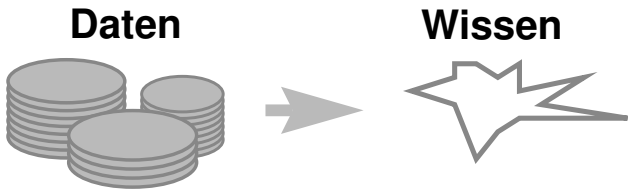


Abbildung 1.1: Die übergeordnete Aufgabe des *knowledge-discovery*-Prozesses ist: Wie gewinnt man interessantes Wissen aus einer möglicherweise erdrückenden Fülle von Daten?

So, wie die Aufgabe des Bergbaus darin besteht, „Gold“ aus dem großen Steingemenge zu bergen, so ist die Aufgabe des Dataminings, wertvolles Wissen in großen Datenmengen zu entdecken. Diese Aufgabe wird im Begriff *knowlegde discovery in databases* (**KDD**) noch stärker zum Ausdruck gebracht und zudem wird Bezug auf die typische Speicherform in Datenbanken genommen.

Nach der Definition von Fayyad et al. (1996) beschreibt ***Knowlegde discovery*** den *(i)* nicht-trivialen Prozess der Identifikation von *(ii)* validen, *(iii)* neuen, *(iv)* potentiell nützlichen und letztlich *(v)* verständlichen bzw. umsetzbaren *(vi)* Mustern und Regularitäten aus *(vii)* Datenbeständen.

Im Folgenden wird diese prägnante Beschreibung an verschiedenen Stellen weiter erläutert.

Die Nichttrivialität *(i)* bedeutet, dass *knowlegde discovery* in komplexen Domänen kein vollkommen automatisierbarer Vorgang ist. Neben teilautomatisierten Verfahren schließt er das Vorwissen und das Interpretationsvermögen von Experten der Anwendungsdomäne unverzichtbar mit ein. Daher ist die effiziente Integration des Menschen in den iterativen Prozess der Wissensgewinnung eine Schlüsselherausforderung. Dies impliziert Aspekte der Gestaltung von Benutzerschnittstellen (*human computer interface*, HCI) als auch der Einbindung der Experten in den KDD-Prozess.

Vorgehensmodell CRISP-DM

Um den KDD-Prozess als solchen besser zu planen und zu standardi-

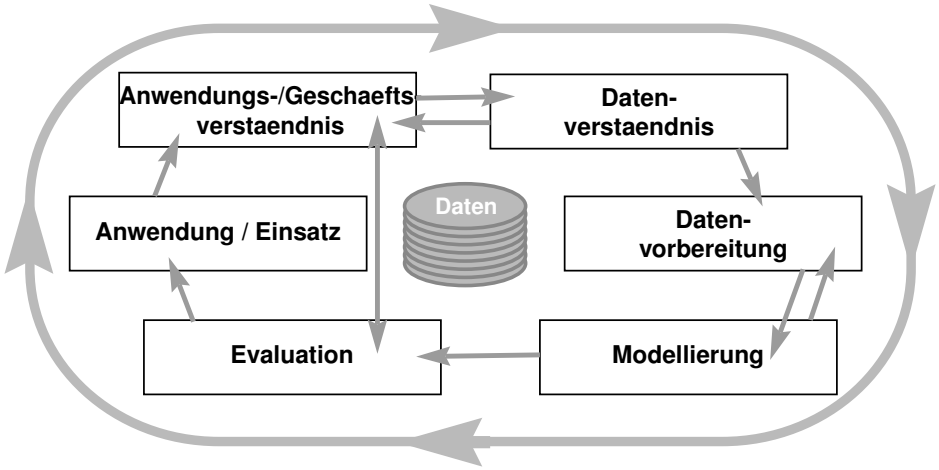


Abbildung 1.2: Der Datamining-Prozess nach dem CRISP-DM-Standard (*CRoss Industry Standard Process for Datamining*) durchläuft mehrere Teilschritte, wobei die Schrittfolge nicht als linear aufeinander folgend, sondern iterativ und situationsgerecht angepasst vorgesehen ist.

sieren, wurden einige Vorgehensmodelle entwickelt, von denen eines der verbreitetsten hier kurz dargestellt werden soll. Der *CRoss Industry Standard Process for Datamining* (**CRISP-DM**) wurde von einem internationalen Industriekonsortium (NCR, Daimler Chrysler, SPSS, OHRA, Chapman et al. 2000) entwickelt. Wie in Abb. 1.2 illustriert, beschreibt CRISP-DM sechs Kernbereiche eines Datamining-Projekts, deren zeitliche Abfolge nicht in sechs streng sukzessiven Phasen vorgeschlagen wird, sondern im Gegenteil, es wird ein iterativer und stark vernetzter Ablauf angelegt:

- Im ersten Schritt eines Projekts gilt es, zunächst ein Verständnis für die relevanten Vorgänge im Geschäfts- bzw. Anwendungsbereich zu gewinnen (*business understanding*). Da das KDD-Ziel ist, *neues* Wissen (*iii*) zu entdecken, ist es bedeutsam eine gute Orientierung darüber zu haben, welche Bereiche bekannt und welche potentiell relevant sind. Das Ergebnis soll eine klare Festlegung der Anforderungen, Projektziele (aus Geschäfts- bzw. Anwendungssicht) und Rahmenbedingungen beinhalten.

  Das Fokussieren dieser Ziele und das Stellen relevanter Fragen sind die wichtigsten Planungsschritte im Datamining-Prozess. Gute Fragen sind solche, deren Antwort irgendwo in den Daten verborgen liegt. Ihre Formulierung erfordert oft die Kooperation von verschiedenen Partnern: Man muss die Geschäfts- oder Wissensziele ver-

stehen, die verfügbaren Daten interpretieren und die Datamining-Verfahren der nachfolgenden Phasen kennen. Mitarbeiter, die in allen drei Bereichen kompetent sind, gibt es zwar ganz selten, aber sie sind äußerst wertvoll;

- der zweite Bereich fokussiert auf die vorhandenen bzw. noch zu beschaffenden Datenbestände. Dies beinhaltet die Sichtung, Dokumentation und überblicksartige Exploration der Daten. Die starke Interdependenz mit der Zielplanung wird durch die Pfeile in beide Richtungen ausgedrückt;

- die Datenvorbereitung schafft integrative Zugänge zu den Daten durch Zusammenführung der (möglicherweise weit verstreuten) Datensätze und deren Aufbereitung in geeignete Datenformate. Dies involviert auch die Beseitigung von Inkonsistenzen, die Behandlung fehlender Werte und die Erweiterung um abgeleitete, später hilfreiche Größen;

- die Modellbildung umfasst die Auswahl der verwendeten Algorithmen und Bewertungsverfahren, die meist iterative Gewinnung der Modelle und deren technische Bewertung;

- bei der Evaluierung steht, im Gegensatz zu vorher, die Beurteilung des gewonnenen Modells aus der Anwendungsperspektive im Vordergrund. Hierzu werden die gefundenen Muster adäquat präsentiert und vom Experten entsprechend der definierten Projektziele bewertet. Die Zwischenschritte werden überprüft und die nächsten Schritte geplant;

- wenn die Evaluierung erfolgreich ist, werden die Ergebnisse dokumentiert und ggf. ein Umsetzungsplan für das gefundene Wissen erstellt und ausgeführt. Dies kann zum Beispiel die Durchführung einer Werbekampagne (s.u.) sein oder die Integration einer Risikobewertung in ein klinisches Informationssystem (s. Kap. 9).

Anwendungsbeispiele für Datamining

Ein typisches und kommerziell sehr erfolgreiches Einsatzfeld für Datamining Verfahren ist der Bereich Marketing und Kundenbindung – (neudeutsch:) *Customer Relationship Management* (**CRM**).

Kampagnenplanung: Zur Optimierung der Marketingaktivität werden Kundendaten analysiert, um z.B. herauszufinden, welche typischen Kunden

profitabel sind oder werden könnten. Aus der Schätzung der Empfänglichkeit für eine bestimmte Werbemaßnahme wird genau jenes Kundensegment bestimmt, in das in einer Marketingaktion investiert wird, oder umgekehrt, nicht investiert wird.

Churn management ist eine dazu verwandte Anwendung mit umgekehrtem Vorzeichen: Für eine Telefongesellschaft gilt es z.B., der Kündigung eines Dienstleistungsvertrages gezielt entgegenwirken zu können. Indem man die mögliche Unzufriedenheit des Kunden abzuschätzen versucht, kann die Kundenbindung durch konkrete Problembehebung oder gezielte Freundlichkeiten (Anruf, Brief, spezielle Angebote etc.) wieder verbessert werden. Die Datenbasis kann sich hier auf die Vertragsdetails und die gesamte Beziehungshistorie mit den Kunden ausweiten.

Webshops: Die Entwicklung des Internets hat die vollautomatische Rund-um-die-Uhr Onlinevermarktung von Produkten und Services ermöglicht. Man mag annehmen, dass die Vermarktungskosten deutlich geringer sind als im klassischen Geschäft. Dies ist durchaus nicht immer so, aber es ergeben sich durch Datamining-Techniken ganz neue Möglichkeiten, Kunden durch Zusatzservices zu binden. Man unterscheidet zwei grundsätzliche Arten von Käuferverhalten. Der *Sucher* weiß, was er will – er kann online durch gute Datenbanksuchoptionen und -strategien hocheffizient bedient werden. Der *Spontankäufer* dagegen schlendert und wartet, bis er etwas Reizvolles findet, das bei ihm den nötigen Kaufimpuls auslöst. Ein reales Ladengeschäft dekoriert sein Sortiment und präsentiert es dem umherschweifenden Blick des Kunden. Ein Online-Webshop kann dies nicht annähernd in der Blickweite nachbilden, aber er kann versuchen, gleich *etwas Passendes* zu präsentieren.

Mit Datamining-Verfahren kann die Beratungsleistung eines persönlich vertrauten Kundenbetreuers nachgebildet werden. Aufgrund der Erfahrung bei vergangenen Besuchen werden individuelle Kaufempfehlungen unterbreitet, verbunden mit einem breiten Angebot von Zusatzinformation und verzugsloser, freundlicher Bedienung. Dies erfordert nicht nur eine geschickte Dialoggestaltung, in der die relevanten Daten unaufdringlich erhoben werden, sondern auch die Kombination mit evtl. demographischen Daten (z.B. straßenbezogene Schätzungen von sozio-ökonomischen Daten, Kaufkraft und Präferenzen, natürlich unter Wahrung der Datenschutzregelungen) und der Einbeziehung von aktuellen Kauftrends, die in der Gesamtkäuferpopulation sichtbar gemacht werden. Auch bei hohem Besucheraufkommen muss die Gesamtarchitektur der Datenverarbeitung ak-

zeptable Antwortzeiten für den Benutzer gewährleisten.

Sternkartierungsystem SKICAT: Ein gutes Datamining-Beispiel aus dem Bereich der Astronomie ist das SKICAT-System (Fayyad et al. 1996). In diesem Projekt handelt es sich um Digitalbilder vom Sternenhimmel der nördlichen Hemisphäre. Die in sechs Jahren im *2nd Palomar Observatory Sky Survey* (POSS-II) zusammengefügten Bilder umfassen 3 Terrabytes Daten, die eine geschätzte Zahl von 50 Millionen Galaxien, 2 Milliarden Sternen und 5 Tausend Quasaren abbilden. Diese Detailfülle ist für den Menschen kaum mehr erfassbar. Mittels handklassifizierten Bildern wurde das System so erfolgreich trainiert, dass selbst Objekte noch eingeordnet werden konnten, die vom Menschen nun mit höher auflösenden Bildern untersucht werden. Auf diese Weise wurde der bekannte Sternenkatalog um den Faktor 3 bereichert. Ohne diese automatisierten Datamining-Verfahren würde die Auswertung der Bilder nicht nur Jahrzehnte manueller Arbeit in Anspruch nehmen, sondern auch weniger Resultate liefern.

Dies ist eine zunehmend häufige Situation im wissenschaftlichen Bereich: Instrumente können heute problemlos riesige Volumina von Daten generieren – Produktionsraten in der Höhe von Gigabytes pro Stunde sind kein Problem. Damit wächst die Lücke zwischen den Möglichkeiten, Daten zu sammeln, und den Möglichkeiten, sie zu analysieren und als Mensch zu begreifen.

Ziel und Gliederung der Arbeit:

Ziel dieser Arbeit ist, einen Überblick über die wichtigsten Methoden und Verfahren des Datamining zu geben, ohne allerdings einen Anspruch auf Vollständigkeit zu erheben. Diese würde angesichts des sehr aktiven Forschungsfeldes den Rahmen dieser Arbeit sprengen. Besondere Berücksichtigung finden Aspekte der Integration: auf der Ebene der Daten, Methoden und Validierung; auf der Ebene der Einbindung des Menschen in den KDD-Prozess durch effektive Visualisierung und interaktive Navigation als auch auf der Ebene der Eingliederung eines Datamining-Systems in einen medizinischen Anwendungsbereich am Fallbeispiel einer herzchirurgischen Klinik.

Welche Grundformen der Repräsentation und Datenskalierung für Objekte man unterscheiden kann, wird in Kapitel 2 beschrieben. Im Hinblick auf deren Verwendung in späteren Kapiteln wird auf Wege zu Transformation in Ähnlichkeitsdaten und spezielle Kodierungsarten besonders einge-

gangen.

Im Gegensatz zu anderen Paradigmen, z.B. der künstlichen Intelligenz, spielt der Mensch im KDD-Prozess eine ausdrückliche und wichtige Rolle. Seine Fähigkeit, Ergebnisse zu interpretieren, Hintergrundwissen zu kombinieren, den Analyseprozess zu steuern, ist in komplexen Domänen bisher unübertroffen. Entscheidend für seine Effizienz ist die geeignete Präsentation der Informationen.

In Kapitel 3 werden verschiedene Methoden der Datenpräsentation, -exploration und Visualisierung vorgestellt, die teilweise in sich bereits verschiedene Techniken integrieren.

Statistische Grundlagen, denen Kapitel 4 gewidmet ist, dienen nicht nur der Beschreibung von Daten, sondern bilden auch die Basis, um Abweichungen von der Regelhaftigkeit klar determinieren zu können. Sie ermöglichen damit, Hypothesen kompakt zu bewerten und integriert mit Konfidenzangaben darstellen zu können. Die Validierung von Wissen (Merkmal *ii* in der KDD-Definition) ist insbesondere in medizinischen Anwendungsbereichen von großer Bedeutung.

Die wichtigsten Modellformen, Kernverfahren und Algorithmen des Dataminings werden im Kapitel 5 erläutert. Aufgabenstellungen wie Klassifikation, Regression, Clustern sowie die Modellierung von Regeln und Abhängigkeiten werden dargestellt. Der Schritt zum Modell ist meist verbunden mit dem Wechsel von einer extensiven Repräsentation zu einer induktiven, kompakteren Repräsentation.

Kapitel 6 bietet einen Exkurs in einen Raum mit ungewöhnlicher Geometrie: Der hyperbolische Raum bietet u.v.a. effektiv mehr Platz als der uns täglich begegnende, euklidische Raum. Die hyperbolischen Räume lassen sich ausgezeichnet zu einer integrierten Visualisierung und kontexterhaltenden Navigation im so genannten „Poincaré-Bild des $\mathbb{H}^2$“ verwenden.

Voraussetzung für eine praktische Anwendung dieser Darstellungstechnik ist das Generieren von geeigneten Layouts im $\mathbb{H}^2$. In Kapitel 7 wird ein neuartiges Datenprojektionsverfahren, das *Hyperbolic Multi-Dimensional Scaling* (HMDS) zusammen mit zwei weiteren Verfahren, der *Hyperbolic Self-Organizing Map* (HSOM) und dem *Hyperbolic Tree Layout*, erläutert.

In Kapitel 8 werden sie anhand von Anwendungsbeispielen zur Navigation in Dokumentsammlungen vorgestellt. Der Eigenschaftsvergleich motiviert die Vorstellung einer hybriden Architektur, die das HSOM- und das

HMDS-Verfahren integriert und die Anwendbarkeit für das semantische Browsen und Suchen in großen Textkorpora für Texte aus dem Reuters-Newsticker demonstriert.

Ein medizinisches Datamining Projekt aus dem Bereich der Herzchirurgie wird in Kapitel 9 vorgestellt. Im Vordergrund steht hier zunächst die Schaffung eines integrativen Zugangs zu heterogenen und partiell konsistenten Datenbeständen, die in autonomen Fachabteilungen geführt werden. Anhand dieses Fallbeispiels werden typische Fragestellungen und Lösungsansätze aufgezeigt. Durch den Aufbau eines Data-Mart-Systems mit einem integrierten Auswertungsportal im Klinik-Intranet werden u.a. risiko-adjustierte Auswertungen aus Gründen der Qualitätssicherung und zur medizinischen Forschung ermöglicht.

Kapitel 2

Datentypen und Datenrepräsentationen

Es gibt eine große Bandbreite von Möglichkeiten, wie man abstrakte Dinge, Messdaten, Texte, Bilder, oder allgemein Informationen über Objekte digital speichern kann. Mit dem Ziel, sie Datamining-Verfahren zugänglich zu machen, findet man zwei Grundformen, die es zu unterscheiden gilt:

Fall F1: objektbezogene Merkmalsbeschreibung (*feature case*) und

Fall F2: Ähnlichkeitsaussagen zwischen Objektpaaren (*distance case*).

2.1 Merkmalsdaten (Datentyp F1)

Im ersten Fall beschreibt ein Datentupel das Objekt, welches mehrere Eigenschaftsbeschreibungen – genannt Merkmale, *features*, Attribute oder Komponenten – umfasst. Diese Begriffe werden häufig synonym verwendet.

Eine Objekteigenschaft kann durch ein einzelnes Merkmal oder eine Komponentengruppe dargestellt werden. Will man die Objekteigenschaft $\mathbf{x}_i, \mathbf{x}_j$ zweier Objekte i und j vergleichen, so sind drei Grundfragen interessant:

- Gleichheit: Sind $\mathbf{x}_i$ und $\mathbf{x}_j$ gleich? Dies setzt die Anwendbarkeit einer booleschen Prüfoperation $(\mathbf{x}_i = \mathbf{x}_j)$ voraus;
- Ordnungsrelation: Ist $\mathbf{x}_i$ kleiner als $\mathbf{x}_j$? Dies setzt die Anwendbarkeit einer booleschen Operation $(\mathbf{x}_i < \mathbf{x}_j)$ für die Wertausprägungen voraus;
- Distanz- oder Ähnlichkeitsmaß: Wie groß ist der Abstand von $\mathbf{x}_i$ und $\mathbf{x}_j$? Dies setzt die Verfügbarkeit einer allgemeinen Abstandsfunktion $d(\mathbf{x}_i, \mathbf{x}_j)$ voraus. Die gebräuchlichste ist hier die euklidische Distanzmetrik $d(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||$, siehe Abs. 2.2.

Nach den Verfügbarkeiten dieser Funktionen lassen sich verschiedene Datentypen einordnen. Die Hauptarten werden als **nominal**, **ordinal** oder **kontinuierlich** skalierte Daten bezeichnet, die in Tabelle 2.1 näher beleuchtet werden:

Datentyp	$\neq$	$<$	$\|\|$	**Bemerkung**	**Beispiel**
nominal	+ 2b	–	–	Bezeichner	Ort, Name
kategorial	+	–	–	Klassenzugehörigkeit	Hunderasse, PLZ
ordinal	+	+	–	Ordnungsrelation besteht	T-Shirt S, M, L, XL
kontinuierlich	+	+	+	reelle Zahlen $x \in X \subseteq \mathbb{R}$	Koordinaten
Intervall	+	+	+	Bereichseinschränkung	$-1 \leq x < 4$
Verhältnis	+	+	+	Nichtnegativ $0 \leq x$	Prozentskala
diskret	+	+	+	Einzelwerte, z.B. $\mathbb{Z}$	$x \in \{1, 2.3, 4\}$
zyklisch	+	–	(+)	Randwiederholung	Drehwinkel

Tabelle 2.1: Wichtige Datentypen für die paarweise die Gleichheit ($x_i = x_j$) oder Ungleichheit ($x_i \neq x_j$), die Größenrelation ($x_i < x_j$) und eine Distanzfunktion ($||x_i, x_j||$) definiert sein müssen (+), oder nicht (–). Die erweiterte Unterscheidung von Intervall und Verhältniswerten findet sich gelegentlich in der sozialwissenschaftlichen Literatur, ist aber von untergeordneter Bedeutung.

Nominalwerte sind Bezeichner, z.B. ein Name, der mit einem anderen nur sinnvoll auf Gleichheit bzw. Ungleichheit getestet werden kann. Sie sind ähnlich einem **kategorial skalierten Wert**, kurz **Kategorialwert**, der die Gruppenzugehörigkeit kodiert, z.B. für einen Hund die Rasse „Dackel“ oder „Labrador“. Der Name „kategorial“ deutet an, dass auch andere Kategorisierungen hätten gewählt werden können (z.B. „Langhaardackel“).

Er wird oft synonym für „nominal“ verwendet. **Dichotome-** oder **Binärvariablen** sind kategoriale Merkmale mit genau zwei Wertausprägungen (z.B. „ja“/„nein“ bzw. 0/1).

Ordinal skalierte Variablen kennen eine Rangordnung aller möglichen Wertausprägungen, z.B. „klein“ < „mittel“ < „groß“. Eine klare Antwort auf die Frage nach dem Abstand – wieviel kleiner ist „klein“ oder ist der Unterschied zwischen „klein“ und „mittel“ kleiner als zwischen „mittel“ und „groß“ – kann nicht erwartet werden.

Kontinuierlich skalierte Variablen: Auf sie kann eine Distanzfunktion angewendet werden, die Auskunft über den Abstand zweier möglicher Wertausprägungen gibt. Der Wertebereich kann eingeschränkt sein: **Diskrete** Variablen, z.B. eine Stückzahl (aus den nicht-negativen Zahlen $\mathbf{N}_0^+$), werden i.d.R. subsumiert, ohne dass sie eigentlich kontinuierlich sind;
Verhältniszahlen *(ratios)*, z.B. Höhe/Breite sind dimensionslos und nichtnegativ, ebenso wie **Prozentskalen**;
Intervallskalen erlauben Werte aus einem bestimmten Intervall, z.B. [0,1];

zyklische Skalen sind Sonderformen und kodieren z.B. Drehwinkel oder Wochentage. Sie haben zusätzlich zur Intervallbeschränkung eine zyklische Randfortsetzung, welcher in der Distanzfunktion Rechnung getragen werden muss. Alternativ kann eine Umkodierung auf eine kontinuierliche 2D-Kreisposition eine Lösung bieten.

Skalare und vektorielle Daten: Eine einzelne, kontinuierliche Variable wird auch als Skalarwert bezeichnet. Ist eine Gruppe von Beschreibungsgrößen auf einem Vektorraum definiert, spricht man von vektoriellen Daten.

2.1.1 Abbildungen auf Standardskalen

Um unterschiedliche Werteverteilungen zu vergleichen, sie z.B. gemeinsam zu visualisieren oder einer bestimmten Farbskala zuzuordnen, muss man die Daten komponentengerecht normieren. Am häufigsten ist der Zielbereich Einheitsintervall [0,1]. Kontinuierliche Attribute werden dazu linear abgebildet:

$$x_i \mapsto a\,x_i + b \quad \text{mit} \quad a = (\max_{\forall\, j}(x_j) - \min_{\forall\, j}(x_j))^{-1} \text{ und } \quad b = -a\,\min_{\forall\, j}(x_j)\;. \tag{2.1}$$

Muss die Intervallbestimmung auf einem Teildatensatz erfolgen, kann das Problem auftreten, dass später Daten erscheinen, die extremer sind, wodurch die Abbildung das Einheitsintervall dann sprengt. Bei Intervallskalen wird die bekannte Intervallbreite verwendet.

Ordinale Attribute mit n Wertausprägungen werden entsprechend ihres Ranges bewertet. Sei $r(x_i) \in \{1, \ldots, n\}$ der Rang des Wertes x_i, dann kann der Ordinalwert z.B. mit

$$x_i \mapsto \frac{r(x_i) - 1}{n - 1} \tag{2.2}$$

abgebildet werden.

Kategoriale Variablen mit n Wertausprägungen werden durch n Variablen ersetzt. Jede ist einer Kategorie zugeordnet und wird zu 0 oder 1 gesetzt, je nachdem, ob die Kategorie zutrifft. Mit dieser Technik der **Nominalexpansion** wird die konkrete Kategorie durch den entsprechenden Einheitsvektor in $\mathbb{R}^n$ ersetzt. Sind mehrfache Kategorienennungen möglich, werden auch die Ecken des Hyperwürfels $\{0, 1\}^n$ besetzt.

Eine andere Standardskala ist die zentrierte, auf Standardvarianz skalierte Verteilung. Sie wird durch die z-Transformation erzeugt und wird durch Gl. 4.35 beschrieben (S. 60). Diese Skalierung ist insbesondere dafür geeignet, Daten auf ihre Normalverteiltheit hin zu beurteilen. Möchte man eine (grob) normalverteilte Variable in eine uniforme Verteilung im Einheitsintervall abbilden, so bietet sich die Fermi-Funktion Gl. 5.48 an, die die z-transformierten Daten nach [0,1] abbildet.

2.1.2 Abbildung von fehlenden Werten (*missing values*)

Das Auftreten von fehlenden oder unspezifizierten Daten ist ein häufiges Problem, für das in der Praxis eine anwendungsbezogene Lösung entwickelt werden muss. Technisch empfiehlt sich eine geeignete Kodierung:

- in Fragebögen und Tabellen durch Spezialcode, z.B. „99“ (nicht empfehlenswert, s. u.) oder „NN“;
- in SQL Datenbanktabellen mit Spezialwert „NULL“ bzw. „\N“;
- im Hauptspeicher durch die Spezialzahl *Not-a-Number* (**NaN**), die durch Standardprozessoren und Rechenwerke nach IEEE-754 eine

Spezialbehandlung erfährt. Tauchen NaN-Werte als Operanden bei arithmetischen Operationen ($+, -, *, \ldots$) auf, ist das Ergebnis auch NaN, bei booleschen Vergleichsoperationen ($<, \leq, != \ldots$) ist das Ergebnis stets *false*. Damit ergibt sich eine Testmöglichkeit $(x = x)$ auf Wertpräsenz (denn NaN=NaN ergibt *false*).

Der Vorteil dieser Kodierung liegt zum einen in der integrierten Repräsentation, außerdem muss kein Extraspeicher die Information über die Wertgültigkeit aufnehmen. Der andere Vorteil ist die Konsistenz des IEEE-754-Standards. Werden NaN-Werte, z.B. in einer Mittelwertberechnung, nicht spezialbehandelt, sondern aufsummiert, ist das Ergebnis NaN. Ohne diese Regel wäre das Ergebnis falsch, ohne dass es nachher erkennbar ist (s.u.).

- Die Spezialzahl „0“ (o.ä.) ist keine gute Kodierung, denn sie schützt unzureichend vor Fehlinterpretationen. Im vorigen Beispiel der Mittelwertbildung würde dies zu schlicht falschen Ergebnissen führen.

2.2 Distanz- und Ähnlichkeitsmaße (Fall F2)

Ähnlichkeiten (***proximities***) drücken zunächst einmal eine Nähe von Objekten in irgendeinem Raum aus. Manchmal sind es klare metrische Größen, die eindeutig analysierbar sind, manchmal sind Ähnlichkeiten auch schwierig und nur kontrovers interpretierbar. Die Umkehrung sind **Unähnlichkeiten** oder Abstände δ_{ij} zwischen zwei Objekten i und j aus einer Menge M von N Objekten. Die Paarabstände sind nicht-negativ $\delta_{ij} \geq 0$ und Selbstabstände sind Null $\delta_{ii} = 0$. Neben diesen allgemein gültigen Voraussetzungen gibt es weitere Anforderungen, die gestellt werden können. Eine Systematik in zwölf Stufen wurde von Hartigan (1967) entwickelt. Sie reicht von einer schlichten Partitionierung von M in Mengen gleicher Objekte, über Rangbildung, und den Symmetrieanspruch $\delta_{ij} = \delta_{ji}$ – zur Metrik bis hin zur euklidischen Distanz. Die letzte und anspruchsvollste von Hartigans Kategorien wird später zur Minkowski-Metrik (s. Gl. 2.9, siehe auch Erweiterung für nicht-euklidische Distanzmaße in Kap. 6).

Distanzmaße $\{\delta_{ij}\}$ genügen den Ansprüchen einer **Metrik**, wenn sie folgende Eigenschaften bzgl. des Selbstabstands, der Symmetrie

$$\delta_{ij} = 0 \qquad \Leftrightarrow i = j \tag{2.3}$$

$$\delta_{ij} = \delta_{ji} \qquad \forall\, 1 \leq i, j \leq N \quad \text{und zudem} \tag{2.4}$$

$$\delta_{ij} \leq \delta_{it} + \delta_{tj} \qquad \forall\, 1 \leq i, j \leq N\,, \tag{2.5}$$

also die Cauchy-Schwarz'sche Dreiecksungleichung, erfüllen.

Schreibt man die Unähnlichkeiten $\{\delta_{ij}\}$ in eine Matrix **D**, nennt man **D** metrisch, wenn die Elemente obige Eigenschaften erfüllen. Eine nichtmetrische Matrix kann metrisch gemacht werden, indem man alle Nebendiagonalelemente inkrementiert (Cox und Cox 1994):

$$\delta'_{ij} = \delta_{ij} + c \qquad \forall\, i \neq j \text{ mit} \tag{2.6}$$

$$c \geq \max_{\forall\ ijk} |\delta_{ij} + \delta_{ik} - \delta_{jk}|. \tag{2.7}$$

2.3 Erzeugung von Distanz- und Unähnlichkeitsmaßen

Sind Merkmalsdaten (Fall F1) gegeben, können Distanzmaße direkt erzeugt werden. In die umgekehrte Richtung ist dies auch ohne weiteres möglich (s. Abs. 5.10).

2.3.1 Kontinuierliche oder diskrete Daten

Für kontinuierliche oder diskrete Daten bieten sich eine Auswahl von Unähnlichkeitsmaßen an:

Euklidische Distanz oder L_2-Norm:

$$\delta_{ij} = \sqrt{\sum_k |x_{i,k} - x_{j,k}|^2}\,; \tag{2.8}$$

Minkowski-Distanz oder L_λ-Norm:

$$\delta_{ij} = \left\{ \sum_k |x_{i,k} - x_{j,k}|^\lambda \right\}^{1/\lambda}, \tag{2.9}$$

Die Minkowski-Distanz enthält als Spezialfall die euklidische Norm ($\lambda = 2$) und die beiden Folgenden:

die city-block-, oder Manhattan-Norm: Gl. 2.9 mit $\lambda = 1$:

$$\delta_{ij} = \sum_k |x_{i,k} - x_{j,k}| \ ;$$

und die

Maximal Norm: Gl. 2.9 mit $\lambda \to \infty$:

$$\delta_{ij} = \max_{\forall k} |x_{i,k} - x_{j,k}| \ ;$$

Gewichtete euklidische Norm:

$$\delta_{ij} = \sqrt{\sum_k w_k |x_{i,k} - x_{j,k}|^2} \tag{2.10}$$

Häufig werden Komponentengewichte w_k so gewählt, dass unterschiedliche Skalierungen einzelner Dimensionen systematisch kompensiert werden, entweder durch

- Normierung auf Einheitsvarianz $w_k = \sigma_k^{-1}$ oder
- Normierung auf Einheitsbreite $w_k = (\max_{\forall k} x_{i,k} - \min x_{i,k})^{-0.5}$ (äquivalent zu Gl. 2.1).

Andere informationstheoretische und entropiebasierte Überlegungen können die w_k bestimmen (Yu et al. 2001).

Mahalanobis-Distanz: $\delta_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$. Durch Matrixmultiplikation mit der inversen Kovarianzmatrix der Daten $\mathbf{C}^{-1}$ werden alle Achsen varianzgleich skaliert. Zusätzlich zur gewichteten euklidischen Norm werden Rotationen und mögliche Verzerrungen kompensiert. Dies entspricht einer „Sphärisierung“ (*sphereing*) der Daten in Hauptachsenlage und anschließender euklidischer Distanzmessung.

Kosinusmetrik: misst die Winkelseparation zwischen zwei Vektoren

$$\delta_{ij} = 1 - |\cos \angle(\mathbf{x}_i, \mathbf{x}_j)| = 1 - \frac{|\mathbf{x}_i \mathbf{x}_j|}{||\mathbf{x}_i|| \ ||\mathbf{x}_j||} \ ; \tag{2.11}$$

Korrelation: Basierend auf Pearson's Korrelationskoeffizient r in Gl. 4.84 werden zwei Vektoren als Stichprobe eines Variablenpaares untersucht $\delta_{ij} = 1 - |r|$ (s. Abs. 4.13.2).

2.3.2 Binärdaten

Besteht der Datensatz aus n binären Variablen, wird üblicherweise ein Ähnlichkeitskoeffizient gebildet und in ein Unähnlichkeitsmaß transformiert. Zunächst werden zwei Datenobjekte i und j auf die Häufigkeit der vier möglichen Kombinationen jedes der n Binärvariablenpaare untersucht (s. Tab. 2.2) und daraus wird ein Ähnlichkeitskoeffizient berechnet. Auch hier gibt es wieder eine Auswahl von Möglichkeiten (s.a . Cox und Cox 1994):

		Objekt j		
		1	**0**	
Objekt i	**1**	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n=a+b+c+d

Tabelle 2.2: Kontingenztabelle für Binärdatensatz i und j enthält die vorgefundene Anzahl von den vier möglichen Kombinationen.

Einfacher Übereinstimmungskoeffizient: $s_{ij} = \frac{a+d}{a+b+c+d}$;

Jaccard-Koeffizient: $s_{ij} = \frac{a}{a+b+c}$;

Yule-Koeffizient: $s_{ij} = \frac{ad-bc}{ad+bc}$;

Dice-Koeffizient: $s_{ij} = \frac{2a}{2a+b+c}$;

Rogers- oder Tanimoto-Koeffizient $s_{ij} = \frac{a+d}{a+2b+2c+d}$;

2.3.3 Nominal- und Ordinaldaten

Prinzipiell kann für jedes mögliche Wertausprägungspaar ein Übereinstimmungskoeffizient definiert werden. Für die Nominalvariable ist im Standardfall $s_{ij} = 1$ bei Übereinstimmung und ansonsten 0. Ordinalwerte können, je nach Anwendungsfall, evtl. als Enumeration der Rangfolge gewertet werden.

2.3.4 Mischdaten und fehlende Einträge

Besteht der Datensatz aus einer Mischung der obigen Datentypen, wird letztlich eine Kombination die Gesamtdistanz bestimmen. Treten fehlende Einträge auf, können die betroffenen Komponentenpaare nicht verglichen werden. Gower schlug (1971) einen verallgemeinerten Ähnlichkeitskoeffizient

$$s_{ij} = \frac{\sum_{\forall k} v_{ijk} s_{ijk}}{\sum_{\forall k} v_{ijk}} \text{ mit } v_{ijk} = \left\{ \begin{array}{ll} 1 & \text{wenn } s_{ijk} \text{ valide} \\ 0 & \text{sonst} \end{array} \right\} \tag{2.12}$$

vor, der das Problem wie eine Extrapolation löst. s_{ijk} ist das auf die Komponente k bezogene Ähnlichkeitsmaß und v_{ijk} ist eine binäre Indikatorvariable, die das Vorhandensein der beiden k-ten Komponenteneinträge in Datensatz i und j und die Durchführbarkeit des Vergleiches mit 1 anzeigt.

Bei fehlenden Komponenteneinträgen werden damit die Lücken des Vergleichsergebnises effektiv durch den Mittelwert der übrigen ersetzt. Dieses Konzept ist unmittelbar auf Unähnlichkeitsmaße anwendbar. Möchte man eine Gewichtung der Komponenten k einfügen, sind die Faktoren jeweils im Zähler und Nenner von Gl. 2.12 einzufügen.

Transformation von Ähnlichkeits- in Unähnlichkeitsdaten gestalten sich einfach, je nach Bedarf z.B. zu:

$$\delta_{ij} = 1 - s_{ij} \tag{2.13}$$

$$\delta_{ij} = c - s_{ij} \quad \text{mit Konstante } c \tag{2.14}$$

$$\delta_{ij} = (1 - s_{ij})^{0.5}$$

2.4 Spezielle Datenrepräsentationen und Distanzmaße

Im Folgenden seien einige Beispiele spezieller Datenrepräsentationen und Distanzmaße erläutert, die in der Praxis von großer Bedeutung sind. Zunächst stehen textuelle Informationen auf der Ebene der Zeichenkette, des Wortes und des Textdokumentes im Vordergrund. Es folgen Verfahrensweisen für Bilddaten und Zeitserien.

2.4.1 Editdistanzen auf Zeichenketten

Die Editdistanz zweier Symbol- oder Zeichenketten s_1, s_2 quantifiziert den Aufwand, die eine Kette in die andere umzuwandeln. Die beiden Beispiele

„Frau Muster-Schmidt“ ↔ „Fr. Muster Schmitt“
„AACGTCGTAGCTGGT“ ↔ „AACGTCGTTGCTCGT“.

verdeutlichen die Anwendungsgebiete: Ist der erste Namensunterschied möglicherweise ein Übertragungsfehler? Durch Erstellen von Listenpaarungen können mittels der Editdistanz systematisch Korrekturvorschläge erstellt werden. Das zweite Stringpaar kodiert symbolhaft die Abfolge von Nukleinsäuren auf zwei DNA-Abschnitten. Das *Human Genom Project* ist ein prominentes Beispiel für die Bedeutung von enormen Mengen von bioinformatischen Symbolketten. Wichtig sind hier Verfahren, die gleiche oder ähnliche Abschnitte effizient finden und ggf. ihre Unterschiedlichkeit bewerten.

Die Kernidee ist dabei die Kostenbewertung von elementaren Editieroperationen, ein Kette in die andere umzuwandeln:

- Einfügung eines Zeichens: Kostenfaktor c_{ins},
- Löschen eines Zeichens: Kostenfaktor c_{del} und
- Austausch eines Zeichens: Kostenfaktor c_{mod}

und die Summation über alle nötigen Teilschritte.

Die gesuchte Editdistanz beziffert das günstigste aller möglichen Umbauverfahren vom ersten zum zweiten String. Im Standardfall wird eine 2D-Matrix der Größe $|s_1| \cdot |s_2|$ systematisch aufgebaut und bewertet. Da diese quadratische Skalierung ein Hemmschuh ist, wird der hohe Bedarf an hochskalierbaren Verfahren in der Bioinformatik deutlich. Als Beispiel für die Weiterentwicklungen zu schnellen Spezialverfahren mit und ohne Heuristiken seien die so genannten „Suffixbäume“ genannt (Giegerich, Kurtz und Stoye 1999).

2.4.2 Textdokumente und das *Bag-of-Words*- oder Vektorraummodell

Die nächste Ebene vom Zeichen zum Text ist die Repräsentation der Wörter. Hier hat sich das Standardmodell des „Wortsackes“, des *bag-of-words*-Modells etabliert. Der Name deutet schon auf die Kernidee: Man verwirft sämtliche semantische Bedeutungen, die in der Sequenz der Wörter stecken, und beschränkt sich allein auf das Erscheinen von Wörtern. Dies ist eine massive Informationsreduktion, doch der Erfolg dieses Verfahrens adelt das Konzept (Salton und Buckley 1988).

Der Zweitname „Vektorraummodell“ deutet auf die Beschreibungsform eines Textes als hochdimensionaler Vektor hin. Jede Komponente zählt das Auftauchen eines bestimmten Wortes. Zunächst wird der Vektorraum konstruiert, indem festgelegt wird, welche Wörter verwendet werden sollen. Hierzu bildet man die duplikatfreie Vereinigungsmenge aller Terme aus allen N Dokumenten des Textkorpus und entwickelt daraus das Wörterbuch (*dictionary*).

Erstellung des Wörterbuchs: Um der Vielzahl potentieller Deklinationen und Konjugationen von Wörtern Rechnung zu tragen, wird eine Wortstammreduktion durchgeführt z.B. *{„country“, „countries“}* $\rightarrow$ *„countri“* und anschließend werden die Erscheinenshäufigkeiten bestimmt. Als nächstes werden bedeutungsfreie Terme anhand einer *Stoppwortliste* eliminiert (z.B. *„e.g.“, „i.e.“, „and“*). Wesentlich dabei ist, dass das Erscheinen von sehr häufigen „Allerweltsworten“ nicht sehr aussagekräftig für einen Text ist (z.B. *„said“*, „kann“, „Jahr“). Umgekehrt sind sehr seltene Spezialworte, wie z.B. „Steuervergünstigungsabbaugesetz“, zwar möglicherweise sehr bedeutungsvoll, aber diese Bedeutung tritt nur selten zum Vorschein. Im anwendungspraktischen Kompromiss trimmt man das Wörterbuch, indem man die häufigsten und die seltensten Terme exkludiert (s.u.). Das resultierende Wörterbuch mit k Wörtern legt die Bedeutung der k Dimensionen des Vektorraumes $\mathbb{R}^k$ fest.

“Term frequency $\times$ invers document frequency” Schema: Schon 1932 hat sich der Linguistikprofessor George Zipf an der Harvard-Universität mit Studien zur Wortfrequenz befasst und ist dabei auf allgemeine Skalierungsphänomene gestoßen. Sortiert man Wörter nach ihren Häufigkeiten, so gibt es einen näherungsweise proportionalen Zusammenhang zwischen dem Rang (der Worthäufigkeit) und dem Logarithmus der Häufigkeit (Zipf 1945, siehe auch Mandelbrot 1983 für eine Diskussion und Er-

weiterungen). Bi et al. (2001) fand eine Verallgemeinerung mit diskreten Gaußverteilungen (DGX), die im Bereich hochfrequenter Worte besser passen und deren Parametrisierung auch als Dokumentindikatoren verwendet werden können.

Aus dem Zipf'schen Gesetz kann man einen groben funktionalen Zusammenhang zwischen Wortwichtigkeit und Häufigkeit ableiten. Tritt ein Wort per se häufig auf, ist die Anwesenheit im konkreten Fall weniger hoch einzuordnen, als wenn ein relativ seltenes Wort auftaucht. Die Quintessenz ist das *term frequency* $\times$ *invers document frequency* (**TFIDF**)-Schema zur Repräsentation jedes Textdokumentes als Vektor $f_{\boldsymbol{t}} \in \mathbb{R}^k$ mit der i-ten Komponente

$$f_{\boldsymbol{t},i} = \mathit{TF}(t, w_i) \log\left(\frac{N}{\mathit{DF}(w_i)}\right). \tag{2.15}$$

$\mathit{TF}(t, w_i)$ ist die Termfrequenz (*term frequency*) und zählt die Anzahl der Vorkommen von Term $w_i \in \{1, \ldots, k\}$ im Dokument t. $\mathit{DF}(w_i)$ benennt die Dokumentfrequenz (*document frequency*) und zählt die Dokumente, in denen der Term w_i auftritt.

Die Dokumentdistanz und damit die Unähnlichkeit zweier Dokumente wird mit der Kosinusmetrik Gl. 2.11

$$\delta_{ij} = 1 - \cos(\vec{f}_{\boldsymbol{t}_i}, \vec{f}_{\boldsymbol{t}_j}) = 1 - \vec{f}'_{\boldsymbol{t}_i} \vec{f}'_{\boldsymbol{t}_j} \tag{2.16}$$

bestimmt, die am effektivsten implementiert wird, indem man die normalisierten Merkmalsvektoren

$$\vec{f}' = \frac{\vec{f}}{||\vec{f}||} \tag{2.17}$$

speichert.

2.4.3 Bildrepräsentation

Die Basisdarstellung von Bildern ist die 2D-Pixelmatrix mit Grau- oder Farbwerteinträgen. Wegen der großen Datenfülle erfolgt die persistente Speicherung häufig in komprimierter Form: Je nach Bedarf werden zum Teil Informationsverluste in Kauf genommen. Die Speichertiefe wird in bit gemessen und beschreibt die Wertdiskretisierung pro Farbebene (meist 8 bit, seltener 1, 2, 12 oder 16 bit).

Der Bereich Computersehen (*computer vision*) ist ein zunehmend wichtiger Bereich von anwendungsnaher Informationsverarbeitung. Er umfasst ein sehr breites Anwendungs- und Methodenspektrum, darunter:

- Extraktion von Bildmerkmalen, z.B. von Ecken, Kanten, Konturen, Linien bestimmter Elementarformen;
- Segmentierung in zusammengehörigen Bildbereichen, z.B. nach Farbe, Form und Textur und z.T. unter Berücksichtigung von Gestaltgesetzen; Trennung von Figur und Hintergrund;
- Klassifikation von Bildbereichen, u.a. Objektidentifikation, z.B. automatische Bestimmung der Bebauungsart in Satellitenbildern oder Gewebebeurteilung in Röntgen- und NRM-Bildern bzw. -Bildfolgen;
- 2D- und 3D-Objekterkennung in Bildern oder Bildfolgen;
- Extraktion von Objektpositions- und -orientierungsinformation, z.B. zur industriellen Qualitätskontrolle oder zur Robotersteuerung (Walter und Arnrich 2000);
- Erkennen von Personen zur Aufmerksamkeitssteuerung oder Zugangskontrolle;
- Erkennung von Gestik als weitere Computereingabemodalität, z.B. Hand- und Fingerzeigegestik zur Kommandierung von Roboter-Greifsystemen (Littmann, Meyering, Walter, Wengerek und Ritter 1992; Walter, Nölker und Ritter 2000);
- Ähnlichkeitssuche in Bildarchiven (*image/video retrival*). Welche Bilder sind ähnlich zu einer Bildvorlage oder -beschreibung? Der Begriff Ähnlichkeit ist hier in besonderer Weise situationsabhängig, z.B. können Farbe, Textur oder Motivarten eine Rolle spielen.

Ist die relative Kamera-Objekt-Lage nicht fest definiert, treten Verschiebungen, Rotationen und perspektivische Verzerrungen auf, die sich auf Pixelmatrixebene nichttrivial auswirken (hierzu mehr in der Standardliteratur, z.B. Marr 1982; Fischler und Frischein 1987; Sonka et al. 1998; Handels 2000; Jähne 2001). Eine Grundoperation ist die lokale Korrelationsanalyse. Hier bildet man das Faltungsprodukt des Bildes mit einem systematisch verschobenen Vergleichsbild und wertet das Ergebnis über

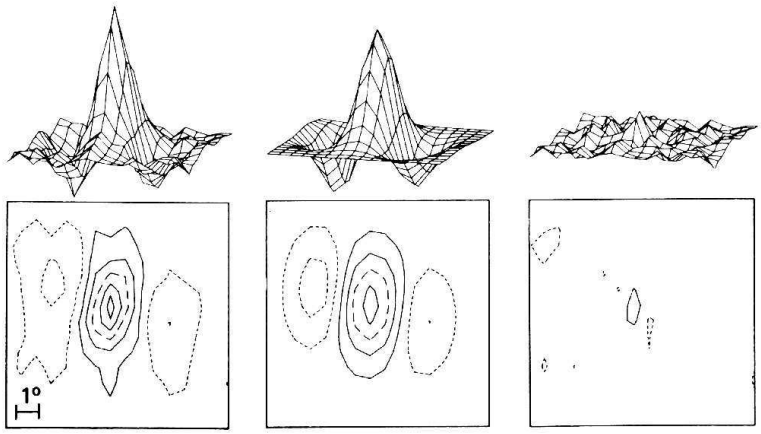


Abbildung 2.1: Gaborfilter in der Biologie: *Links:* experimentell bestimmtes Antwortverhalten von „simple cells“ im Primären Visuellen Kortex einer Katze. *Mitte:* adaptierter Gaborfilter. *Rechts:* Differenz der adaptierten Filterwerte von den experimentellen Daten.

die Verschiebung aus. Die Vergleichsbilder sind meist anwendungsspezifische Objektteilbilder oder auch Standardmatrizen zur Textur- oder Kantenfindung (auch Faltungskern oder -matrix genannt).

Gaborrepräsentation und -wavelets: Eine wichtige Repräsentationsart ist die Transformation in den Ortsfrequenzraum durch eine 2D-Fouriertransformation. Lokalisierte Variationen sind die so genannten Wavelet- und Gabortransformationen. Diese gehen auf Denis Gabor (1946) zurück und finden sich interessanterweise auch in biologischen Sehsystemen, wie in Abb. 2.1 illustriert.[1] Die Gabor-Faltungskerne werden durch eine ebene Welle beschrieben, die von einer elliptischen Gaußglocke moduliert wird.

$$\Psi_{\lambda\sigma\alpha}(x,y) = \exp\left(-\frac{x^2+\alpha^2y^2}{2\sigma^2}\right)\left[\exp\left(-2\pi i\frac{x}{\lambda}\right) - \exp\left(-\frac{\sigma^2}{2\alpha}\right)\right] \quad (2.18)$$

Die Welle hat in x-Richtung die Wellenlänge λ; die Gaußfunktion längs die Breite σ und quer σ/α, wobei α die Streckung angibt. Der letzte konstante Term macht den Filter invariant gegenüber einer Verschiebung der Grauwertintensität des Bildes (*DC free*).
Gleichung 2.18 kann als *mother* ***wavelet*** bezeichnet werden, und durch die folgende Konstruktionsvorschrift wird eine vollständige Familie von selbst-

[1] Aus Jones und Palmer 1987, siehe auch Daugman und Downing 1995.

ähnlichen *daughter wavelets* (manchmal auch *jet* genannt) erzeugt:

$$\begin{aligned} \Psi_{mpq\theta}(x,y) = &\ 2^{-2m}\Psi(x',y') \\ x' = &\ 2^{-m}[+x\cos\theta + y\sin\theta] - p \\ y' = &\ 2^{-m}[-x\sin\theta + y\cos\theta] - q. \end{aligned} \tag{2.19}$$

Der ganzzahlige Parameter m bestimmt die Ausdehnung und Frequenz der Welle, p und q die Translation und θ den Rotationswinkel. Jede Zelle kann jetzt durch eine Gaborfunktion mit dem Zentrum (p,q), der Wellenlänge $\lambda/2^m$, der Richtung θ und einer einhüllenden Gaußglocke mit Breite σ und σ/α modelliert werden. Um die Faltung für alle Pixelorte (p,q) durchzuführen, kann sie als Multiplikation im Fourierraum durchgeführt werden.

Eine starke Dimensionsreduzierung des Bildes ist durch Selektion des *jets* und die Auswertung dieses Satzes von Gaborfiltern an spärlich platzierten Orten möglich. Diese Information kann z.B. unmittelbar in einem neuronalen Netz zur Lagebestimmung von Objekten verwandt werden (Walter et al. 2000).

Gaborfilter eignen sich auch sehr gut zur Texturbeschreibung in ihrer lokalen Frequenz- und Richtungszusammensetzung (Fogel und Sagi 1989; Ontrup und Ritter 1998) und bilden damit die Basis für **Texturunähnlichkeitsmaße**. Texturdistanzen werden als gewichtete euklidische Norm über eine zu bestimmende Menge von Filterantworten definiert (Gl. 2.10).

Farbräume und Farbunähnlichkeiten: Spektrale Körpereigenschaften von Selbstleuchtern oder streuenden Oberflächen können viel über den Körper aussagen. Dies macht sich die Natur zunutze und stattete uns mit einem trichromatischen Sehsystem aus. Mit drei Grundtypen von Sehzapfen sind wir in drei Spektralbänden farbempfindlich und sehen „Farben“. Farbreproduktion wird dadurch stark vereinfacht, denn drei Grundfarben genügen, um einen Bereich (den „Gamut“) der möglichen Farben zu erzeugen. Bei additiver Farbmischung, z.B. bei Bildschirmen, reichen Rot, Grün und Blau, bei subtraktiver Farbmischung im Printbereich genügen Gelb (*Yellow*), Cyan, Magenta und Schwarz (*Black*; zur Kontrastbesserung und aus Kostengründen). Dies begründet die Verbreitung der gerätespezifischen Farbräume RGB und YCMB. Sie bedürfen einer Kalibrierung, um in einem absoluten, z.B. in dem (CIE 1931) XYZ-Farbraum affin abgebildet zu werden (siehe Abb. 5.24b). Verfahrenstechnisch einfach ist die Transformation in den HSI/HSV-Raum, der den Farbton (*hue*), die Sättigung und die Helligkeit (*intensity, value*) der Farbe sehr schnell zu RGB umrechnet. Allerdings ist die resultierende Farbmetrik willkürlich und entspricht nicht

der Farbsensibilität des Menschen. MacAdam untersuchte (1942) die Mindeständerung an Farbvalenzen, die einen wahrnehmbaren Unterschied in Farbton oder Sättigung erkennen lassen, und fand große Inhomogenität. Dies gab Anlass zur Festlegung von perzeptuell adäquaten Farbräumen, die Farbabstände homogen bewerten. Verbreitung fanden die Farbräume der L*u*v (CIE-1976) und insbesondere der L*a*b (CIE-1978), die beide aus nicht-linearen Transformationen resultieren und die geforderte Abstandstreue in euklidischer Norm gut umsetzen. Sie finden Anwendung bei der Messung objektiver Farbunterschiede (bedeutungsvoll in der Druckindustrie), bei der gezielten Farbmanipulation in Bildern (Druckvorstufe) und bei der Konstruktion von Farbpaletten, die perzeptuell adäquat sind, z.B. Wyszecki und Styles (1982) und Hunt (1995).

EMD-Metrik: Yassi Rubner et al. (1998) (2000) entwickelten perzeptuelle Metriken für Bilder in Datenbanken. Ausgehend von sehr stark vergröberten Bildern und Pixelhistogrammen im CIE-L*a*b-Farbraum wird eine so genannte „Signatur“ für jedes Bild gebildet. Ursprünglich besteht diese aus wenigen (7-12) Clusterzentren im Farbraum und deren Gewichtung.

Mit der *Earth-Mover-Distance* (EMD) wird nun die Distanz zweier Verteilungen anhand solcher kompakten Signaturen berechnet. Sie ist definiert als die minimale „Arbeit“, um eine Signatur in die andere umzuformen – in Analogie zum Erdarbeiter, der „sparsam“ umschaufelt. Vorteilhafterweise können die beiden Signaturen unterschiedliche Länge und unterschiedliches Gesamtgewicht besitzen. Jan Puzicha et al. (1999) untersuchten die EMD und weitere Bilddistanzmaße in umfangreichen Simulationen. In Abb. 8.8 dient die EMD-Metrik zur Anordnung von Bildern in der hyperbolischen Ebene (s.S. 215)

2.4.4 Repräsentationen von Zeitserien

Für zeitliche Beschreibungen eines Systems oder eines Signals gibt es zwei Hauptrichtungen der Datenrepräsentation: Zustandsraumeinbettung und Spektraldarstellung.

Zunächst wird eine (oder mehrere) Observable des Systems $x(t)$ gewählt (z.B. Ort, Signalpegel) und „stroboskopisch“ mit (vorteilhafterweise) konstanter Frequenz $1/\Delta t$ aufgezeichnet. Aus dieser Zeitsequenz $\{x(t_i)\}$

mit $t_i = i\Delta t$ $(i \in \mathbf{N})$ wird nun ein Zeitabschnitt der Breite m betrachtet

$$\mathbf{x}_{state}(t) := \begin{pmatrix} x(t) \\ x(t-\Delta t) \\ x(t-2\Delta t) \\ \vdots \\ x(t-(m-1)\Delta t) \end{pmatrix} \in X \subseteq \mathbf{R}^m. \qquad (2.20)$$

und in den Zustandsraum X_{state} „eingebettet". Dieser Ansatz spielt eine wichtige Rolle, wenn man den vollständigen Zustand einer Systemdynamik beschreiben möchte. Hiermit lassen sich zum Beispiel determinierte chaotische Systeme modellieren und Aussagen über die nächste Zukunft treffen. Ist die Dimensionalität d der so genannten „Attraktordynamik" bekannt, so genügt nach Takens (1981) die Einbettungsdimension $m > 2d + 1$. In der Praxis werden Δt und m oft durch Ausprobieren bestimmt, indem man mit einem kleinen, dann steigenden m anfängt, bis gute Ergebnisse gefunden werden. Systematischere Prozeduren finden sich aufgrund der Messung der fraktalen Dimension r des Attraktors (Grassberger und Procaccia 1983) und informationstheoretischer Überlegungen, siehe auch z. B. Fraser und Swinney 1986; Broomhead und King 1986; Walter 1991, sowie Weigend und Gershenfeld 1994.

Die zweite wichtige Darstellung ist die *Short-Time-Fourier*-Transformation (STFT), eine spektrale Analyse in Zeitabschnitten. Das Konzept ist völlig analog zur Gaborfilterung (Gl. 2.18), hier auf der Zeitachse. Aus Performanzgründen wird oft eine einfache Hüllkurve verwendet: eine Rechteckfunktion der Breite $m = 2^i$ oder eine trapezförmige Modulation. Dank der Erfindung der *Fast-Fourier-Transformation* (FFT) lassen sich sehr effektiv Frequenz-Zeit-Repräsentationen des Signals finden (Press et al. 1988). Deren graphische Darstellungen als Intensitätsbild werden auch „Spektrogramme" genannt. Fasst man hörphysiologisch orientierte Frequenzbänder zusammen, kommt man zur perzeptuell adäquaten „Bark"-Skala und durch nicht-lineare Transformationen zu den Cepstral-Koeffizienten, die im Bereich der maschinellen Verarbeitung gesprochener Sprache vorteilhaft sind.

Kapitel 3

Datenpräsentation und -exploration

„In der Tat ist der Mensch ein Augentier.“

(Herbert W. Franke)

„Imagination oder Visualisierung und besonders die Benutzung von Diagrammen haben einen entscheidenden Anteil an der wissenschaftlichen Forschung.“

(René Descartes, 1637)

Wenn wir Daten explorieren, wollen wir uns ein mentales Bild schaffen, um Eigenschaften über den Weltausschnitt zu erfahren, der von den Daten mutmaßlich beschrieben ist. Dies lässt sich auch direkt als „visualisieren“ umschreiben. Der Begriff „Visualisieren“ umfasst neben diesem Aspekt *(i)* des mentalen Prozesses, sich etwas vorzustellen, noch zwei weitere: *(ii)* Dinge sichtbar zu machen, die wir mit dem Auge nicht unmittelbar sehen können, z.B. mittels einer Wärmebildkamera (Licht in anderem Spektralbereich zu sehen), eines Mikroskops (zur räumlichen Bündelung von Licht) oder eines Tunnelelektronenmikroskops (Abtastung von Oberflächen auf atomarer Ebene). Im Folgenden steht der dritte Aspekt im Vordergrund: *(iii)* das Sichtbarmachen von abstrakten Dingen, die i.d.R. keine natürliche, räumliche Entsprechung besitzen.

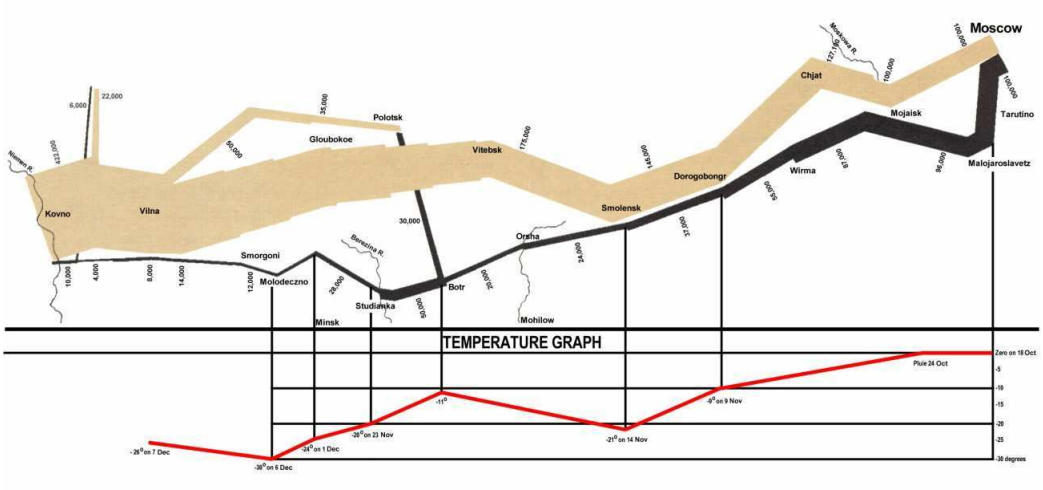


Abbildung 3.1: Der französische Ingenieur Charles Joseph Minard (1781-1870) illustrierte Napoleon's verlustreichen Russlandfeldzug von 1812: Die Breite des Bandes symbolisiert die Mannstärke auf dem (*braun*) Hin- und (*schwarz*) Rückweg. Die Liniengraphik *unten* zeigt Temperaturenverläufe und Zeitmarken auf dem Rückzug im Winter 1812.

Ziele:
Die Datenpräsentation kann mehreren Zwecken dienen (Keim 2001):

Explorative Analyse: Ausgangspunkt sind die vorliegenden Daten, ohne dass bereits eine konkrete Hypothese vorliegt. Der Prozess sucht typischerweise interaktiv nach Strukturen und Trends etc. Das Resultat sind Visualisierungen, die Hypothesen über die Daten liefern. Im darauf folgenden Schritt gilt es dann, diese auf ihre Gültigkeit zu prüfen und z.B. statistisch zu validieren (s. Kap. 4);

Konfirmatorische Analyse: Neben den Daten besitzt man bereits Hypothesen, die man dazu nutzt, die Untersuchung der Daten zielgerichtet zu lenken. Das Resultat sind Visualisierungen, die die Hypothesen möglichst eindeutig bestätigen oder widerlegen;

Ergebnispräsentation: Ausgangspunkt sind Daten und Fakten, für die es gilt, die besten Darstellungen auszuwählen. Das Ergebnis sind hochwertige und prägnante Präsentationsdarstellungen, die die Faktenlage auf der gegebenen Datenbasis illustrieren.

Als Beispiel bildet Abb. 3.1 eine Darstellung von Charles Minard nach, die als Meilenstein der statistischen Datenpräsentation und beste hi-

storische Graphik gilt (Tufte 1983). Sie integriert die Darstellung von Zeit, Ort, Temperatur und Truppenstärke von Napoleon's Russlandfeldzug im Jahre 1812[1] (Tufte 1983).

Wahrnehmungsmodalitäten: Neben dem visuellen Sinn stehen uns die Geschmacks-, Tast- und Hörwahrnehmung als Perzeptionsmodalitäten zur Verfügung, um relevante Signale unserer Umwelt zu analysieren und zu interpretieren. Die menschliche Evolution hat hervorragende Strukturen zur Signalverarbeitung, Mustererkennung und Interpretation herausgebildet. Bei der Präsentation von Daten geht es darum, Darstellung- und Interaktionsformen zu finden, die an bestehende Signalverarbeitungskapazitäten ankoppeln.

Die Ankopplung an den Geschmacks- und Tastsinn erscheint (heute noch) gänzlich unpraktikabel. Der auditive Kanal ist dagegen interessant, da Displays leicht verfügbar sind. Unser Hörvermögen ist darauf spezialisiert, zeitliche Signalmodulationen in physiologisch bedingten Bereichen sehr fein zu differenzieren. Die Vielfalt der Möglichkeiten, künstliche Daten durch Sonifikation geeignet erfahrbar zu machen, wurden z.B. von Hermann (2002) untersucht.

Statische und dynamische Präsentationen: Eine statische Präsentation ist im Vergleich zu einer dynamischen i.d.R. meist einfacher und weniger unterhaltsam – hat aber durchaus auch ihre Vorzüge. Insbesondere wenn mehrere Personen der Darstellung folgen, behält jeder für sich die Autonomie über den Darstellungsraum. Durch schnelle visuelle Sakkaden (Blickbewegungen) kann jeder die seine eigene Aufmerksamkeit bewusst und unbewusst lenken.

Bei einer animierten Darstellung, etwa einer Animation, einem Film oder einer Sonifikation, ist dies zunächst nicht möglich: Man hat keine Autonomie über die Zeit – sie läuft. Erst durch zusätzliche Interaktionsformen wird eine Steuerung wieder möglich und Kontrolle über den Darstellungsraum zurückerlangt. Interessanterweise wird zur Steuerung der Zeitachse neben inkrementellen Weiterblätter- und Weiterspul-Aktionsmetaphern gern auf eine räumliche Metaphore (z.B. Schieberegler oder Indexwahl) zurückgegriffen, die wieder eine sakkadenartige Sprungfunktion anbietet.

Interaktionstechniken: Zur effektiven Exploration von Daten ist die Integration von Visualisierung und Benutzersteuerung zu einer interaktiven

[1] http://www.napoleonic-literature.com/

Darstellung unerlässlich. Neben der Auswahl, *was* dargestellt wird (ggf. Teilmengen, Filter), sollte das *wie*, also die Art der Abbildung (Ort, Attribute, Projektionen), steuerbar sein. Erweiterte Interaktionstechniken beinhalten darstellungsinterne oder -externe Navigationsmöglichkeiten, z.B.: *panning* und *zooming* zur Steuerung des Ausschnitts und der Auflösung; *linking* zu Hypertext-artigen Verknüpfungen mit weiterführenden Dokumenten oder Darstellungen etc.; *brushing* erlaubt die Markierung von Datenteilmengen, die sich simultan auf andere Darstellungen auswirkt.

2D-, $2\frac{1}{2}$D- und 3D-Darstellung: Auch wenn wir davon ausgehen, den drei-dimensionalen (3D-) Raum sehen zu können, beruht unser Seheindruck auf der Fusion zweier 2D-Bilder von der Netzhaut unserer Augen. D.h. unsere Tiefenwahrnehmung beschränkt sich auf die Entfernungseinschätzung zum nächstentfernten blickdichten Objekt. Das prinzipielle Unvermögen verdeckte Dinge (dahinter) zu erfassen, wird gelegentlich mit dem Dimensionsausdruck „$2\frac{1}{2}$D“ reflektiert.

So genannte 3D-Darstellung bezeichnet tatsächlich stereoskopische Anzeigegeräte, die zwei Bilder generieren und für ihre augengerechte Sichtbarkeit sorgen. Verschiedene Verfahren sind heute technisch ausgereift und verfügbar, aber wegen des Aufwandes und der damit einhergehenden Einschränkungen (z.B. meist Brillenzwang) nur bedingt verbreitet.

2D-Darstellungen mit perspektivischen 3D-Effekten, photorealistischen Beleuchtungseffekten und Schattenwürfen sind, dank günstiger und leistungsfähiger Graphikprozessoren, leicht erzeugbar und daher auch stark verbreitet. Die Tiefenwahrnehmung sollte aber nicht überschätzt werden: Sie ist auf Objekte beschränkt, die dem Betrachter vertraut sind. Ferner ist sie nicht sehr präzise und leicht verwirrbar – wie man an berühmten optischen Trugbildern, z.B. dem Neckar-Würfel, sehen kann.

Visuelle Attribute: Für statische 2D-Darstellungen gibt es nach Jacques Bertin (1982) prinzipiell verschiedene **visuelle Variablen**, die konstruktiv zur Generierung graphischer Darstellungen eingesetzt werden können:

- die Position auf der Ebene,
- die Ausdehnung (Länge, Fläche, Volumen),
- den Helligkeitswert,
- die Farbe,

- die Musterung oder Textur,
- die Richtung oder Orientierung sowie
- die Form des Elementes.

Sie alle sind geeignet, Raumregionen zu markieren und zu differenzieren.

Kartographische Metaphore: Durch kulturelle Prägung sind wir mit der systematischen räumlichen Anordnung von abstrakten Dingen eng vertraut: Ähnliche Dinge werden wie benachbarte Dinge verstanden und behandelt. Dies erzeugt Ordnung und hilft beim Suchen und Navigieren. Zum Beispiel kann es den gemeinsamen Einkauf von Äpfeln und Birnen, b.z.w. Schrauben und Nägeln ganz praktisch beschleunigen, da Ähnliches benachbart zu erwarten ist.

Auch der Umgang mit geographischen Karten ist uns durch Landkarten und Stadtpläne sehr vertraut. Diesen Umstand machen sich zahlreiche Visualisierungsverfahren systematisch zunutze. Sehr erfolgreich ist hier die 2D selbstorganisierende Merkmalskarte (SOM), die in Abschnitt 5.9 erläutert wird. Beispiele sind das WEBSOM-Projekt, das eine komprimierte Karte von Artikeln aus Newsgruppen erstellt (Kohonen et al. 2000) und das System von Skupin (2002) zur Verortung von Konferenzpapieren. In Kap. 7 und 8 werden Entwicklungen dargelegt, die dieses Konzept in einen besonderen Raum, den hyperbolischen Raum, integrieren.

3.1 Einige multivariate Visualisierungstechniken

Im Folgenden wird eine Reihe von bekannten und weniger bekannten Darstellungformen von n-dimensionalen Datensätzen vorgestellt. Für eine weiterführende Darstellung sei z.B. auf das Tutorial von Keim (2001) oder die Textesammlung in Card et al. (1999) und Fayyad et al. (2002) verwiesen.

Die Beispielvisualisierungen beziehen sich, wenn nicht anders angegeben, auf den in Tab. 3.1 beschriebenen Automodelldatensatz (Henderson und Velleman 1981). Er umfasst verschiedene Datentypen: Neben dem Nominalwert Modellname und dem kategorial skalierten Hersteller-

Origine-land	Model Name	Milage [mi/gal]	Weight [lb]	Gear-ratio	Power [hp]	Disp. [inch^3]	Cyl. #
USA	AMC Concord D/L	18.1	1.705	2.73	120	258	6
USA	AMC Spirit	27.4	1.335	3.08	80	121	4
Germany	Audi 5000	20.3	1.415	3.9	103	131	5
Germany	BMW 320i	21.5	1.3	3.64	110	121	4
USA	Buick Century Special	20.6	1.69	2.73	105	231	6
USA	Buick Estate Wagon	16.9	2.18	2.73	155	350	8
USA	Buick Skylark	28.4	1.335	2.53	90	151	4
USA	Chevette	30	1.0775	3.7	68	98	4
USA	Chevy Caprice Classic	17	1.92	2.41	130	305	8
USA	Chevy Citation	28.8	1.2975	2.69	115	173	6
USA	Chevy Malibu Wagon	19.2	1.8025	2.56	125	267	8
USA	Chrysler LeBaron Wagon	18.5	1.97	2.45	150	360	8
Japan	Datsun 210	31.8	1.01	3.7	65	85	4
Japan	Datsun 510	27.2	1.15	3.54	97	119	4
Japan	Datsun 810	22	1.4075	3.7	97	146	6
USA	Dodge Aspen	18.6	1.81	2.71	110	225	6
Japan	Dodge Colt	35.1	0.9575	2.97	80	98	4
USA	Dodge Omni	30.9	1.115	3.37	75	105	4
USA	Dodge St Regis	18.2	1.915	2.45	135	318	8
Italy	Fiat Strada	37.3	1.065	3.1	69	91	4
USA	Ford Country Squire Wagon	15.5	2.027	2.26	142	351	8
USA	Ford LTD	17.6	1.8625	2.26	129	302	8
USA	Ford Mustang 4	26.5	1.2925	3.08	88	140	4
USA	Ford Mustang Ghia	21.9	1.455	3.08	109	171	6
Japan	Honda Accord LX	29.5	1.0675	3.05	68	98	4
Japan	Mazda GLC	34.1	0.9875	3.73	65	86	4
USA	Mercury Grand Marquis	16.5	1.9775	2.26	138	351	8
USA	Mercury Zephyr	20.8	1.535	3.08	85	200	6
USA	Olds Omega	26.8	1.35	2.84	115	173	6
France	Peugeot 694 SL	16.2	1.705	3.58	133	163	6
USA	Plymouth Horizon	34.2	1.1	3.37	70	105	4
USA	Pontiac Phoenix	33.5	1.278	2.69	90	151	4
Sweden	Saab 99 GLE	21.6	1.3975	3.77	115	121	4
Japan	Toyota Corona	27.5	1.28	3.05	95	134	4
Sweden	Volvo 240 GL	17	1.57	3.5	125	163	6
Germany	VW Dasher	30.5	1.095	3.7	78	97	4
Germany	VW Rabbit	31.9	0.9625	3.78	71	89	4
Germany	VW Scirocco	31.5	0.995	3.78	71	89	4

Tabelle 3.1: Multivariater Automarkendatensatz nach Henderson (1981). Er umfasst das primäre Herkunftsland, den Markennamen, die Kraftstoffeffizenz (*milage*) in miles/gallon, das Gewicht in 1000 lbs, das Getriebeverhältnis im höchsten Gang, die Motorleistung in PS, den Hubraum (*displacement*) in Kubikinch und die Zylinderzahl von 38 Automodellen.

land gibt es sechs kontinuierlich skalierte Variablen, siehe Tabellenlegende 3.1.

Standardgraphiken zur Visualisierung von insbesondere uni- und bivariaten Datensätzen gibt es in vielen verfügbaren Variationen, z.B. in vielen Statistikpaketen und Tabellenkalkulationsprogrammen. Zu ihnen zählen

- 1D-Linien- und Balkengraphiken mit Varianten: horizontal oder vertikal; Balkenlänge proportional zum y-Wert oder zum relativen Gruppenanteil; mehrere $y_1, y_2, \ldots$-Werte durch parallele Balkengruppen (s. Abb. 3.2a) oder gestapelte Balken; zur Erzeugung von räumlichen Effekten werden Balken auch in 3D-Projektionen gezeigt (insbesondere zu Präsentationszwecken); Balken in Matrixanordnung, aufge-

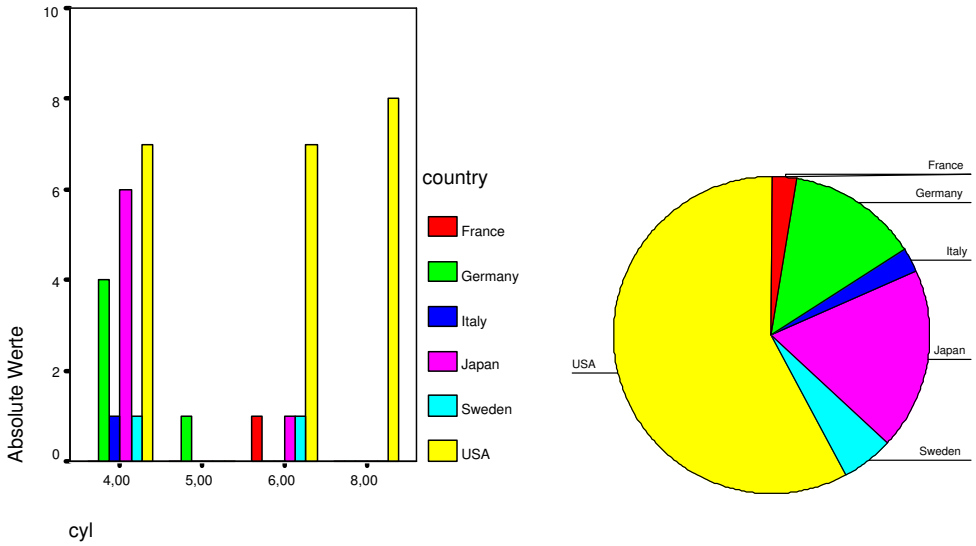


Abbildung 3.2: *(a, links:)* Ein Balkendiagramm der Modellzahl nach Zylinderanzahl und Herkunftsland. *(b, rechts:)* Das Tortendiagramm nach Herkunftsland gibt einen schnellen Anteilsüberblick.

reiht entlang zweier kategorialer Attribute;

- Tortengraphiken zur Visualisierung von Anteilen durch Kreissegmente; Variationen: Explosionszeichnungen für exponierte Darstellungen; Nutzung der Tortensstückhöhe als weiteren Parameter; Reduktion auf Kreisringfläche und Darstellung mehrerer $y_1, y_2, \ldots$-Anteile in konzentrischen Kreisringen;

- 2D-xy-Streudiagramme oder ***scatter plots***; Varianten sind überlagerte Darstellungen (Mehrfachplots mit $(x_1, y_1), (x_2, y_2) \ldots$);

- xy-Liniengraphiken sind durch Linien verbundene Streudiagramme; Varianten: Darstellung mehrerer Kurven; Fläche unter der Linie ist koloriert; mehrere Werte übereinander oder auch (als Flächen) gestapelt;

- „3D“-Plots: Parametrisierte xyz-Oberflächenplots für Daten mit 2D-Gitterstruktur; benachbarte Punkte werden durch Linien und/oder Flächen verbunden; z.T. komplexe 3D-Renderingmodelle; s. Abb. 5.20, S. 139).

Standardsoftwarepakete bieten hierzu vielfältige Optionen zur Attribuierung, Beschriftung, Kolorierung und zu Textureffekten.

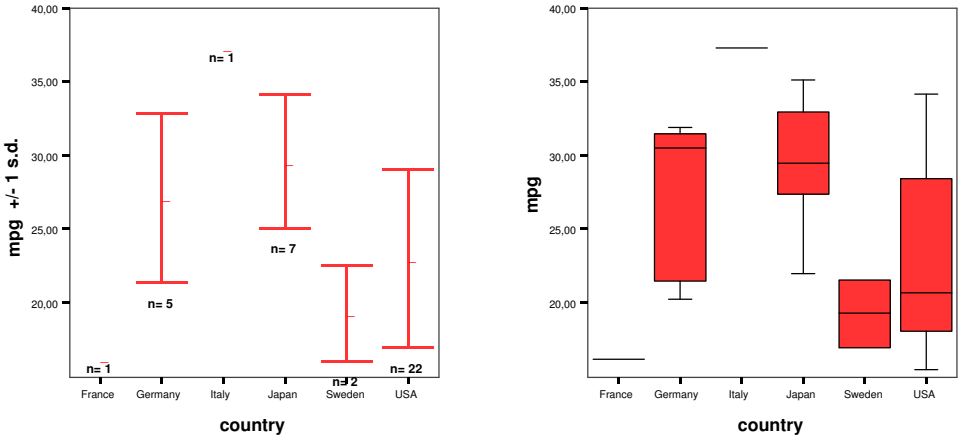


Abbildung 3.3: *(a, links:)* Die Milagedurchschnittswerte als Fehlerbalkenplot versus dem Herkunftsland. Da verschiedene Konventionen für Fehlerbreiten üblich sind, ist die spezifische Angabe des Typs wichtig, hier $\pm 1\, s.d.$ (nicht $s.e.$, 95 %-CI, etc., s. Abs. 4.2.2, 4.4). *(b, rechts:)* Der Boxplot zeigt eine andere Sicht auf dieselben Daten wie *(a)*. Die Endstriche notieren hier die Extremwerte. Die Box umschließt 50 % aller Fälle zwischen der $Q_{25\%}$- und der $Q_{75\%}$-Quantile (s. Abs. 4.2.3, S. 57) mit der Innenmarke des Medianwertes $Q_{50\%}$.

3.1.1 Zusatzinformationen: Fehler- und *Boxplots*

Für die Darstellung von statistischer Zusatzinformation haben sich Konventionen eingebürgert. Abb. 3.3a zeigt einen Fehlerplot, bei dem der Mittelwert und zusätzlich ein Unsicherheitsmaß, hier die Standardabweichung, durch „H“-förmige Markierungen notiert wird. Die statistischen Details werden in Kap. 4 genauer erläutert.

Eine hochaggregierte Darstellung der Datenverteilung bietet der **Boxplot**, s. Abb. 3.3b. Im Gegensatz zur Darstellung des Mittelwertes und der Standardabweichung in Abb. 3.3a, zeigt er eine Ausreißer-robuste Ansicht der Daten. Er markiert das aufgespannte Intervall, die „Mitte“ (Median) und die „mittleren“ 50 % der Daten ($Q_{25\%}$- und $Q_{75\%}$-Quantile) durch den Kasten und den Markern. Weitere Erläuterungen siehe Abs. 4.2.3, S. 58.

Es gibt natürlich viele Speziallösungen, z.B. werden in der Finanzbranche die Tagesverläufe von Aktienkursen häufig verdichtet durch kompakte vertikale *High-Low*-Striche mit kleinen eingehenden *open*- und ausgehenden *close*-Markierungen notiert, um die Tagesvariabilität des Börsengeschehens zu erfassen.

3.1.2 *Scatterplots*-Matrix

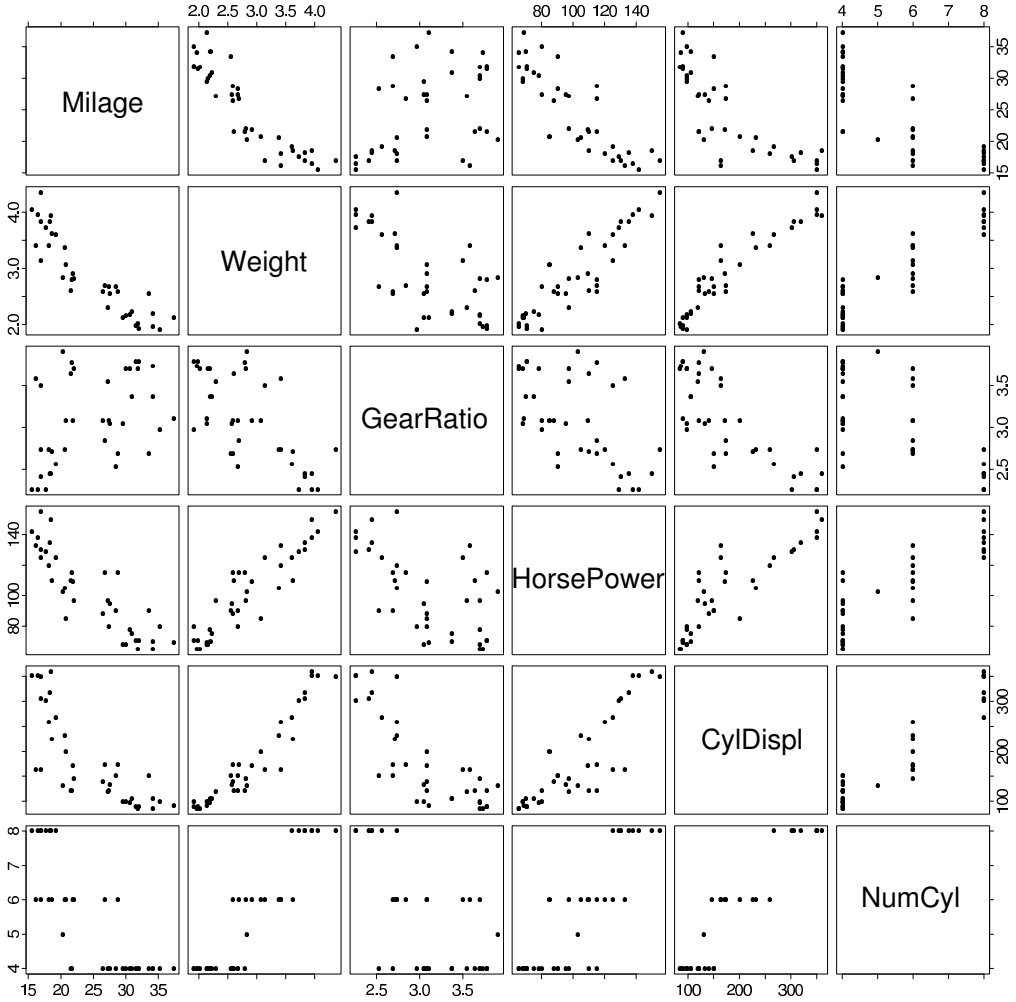


Abbildung 3.4: Scatterplots-Matrix für den Autodatensatz in Tab. 3.1.

Eine Scatterplots-Matrix ist eine quadratisch angeordnete Kollektion von kleinen Streudiagrammen, deren i, j-tes Element einen Scatterplot

vom i-ten versus dem j-ten Merkmal zeigt. Diagonalelemente zeigen das jeweilige Merkmalshistogramm oder sind einfach ein Platzhalter für den Merkmalsnamen. Diese Matrix gibt einen schnellen Überblick über die Datenverteilungen und Paarkorrelationen. Steigt die Merkmalszahl, wird die gesamte Matrix unhandlich. Treten gleiche Wertepaare mehrfach auf, bleibt dies verborgen. Insbesondere bei Binärdatenpaaren wird das Streudiagramm unbrauchbar. Abhilfe schafft die Addition eines geringen Rauschens (*jitter*), das eine Verschmierung der Daten und damit eine visuelle Verbreiterung der Punkte erzeugt. Bei sehr hoher Datenzahl muss eine *Subset*-reduktion erfolgen oder es müssen Intensitätsbilder durch 2D-Histogrammbildung erzeugt werden.

Klare Korrelationen können in Abb. 3.4 unschwer identifiziert werden: z.B. zwischen Fahrzeuggewicht und Motorleistung, Hubraum und Zylinderzahl und negativ korreliert zur Milage.

3.1.3 Parallele Koordinaten und *Andrews-Plots*

Betrachtet man eine Datentupel, so ist dies ein Punkt im hochdimensionalen Raum. Die Idee der **parallelen Koordinaten** ist die Umgestaltung der normalerweise senkrechten Achsen in parallele Achsen (Inselberg 1985).

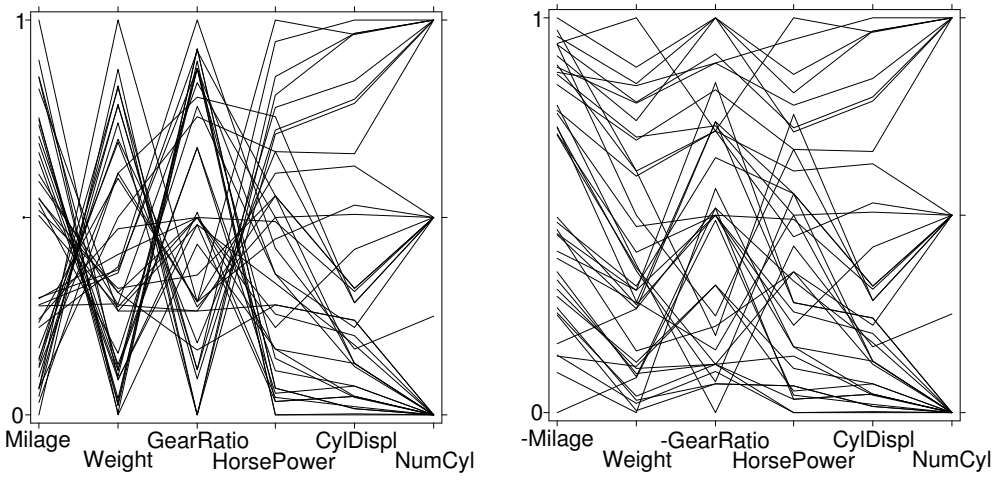


Abbildung 3.5: Parallele Koodinaten Darstellung. *(a, links:)* Originale Reihenfolge. *(b, rechts:* Nach Invertierung zweier Achsen (Milage, GearRatio) entsteht ein korrelierteres Bild.

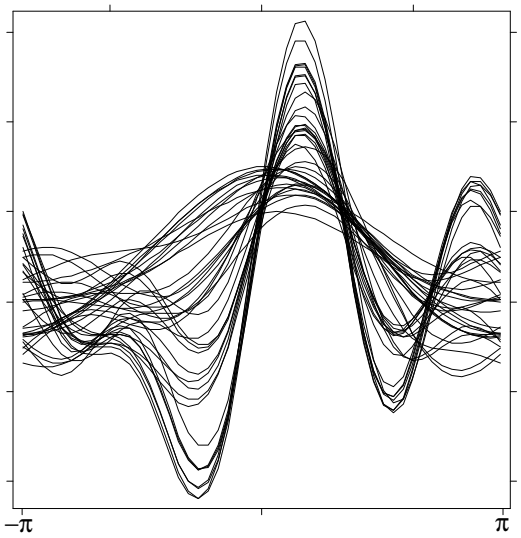


Abbildung 3.6: Andrews-Plot. Jedes Automodell ist als Kurve gezeichnet, wobei die Wertausprägungen die Frequenz und Phasenlage bestimmen.

Abb. 3.5 zeigt das Ergebnis, nachdem alle Merkmale auf die [0,1] Einheitsintervalle abgebildet wurden (Abs. 2.1.1) und die originalen Raumpunkte nun als Linienzüge durch alle Achsen gezeichnet sind.

Die drei letzten Koordinaten sind klar korreliert, wohingegen die ersten in Abb. 3.5a links sich kreuzen, also offensichtlich negativ korreliert sind. Durch Invertierung der *Milage* und *GearRatio* werden die Linien (*rechts*) besser entwirrt und die Korrelationen werden klarer erkennbar. Die Achsanordnung kann entscheidend für die Erkennbarkeit von Korellationen sein.

Der **Andrews Plot** interpretiert die standardisierten Werte als Spektralkomponenten und zeigt eine Zeitkurve. Die Kurve in Abb. 3.6 wird durch die trigonometrische Funktion (Andrews 1972)

$$f_{\mathbf{X}_i} = 2^{-0.5}\, x_{i,1} + x_{i,2}\sin(t) + x_{i,3}\cos(t) + x_{i,4}\sin(2t) + x_{i,4}\cos(2t) + \cdots \quad (3.1)$$

gebildet und im Intervall $[-\pi, \pi]$ geplottet. Die Zuordnung der Komponenten ist bedeutungsvoll und sollte vom wichtigen zum unwichtigen Parameter erfolgen. Eine wichtige Eigenschaft ist die Erhaltung der Paardistanzen zweier Punkte

$$\int_{-\pi}^{\pi} (f_{\mathbf{X}_1}(t) - f_{\mathbf{X}_2}(t))^2 dt = \pi ||\mathbf{x}_1 - \mathbf{x}_2||_{L2}^2. \quad (3.2)$$

D.h. zwei ähnliche Punkte werden auf ähnliche Kurven abgebildet. In Abb. 3.6 erkennt man deutlich Bänder, die auf einige wenige Cluster hindeuten.

3.1.4 Ikonographische Darstellungen

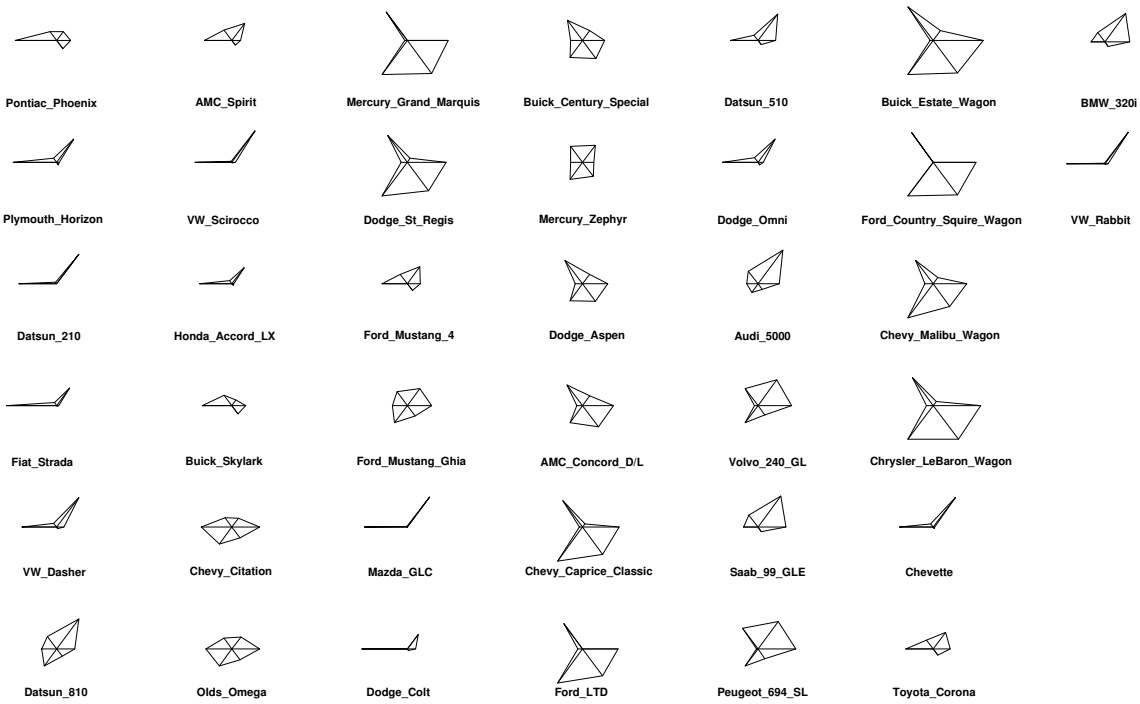


Abbildung 3.7: Ikonographische Darstellung: *Star Glyphs* für die Automodelle in Tab. 3.1.

In einer ikonographischen Darstellungen wird jeder Datentupel durch ein Bild oder Glyph repräsentiert. Stellt man sie gitterartig nebeneinander, kann man sie vergleichend beobachten.

Sterndarstellungen – star plots: Ein Einheitskreis wird durch n-viele (Anzahl der Komponenten) Radialstrahlen gleichförmig geteilt. Jedem Strahl wird systematisch eine normierte Komponente zugeordnet. Die konkreten Werte werden dann auf dem Strahl abgetragen und durch Linien tangential verbunden (Fienberg 1979).

Abb. 3.7 zeigt, dass die Ähnlichkeit zwischen Modellen durch die Formähnlichkeit der Glyphen visuell leicht erkannt werden kann. Die Methode wird ungeeignet, wenn zu viele Komponenten oder zu viele Beobachtungen verglichen werden sollen.

Chernoff-Gesichter nutzen den Vorteil unserer natürlich geschärften Wahrnehmungsfähigkeit für menschliche Gesichter. Chernoff (1973) schlug Ge-

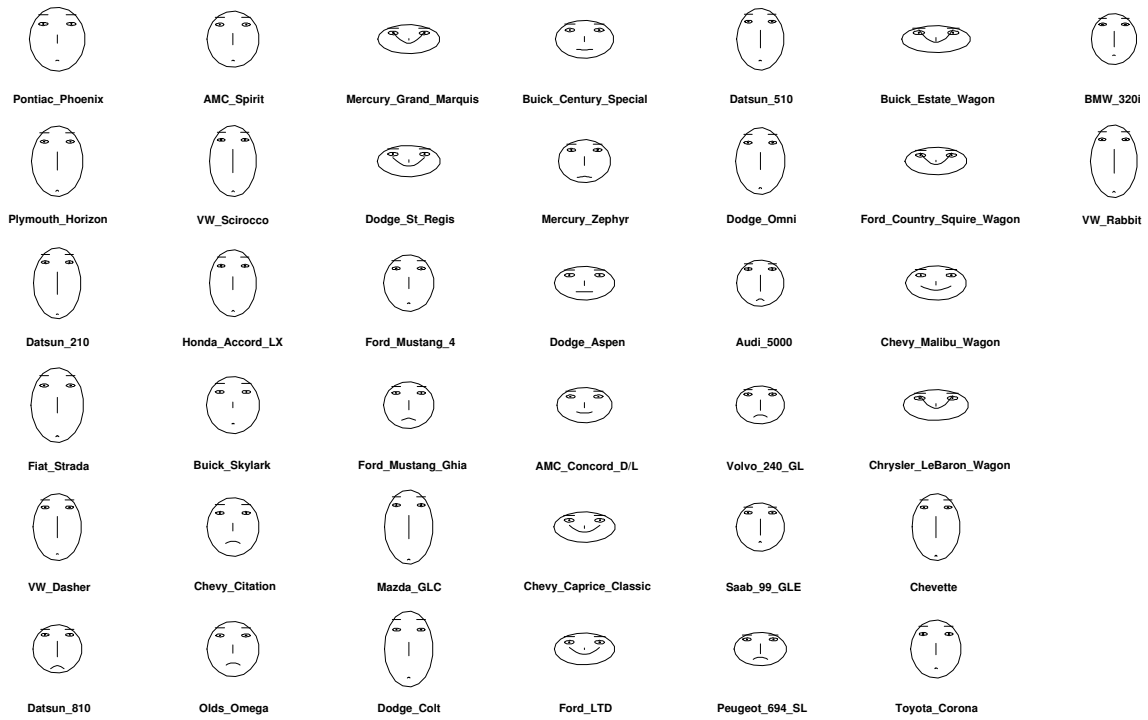


Abbildung 3.8: Ikonographische Darstellung: Chernoff-Gesichter repräsentieren die sechs Attribute jedes Automodells in Tab. 3.1.

sichtskarrikaturen vor, die bis zu 15 Attribute (und mehr) repräsentieren und an Datenkomponenten geknüpft werden können. In absteigender Wichtigkeit sind dies

Gesichtsfläche;
Gesichtsform;
Nasenlänge;
Mundort;
Mundkrümmung (Lachmund);
Mundbreite;
Ort, Abstand, Winkel, Form und Breite der Augen;
Pupillenort;
Ort, Winkel und Breite der Augenbrauen.

Abb. 3.8 gibt den Automodellen Gesichter. Vergleicht man die Chernoff-Gesichter mit den Sternglyphen, stellt man fest, dass der *star plot* den Vorteil der Neutralität und der besseren Vergleichbarkeit genießt. Da einige Gesichtsmerkmale emotionstragend sind, z.B. der lachende oder traurige Mund, die entspannten oder aggressiven Augenbrauen, besteht die große

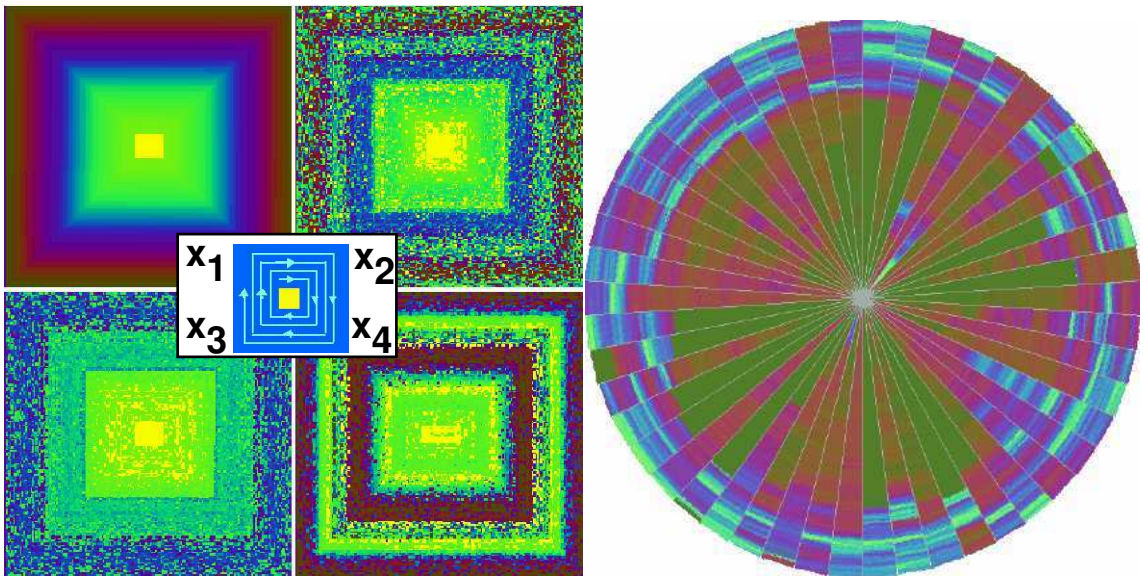


Abbildung 3.9: Pixelorientierte Visualisierung in *(a, links)* Spiral- und *(b, rechts)* in Kreissegmentform. *(Links:)* Vier Attribute eines Datensatzes werden, z.B. durch eine Datenbankabfrage ausgewählt: $x_{1i}, \ldots, x_{4i}$, die Reihenfolge wird hier nach dem ersten Merkmal x_{1i} sortiert. Nach Normierung und Farbkodierung werden die Merkmalsspalten spiralig in die quadratischen Felder eingeordnet. Merkmalskorrelationen machen sich als sich partiell wiederholende Muster bemerkbar. *(Rechts:)* Die Einreihung erfolgt hier innerhalb jedes Kreissegmentes auf einer mäanderförmigen Linie (zick-zack von innen nach außen). Visualisiert werden hier 50 Aktienkurse der Frankfurter Börse. Durch chronologische Aufreihung werden anhand der Kreisringstrukturen deutliche Synchronisationen in den späten Phasen des Zeitraums 1994–1995 erkennbar.

Gefahr, dass mehr wahrgenommen wird, als vorliegt. Zudem sind die Komponentenzuordnungen willkürlich und die Wertausprägungen nicht quantitativ beurteilbar.

Ikonographische Darstellungen können in Streudiagrammen integriert werden. Weitere Formen sind die so genannten *stick figures*, kleine Strichfiguren, deren Verbindungswinkel parametergesteuert wird. *Color icons* stellen das multivariate Datenobjekt als ein Rechteck dar, das aus kleinen Farbfeldern zusammengesetzt ist.

3.2 Pixelorientierte Visualisierungen

Steigt die Datenmenge, ist es sinnvoll, den Platz pro Datum zu minimieren. Pixelorientierte Techniken sind konsequent und reduzieren die Fläche bis auf die Größe eines Pixel, wobei jede Wertausprägung farbkodiert wird. Im Gegensatz zu den *color icons* wird die Objekt-Merkmals-Relation umgekehrt, d.h. ein Merkmal j von allen Daten füllt eine Fläche F_j, wobei die Werte systematisch entlang einer die Fläche F_j füllenden Linie aufgereiht werden (Keim und Kriegel 1994). Abb. 3.9 illustriert das Konzept anhand zweier Beispiele (aus Keim 2001).

3.3 Die interaktive „Tabellenlupe“

Die pixelorientierten Techniken nutzen die Anzeigefläche maximal aus. Der Zusammenhang der Merkmale auf Datenobjektebene ist aber nicht mehr direkt erkennbar. Dieses Problem löst die „Tabellenlupe“ oder ***table lens*** (Rao und Card 1994) durch eine elegante Kombination von Interaktion und einer möglichen Flächenreduktion auf Linienstärke unter einem Pixel Höhe pro Datenobjekt (s.u.).

Konzeptionell ist es eine graphische Darstellung von Tabelleneinträgen mittels Balken geeigneter Länge und Positionierung. Sie kann die textuelle Eintragsdarstellung *(i)* ergänzen oder *(ii)* ersetzen. Die Balkenlänge wird proportional zum jeweiligen Wert bestimmt (normiert auf die Intervallbreite, die von den Spalteneinträgen aufgespannt ist). Spalten mit kategorialen Datentypen werden zuvor wertmäßig enumeriert.

Die Tabellenlupe ermöglicht, die Daten zum einen ganz oder partiell im Graphikmodus *(i)* als Text+Balken oder im Modus *(ii)* nur mit Balken zu betrachten. Die entscheidende Zutat ist die Interaktionsform der Zeilenumsortierung, ausgelöst durch einen Mausklick auf die Kopfzeile der gewünschten Spalte. Ein weiteres Klicken invertiert die Sortierrichtung (auf- bzw. absteigend).

Damit ist die Tabellenlupe auch als integriertes Explorationswerkzeug nutzbar. Die einfache Benutzersteuerung wird anhand einiger exemplarischer Analyseschritte aufgezeigt, siehe Abb. 3.10a–h. Der Datensatz umfasst hier 392 Autotypen im Zeitraum 1971–1983, er ist strukturgleich mit dem in Tab. 3.1, aber größer und (leider) typanonymisiert.

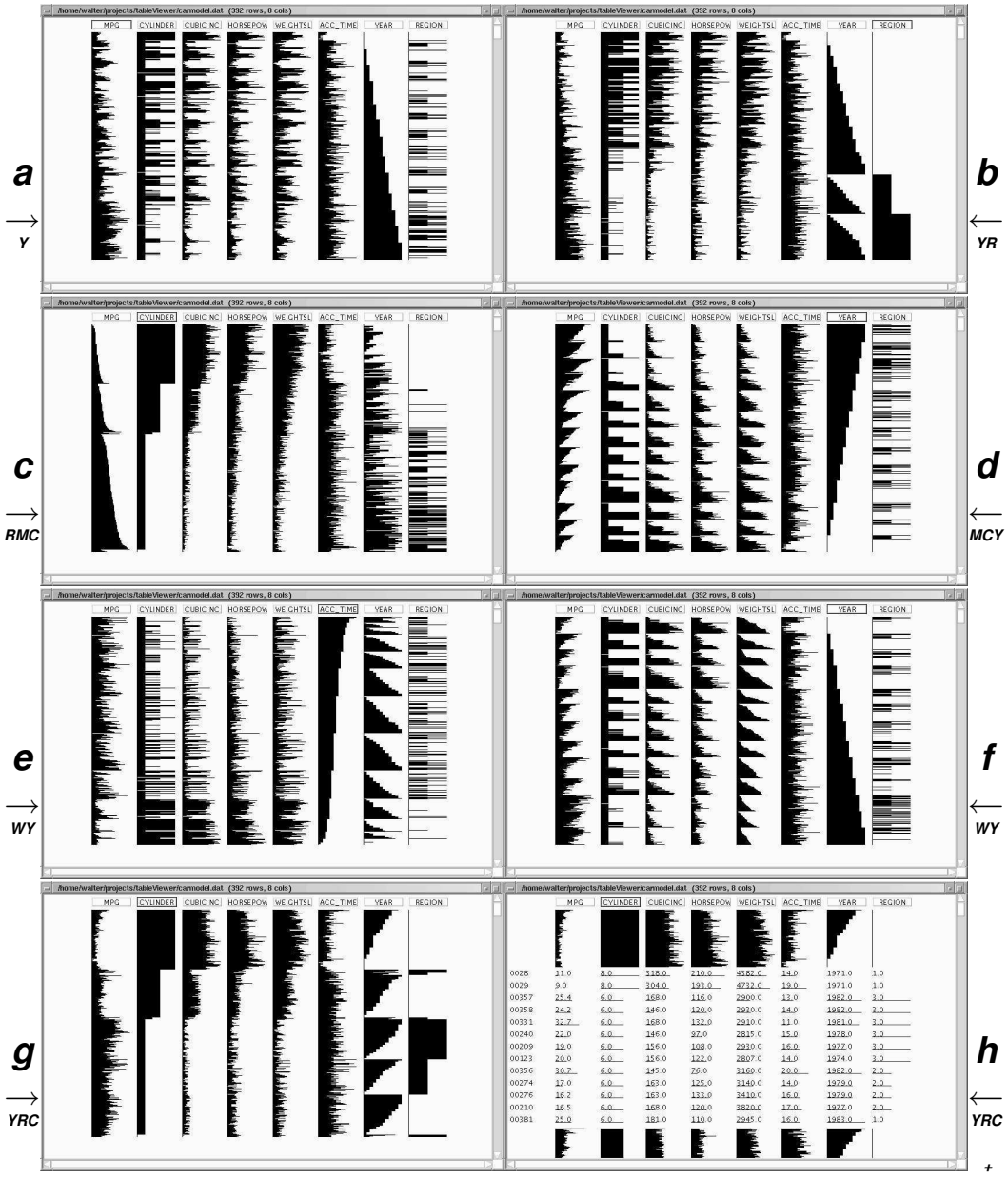


Abbildung 3.10: (*a–h*) Ansichten der Tabellenlupe nach diversen sequenzstabilen Umsortierungen. Datensatz mit 392 Automodellen und den acht Attributen ***M**pg, **C**ylinder* {3, 4..8}, *cubic**I**nch, **H**orsePower, **W**eightLb, **A**ccelerationTime, **Y**ear* [1971, 1983], ***R**egion* { US, EU, J }. Die Sequenz relevanter Sortierdurchgänge ist mit deren Attributinitialen markiert, s.a. Text.

Die ursprüngliche chronologische Datenreihung ist in Teilbild *a* abgebildet. Abb. 3.10b entsteht durch Spaltensortierung nach *Region*. Eine wich-

tige Besonderheit ist, dass der Sortierprozess indexstabil ist, d.h. er erhält die Datenreihung von wertgleichen Gruppen. Dadurch bleiben alle Automodelle aus der *Region=0* (=USA) weiterhin chronologisch geordnet, ebenso die des Blocks *Region=1* (europäische Herkunft) und *Region=2* (Japan).

In Abb. 3.10a zeigt sich an der geraden, stufenartigen Reihung der *year*-Spalte, dass pro Jahr etwa gleich viele Autos Eingang in den Datensatz gefunden haben. Nach dem *Region*-Sortieren teilt sich die Verteilungskurve in drei bedingte Verteilungskurven: bedingt nach Herkunft. An der konvexen unteren kann man ablesen, dass gegen Ende des Betrachtungszeitraums [1971, 1983] der Anteil japanischer Produkte auf Kosten des amerikanischen Marktanteils zugenommen hat.

Abb. 3.10c ist zwei Clicks entfernt: Die Sortierung nach *Mpg* (*miles per gallon*) und *Cylinder* zeigt wieder mehrere bedingte, kumulative Verteilungsfunktionen – hier die Kraftstoffergiebigkeit nach der Zylinderzahl. Die *S*-förmigen Kurven sind typisch für eine Normalverteilung. An der Zylinderspalte erkennt man die Dominanz der Vierzylinderkonstruktion und ein paar Dreizylinder, die sich aber verbrauchsmäßig nicht behaupten (geringe *Mpg*).

Sortiert man nun nach Zeit, erkennt man an Abb. 3.10d die Folgen der Ölkrise: im Lauf der Jahre sanken Zylinderzahl, Hubraum (*Cubicinch*), Verbrauch und Leistung (*Horsepower*). Der Trend ist klar, viele Details lassen sich mit wenigen Sortierungsschritten studieren. Abb. 3.10e entsteht durch Sortierung nach Zylinderzahl, Jahr, dann Beschleunigung (*Acctime*), Abb. 3.10f durch Sortierung nach Gewicht (*Weightlb*) und Baujahr.

Abb. 3.10g und h entstehen durch Sortieren nach Jahr, Herkunft und Zylinderzahl. Es wurden auch einige Achtzylindermotoren außerhalb der USA gebaut: wenige, verhältnismäßig spät und mit etwas geringerem Verbrauch. Diese Details deuten sich in dem obersten *Region*>1 Streifen bereits an. In der rechten Graphik ist dieser schmale Streifen in tabellarischen Originaldaten expandiert (leider sind die Modellnamen nicht explizit verfügbar und werden links durch die IDs ersetzt). Vergleicht man die vollständige Balkendarstellung *(ii)* in Abb. 3.10g und das partielle Hineinzoomen *(i)* in Abb. 3.10h, fällt der gute Kompressionsfaktor der Tabellenlupe auf. Nur wenige Textzeilen passen in den Screenshot und die zweite Datenhälfte ist schon nicht mehr sichtbar.

Die Hauptcharakteristika der Tabellenlupe (*table lens*) sind

- Ein vollständiger Datentupel lässt sich auf einer dünnen Zeile abbilden, ohne dass Aggregationsmechanismen bemüht werden müssen. Bei Bildschirmdarstellung ist die Minimalgröße von einem Pixel durchaus noch komfortabel (Experimente zeigen, dass mittels blockweiser vertikaler Mittelung und Graustufung eine weitere Verdichtung von 4:1 bis 16:1 noch sinnvoll ist);
- Mehrere hundert Datensätze können damit auf einen Blick verglichen werden (im Extremfall bis 10^4);
- Durch interaktiv ausgelöste Sortierungsschritte ergeben sich direkt ablesbare Verteilungsfunktionen;
- Durch eine einfache Sequenz von Sortierungskriterien werden bedingte Verteilungsfunktionen sichtbar. Sie sind sinnvoll und übersichtlich, wenn die Bedingung den Datensatz in wenige, etwa ausgewogen große Gruppen aufteilt. Die Bedingung muss also eine kleine Kardinalität besitzen und kann als ein Attribut oder als kartesisches Produkt zweier Merkmale bestimmt werden. Zuletzt wird interaktiv nach diesen bedingenden Merkmalen sortiert, was dann eine direkte visuelle Abhängigkeitsanalyse von zwei bis drei Merkmalen unterstützt;
- durch Einblenden von alphanumerischer Darstellung werden ein interaktives Hineinzoomen und eine integrierte Darstellung der kompletten Originaldaten ermöglicht.

3.4 Integrierte *Missing-Value*-Statistik und Assoziationsanalyse

Insbesondere bei medizinischen Datensätzen ist das Auftreten von fehlenden Werten (*missing values*) nicht ungewöhnlich. Für die folgende Datenauswahl, Analyse und Interpretation ist die Kenntnis *(i)* von möglichen Korrelationen der Abwesenheit zweier Merkmale, *(ii)* der Verteilung von Merkmalen und *(iii)* der Korrelationen zwischen der Abwesenheit eines Merkmals und der Verteilung anderer Merkmale bedeutsam.

(iv) Ob zwei Merkmale eine lineare Korrelation aufzeigen oder ob Verteilungen gänzlich unabhängig sind, kann man durch entsprechende Assoziationsanalysen ermitteln (Vorstellung in Abs. 4.12 und 4.13).

(v) Ferner ist man an der Stichhaltigkeit aller Aussagen interessiert, die sich durch Berechnung von Signifikanzmaßen validieren lassen (p-Werte).

(vi) Um den Überblick über große Ergebnismengen zu beschleunigen, ist die Einfärbung der Ergebnisse nach Wert oder Signifikanz eine Hilfe.

Ein Lösung, die all diese Wünsche integriert, ist in Abb. 3.12 (S. 47) dargestellt. Das Werkzeug verbindet die relevanten statistischen Algorithmen (s. Kap. 4) und macht die Präsentation der Ergebnisse als auch die Steuerung des Analyseprozesses durch einen Standard-Web-Browser möglich. Der zweistufige Ablauf wird in Abb. 3.12 erläutert.

3.5 Integrierte Assoziationsanalyse

Das in Abb. 3.13 (S. 48) dargestellte Analysewerkzeug dient der interaktiven Assoziationsanalyse von Merkmalspaarungen. Es vereinigt Konzepte der Scatterplot-Matrix, der Farbkodierung von Merkmalsassoziationen und der Detailinspektion von selektierten Merkmalspaaren. Verschiedene Anzeigemodi stehen für das i, j-te Matrixelement zur Auswahl:

- Anzeige von Assoziationsmaßen, die später in Abs. 4.12 erläutert werden: Pearson's Korrelationskoeffizient $|r|$, Cramers V oder χ^2. In Abb. 3.13 kommt der normierte Unsicherheitskoeffizient $U(x_i|x_j)$ durch Farbkodierung zum Ausdruck (Gl. 4.80);

- diverse Farbpalletten, z.T. in nicht-linearen Transformationen. Links in Abb. 3.13 ist der aktuelle Farbcode dargestellt;

- Mittels Mausselektion in der Matrix wird ein Streudiagramm $x_i(x_j)$ im unteren Teil eingeblendet (bzw. für $i = j$ ein Histogramm). Zusätzlich werden die Assoziationskennzahlen für i, j tabellarisch rechts unten in Abb. 3.13 gelistet;

- Optional wird eine lineare Regression berechnet und gemeinsam mit Konfidenzbereichen in das Streudiagramm eingezeichnet. Die Details der Darstellung und der 95 %-Konfidenzbänder werden in Abb. 5.8 und Abs. 5.7.1 dargestellt.

3.6 Trellis-Darstellung

Trellis-Darstellungen sind systematische Anordnungen von Streudiagrammen. Im Gegensatz zur Scatterplot-Matrix werden nicht stets neue Merkmale ausgewählt, sondern sie bleiben konstant. Es werden jedoch systematisch Datenteilmengen in Abhängigkeit von anderen Merkmalen gebildet. Abb. 3.11 zeigt ein Beispiel mit einer Selektion auf einem Intervall (Zeit) und einer Kategorie (Region) pro Diagramm.

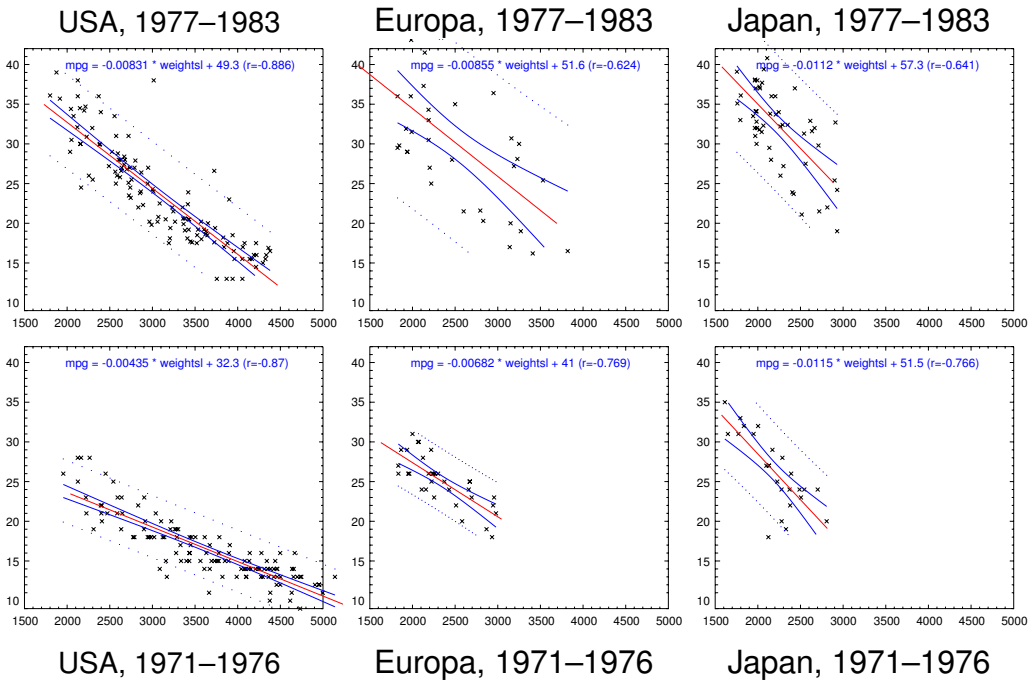


Abbildung 3.11: Trellis-Darstellung: Streudiagramm-Matrix mit Regressionsgeraden und Konfidenzintervallen. Hier ist die *milage* (Mpg) versus dem Gewicht für drei Herkunftsregionen (horizontal orientiert) und zu zwei Zeiträumen (<1977, $\geq$1977 vertikal orientiert) für den Automodell-Datensatz aus Abb. 3.10 dargestellt. Da die Achsen gleich skaliert sind, erkennt man klar Zusammenhänge: Z.B. haben neuere Modelle (*obere Reihe*) deutlich bessere Verbrauchseigenschaften (höhere *milage*); europäische Hersteller haben ihre Modellpalette im zweiten Zeitraum noch mehr diversifiziert als die anderen beiden Regionen, und Japan zeigt einen sich verringernden Vorsprung im Verbrauchsspektrum.

Im folgenden Kapitel werden die Grundlagen für die statistische Beschreibung und die Validierung von Hypothesen und Zusammenhängen

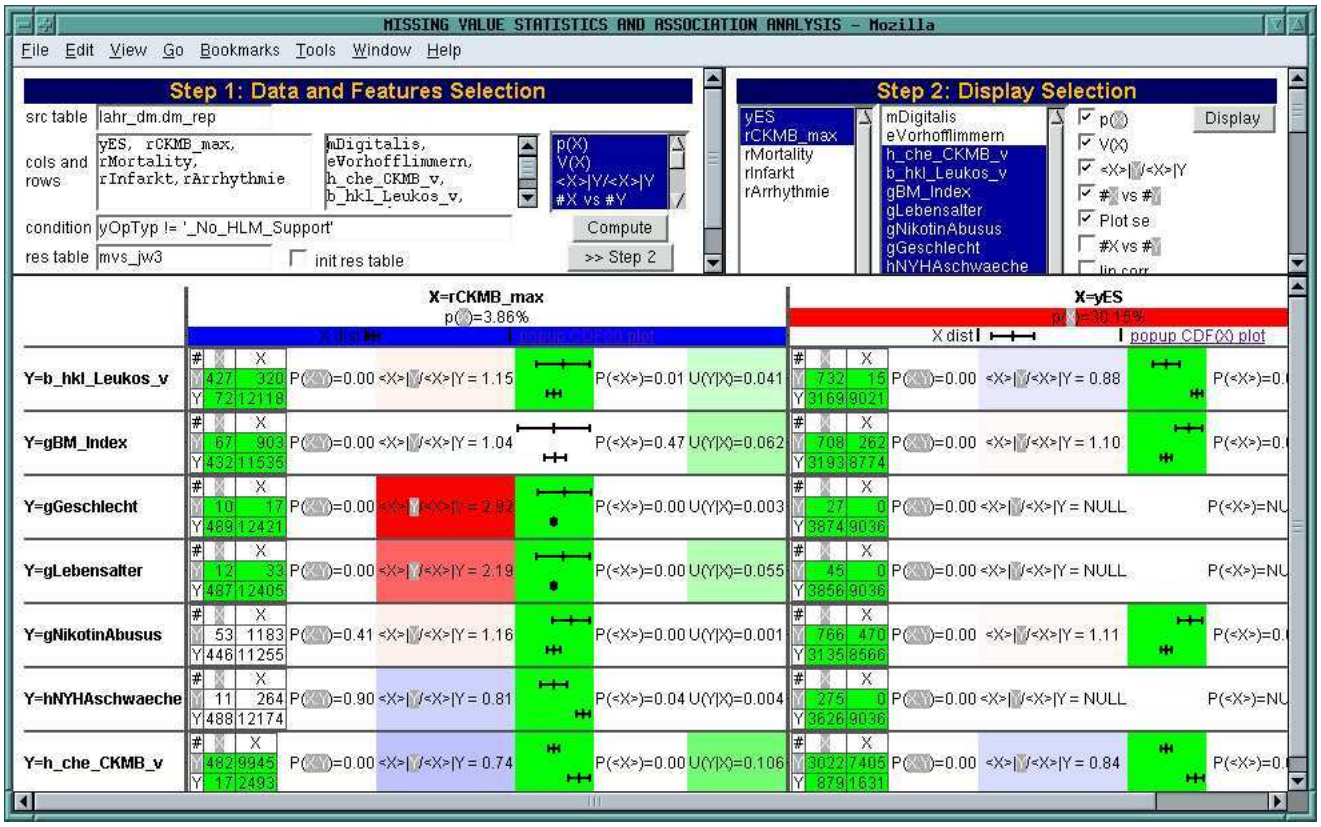


Abbildung 3.12: Die Ausgabe und zweistufige Steuerung dieses Werkzeugs zur Missing-Value- und Assoziations-Analyse ist als HTML-Interface ausgeführt: *(links oben:)* In der ersten Stufe werden die Daten und gewünschten Merkmale selektiert und Vorberechnungen veranlasst. *(rechts oben:)* Nachdem charakteristische Verteilungsparameter berechnet und in eine Datenbank geschrieben wurden, kann in einem zweiten Schritte eine Reihe von statistischen Informationen ausgewählt werden. Im *unteren Teil* werden diese in Matrixform präsentiert. Jede (große) Zelle enthält Einzelwerte, (innere) Tabellen und/oder Inline-Graphiken, die zu einer Merkmalskombination (x, y) gehört. Auch hier ist der Anzeigeplatz eine knappe Ressource, somit kann die Darstellung auf bestimme Merkmale (getrennt nach Spalten und Zeilen) und Inhalte beschränkt werden (via *list-* und *checkboxes rechts oben*). Für eine Teilmenge des medizinischen Datensatzes, der in Kap. 9 erläutert wird, sind für die ausgewählten Merkmalspaare folgende Informationen darstellbar: u.a. die Kontingenztabellen für Abwesenheit (durch Weiß-auf-Grau dargestelltes x oder y), der p-Wert (χ^2-Test), Mittelwertsvergleich $\bar{x}$ für die Gruppe mit und ohne y-Wert mit Standardfehler-Vergleichsgrafik und p-Wert (t-Test), sowie der normierten Vorwärtsentropie $U(x|y)$ (Gl. 4.80). Kennwerte und Verteilungsfunktionen der einzelnen Merkmale sind über die Kopfzeilen verknüpft (Hypertext-links).

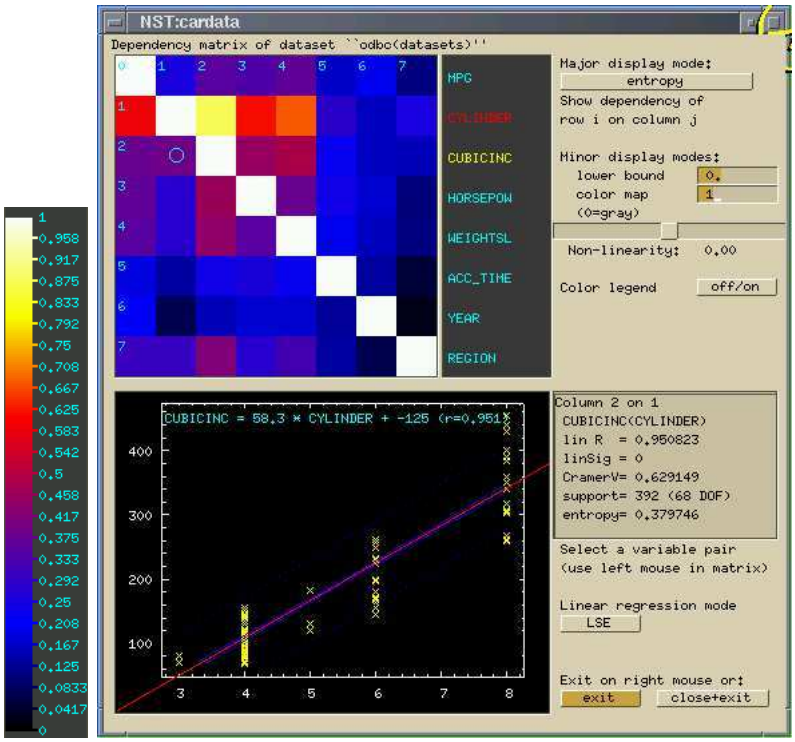


Abbildung 3.13: Analysewerkzeug zur interaktiven Assoziationsanalyse von Merkmalspaarungen – hier angewandt auf den Automodell-Datensatz von Abb. 3.10. Die Farbkodierung in der Merkmalsmatrix zeigt hier die normierte Vorwärtsentropie oder den Unsicherheitskoeffizient $U(x_i|x_j)$. Der Kreis markiert die selektierte Zeile *Cubicinch* ($i = 2$) und Spalte *Cylinder* ($j = 1$). Zu diesem Paar werden dann Assoziationsmaße und das Streudiagramm mit 95 % Konfidenzintervallen dargestellt (weitere Funktionsmerkmale dieser NEO/NST-Implementation s. Text). An diesem Streudiagramm kann man erkennen, dass *Cubicinch* weniger von *Cylinder* determiniert ist als umgekehrt, was sich in den unsymmetrischen Entropiewerten und Farben in i, j und j, i widergespiegelt.

gelegt. Es geht um Hypothesen und Zusammenhänge, die möglicherweise, aufgrund visueller Inspektion, iterativer Exploration und durch effektive Integration von Expertenwissen zutage getreten sind.

Kapitel 4

Statistische Grundlagen

„Wir Wissenschaftler nutzen Statistik oft wie ein Betrunkener, der den Laternenpfahl mehr als Stütze als zur Beleuchtung sucht.“

(Winifred Castle, britischer Statistiker)

Statistik beschäftigt sich mit dem Beschreiben und Interpretieren von Daten sowie dem Testen von Hypothesen anhand dieser Daten. Wo früher das Erzählen von Anekdoten und die Weitergabe persönlicher Erfahrung eine zentrale Rolle in mancher wissenschaftlichen Disziplin spielte, sind heute die formalen Ansprüche an Nachvollziehbarkeit von hoher Bedeutung. Hinzu kommt die Möglichkeit, viele Dinge auch quantitativ zu messen und auszuwerten. Zum Beispiel sind mit der Verbreitung der Disziplin *evidence based medicine* die Ansprüche an die statistische Validierung medizinischer Erkenntnisse drastisch gestiegen, so dass Autoren in medizinischen Journalen meist mindestens einen so genannten p-Wert pro Arbeit referieren, um die Glaubwürdigkeit ihrer Resultate zu untermauern.

In diesem Kapitel sollen wichtige statistische Grundlagen eingeführt werden. Die Kenntnis geeigneter statistischer Methoden ist das Fundament zur Bildung und Evaluierung von Hypothesen und Modellen und erlaubt damit solide Signifikanzbewertungen für das im KDD-Prozess entdeckte Wissen.

4.1 Zufallsexperimente, Wahrscheinlichkeiten und Verteilungen

Das Werfen eines Würfels ist das klassische Beispiel eines kategorialen Zufallsexperiments. Es resultiert ein *zufälliges* Ereignis A aus einer Menge von möglichen Ereignissen Ω. Der Ereignisraum Ω ist die Vereinigungsmenge der Elementarereignisse, hier z.B. der Wurf der Augenzahlen $1, 2, \ldots, 6$ mit der Kardinalität $|\Omega| = 6$.

Der klassische Wahrscheinlichkeitsbegriff ist axiomatisch geprägt und analysiert die Ereignismöglichkeiten. Die **Eintrittswahrscheinlichkeit** P_A eines Ereignisses A ist klassisch

$$P_A = \frac{\text{Anzahl für } A \text{ günstige Ereignisse}}{\text{Anzahl der Elementarereignisse}}, \tag{4.1}$$

hingegen ist die empirische Herangehensweise durch das Experiment geprägt.

Der statistische Wahrscheinlichkeitsbegriff lässt sich pragmatisch aus der **relativen Häufigkeit** herleiten. Die relative Häufigkeit $f(A)$ (engl.: *frequency*) eines Ereignisses A ergibt sich bei N-maliger Wiederholung des Experimentes aus der absoluten Häufigkeit $c(A) = m$ (engl.: *count*)

$$f(A) = \frac{c(A)}{N} = \frac{m}{N}. \tag{4.2}$$

Kann man einen (ggf. unbekannten) Prozess annehmen, der mit einer konstanten Wahrscheinlichkeit das Ergebnis A liefert, so definiert man die **Wahrscheinlichkeit** $P(A)$ als den asymptotischen Grenzwert für große N (Gesetz der großen Zahlen):

$$P(A) = \lim_{N\to\infty} f(A) = \lim_{N\to\infty} \frac{m}{N}. \tag{4.3}$$

Nun betrachten wir allgemein zwei Ereigniseintritte in einem Experiment. Denken wir zum Beispiel an das Einrasten der Roulettekugel in ein Feld. Ein Ereignis sei „rot“, das andere die Zugehörigkeit zu einem bestimmten Zahlenblock.

Die Wahrscheinlichkeit des gemeinsamen Auftretens zweier Ereignisse A und B schreibt man als **Verbundwahrscheinlichkeit** $P(A \cap B)$ meist

in der verkürzten Form $P(A, B)$ (gelegentlich auch als $P(A \wedge B)$. Die **bedingte Wahrscheinlichkeit** $P(A|B)$ beschreibt die Auftretenswahrscheinlichkeit von A, wenn vor der Beobachtung schon die Information des Ereigniseintritts von B vorliegt:

$$P(A|B) = \frac{P(A, B)}{P(B)}. \tag{4.4}$$

Im Kontext dieser Zusatzeinschränkungen spricht man auch von den **unbedingten**, den **a-priori-Wahrscheinlichkeiten** $P(B), P(A)$ und der **bedingten**, der **a-posteriori-Wahrscheinlichkeit**, z.B. $P(A|B)$.

Kennt man die a-posteriori-Wahrscheinlichkeit in die eine Bedingungsrichtung, kann man mit der **Bayes-Regel** die Gegenrichtung berechnen:

$$P(B|A) = \frac{P(A|B)\, P(B)}{P(A)}. \tag{4.5}$$

Dies entspricht auch der manchmal als **Multiplikationsregel** genannten Regel für Wahrscheinlichkeiten:

$$P(A, B) = P(A \cap B) = P(A|B)\, P(B) = P(B|A)\, P(A). \tag{4.6}$$

Ferner gibt es die Additionsregel:

$$P(A \cup B) = P(A \vee B) = P(A) + P(B) - P(A, B). \tag{4.7}$$

Die Bayes-Regel ist ein fundamentaler Baustein für die korrekte Behandlung von bedingten Wahrscheinlichkeitsmodellen (s. Abs. 5.1) und kann gelegentlich zu überraschenden Ergebnissen führen, wie das Beispiel im abgebildeten Kasten 4.1 verdeutlicht.

Die **statistische Unabhängigkeit** von Ereignissen ist ein wichtiges Grundkonzept und besagt, dass es keinen Unterschied für den Eintritt von A macht, ob B vorher eingetreten ist oder nicht.

$$P(A|B) = P(A) \quad \wedge \quad P(B|A) = P(B) \quad \Longleftrightarrow \quad A, B \text{ statistisch unabhängig} \tag{4.8}$$

Die Verbundwahrscheinlichkeit $P(A, B)$ ist dann gleich dem Produkt der a-priori-Wahrscheinlichkeit (auch aus Gl. 4.4, 4.5, 4.8).

$$P(A, B) = P(A)P(B) \quad \Longleftrightarrow \quad A, B \text{ statistisch unabhängig} \tag{4.9}$$

Abbildung 4.1: Anwendungsbeispiel für die Bayes'sche Regel

Was bedeutet es, wenn ein Test 95 % akkurat ist? Heißt dies, das ich mit 95 % Wahrscheinlichkeit die schlimme Krankheit (K_+) habe, wenn der Test positiv (T_+) ist? Klingt dies plausibel, so ist die richtige Antwort jedoch „Nein". Denn im Allgemeinen ist $P(K_+|T_+) \neq P(T_+|K_+)$ und hängt sehr von der a-priori-Infektionshäufigkeit $P(K_+)$ ab.

Sei $P(T_+|K_+) = P(T_-|K_-) = 0.95$ (damit ist die *false-negative-rate* = *false-positive-rate* 5 %; s.a. Abs. 5.7.4 Sensitivität und Spezifität) und die Krankheit träte selten auf $P(K_+) = 0.01$. Die Bayes'sche Formel Gl. 4.5 für die gesuchte Infektionswahrscheinlichkeit ergibt

$$\begin{aligned} P(K_+|T_+) &= \frac{P(T_+|K_+)\,P(K_+)}{P(T_+)} = \frac{P(T_+|K_+)\,P(K_+)}{P(T_+|K_+)\,P(K_+) \;+\; P(T_+|K_-)\,P(K_-)} \\ &= \frac{0.95 * 0.01}{0.95 * 0.01 \;+\; 0.05 * 0.99} = 0.161 \end{aligned}$$

Sie ist vor dem Test 1 % und nach dem Test 16 % (nicht 95 %!). Setzt man Zahlen ein, wird deutlich, warum Massentests (wie für aktuelle Epidemien politisch immer gerne eingefordert) oft wenig Sinn ergeben: In einer Bevölkerung von einer Million Menschen sind 10 000 infiziert und 9 500 positiv getestet; von den 990 000 Gesunden werden aber 49 500 irrtümlich positiv getestet, was zum Verhältnis 9 500 zu 59 000 (=16 %) insgesamt Positiv-Getesteten führt.

Ist die Krankheit noch seltener, $P(K_+) = 0.0001$, aber auch der Test besser, $P(T_+|K_+) = P(T_-|K_-) = 0.99$, ist die Infektionswahrscheinlichkeit $P(K_+|T_+) = 0.0098$ nach positivem Test noch geringer.

Die praktische Abhilfe ist die Erhöhung der a-priori-Infektionshäufigkeit $P(K_+)$ durch Testen erst nach begründetem Verdacht.

4.2 Zufallsvariablen und Wahrscheinlichkeitsverteilungen

Ist das Resultat des Zufallsexperimentes zahlenwertig (und mit natürlicher Kleiner-größer-Ordnung), so lässt es sich in der Regel auf die Menge der reellen Zahlen $\mathbb{R}$ abbilden und als **Zufallsvariable** repräsentieren. Je nach Art der Ergebnismenge unterscheidet man die gelegentlich *kontinuierliche* und *diskrete* Zufallsvariable, s.a. Tabelle 4.1.

Die **kumulative Häufigkeitsverteilung** oder kurz **Verteilungsfunktion** erlaubt die Charakterisierung der Zufallsvariablen X. Die Verteilungs-

funktion $F_X(x)$ gibt die Auftretenswahrscheinlichkeit von Werten $\leq x$ an:

$$V_X(x) = P(X \leq x) \tag{4.10}$$

Im diskreten Fall stellt sich $V_X(x)$ als Summe von Eintrittswahrscheinlichkeiten p_i aller Elementarereignisse A_i dar, die wertmäßig $\leq x$ sind:

$$V_X(x) = \sum_{i:\ x_i \leq x} P(A_i) = \sum_{i:\ x_i \leq x} p_i \tag{4.11}$$

Für eine kontinuierliche Zufallsvariable wird die Summation zum Integral

$$V_X(x) = \int_{-\infty}^{x} p_X(x')dx' \tag{4.12}$$

über die Wahrscheinlichkeitsdichte $p_X(x)$, kurz **Dichte**:

$$p_X(x) = \lim_{b \to 0} P_X(x \leq X < x + b) \tag{4.13}$$

Wie die diskrete Auftretenswahrscheinlichkeit p_i ist die Dichte $p(x)$ nichtnegativ, hier aber unbeschränkt ($p_i \in [0, 1], p(x) \in [0, \infty[$). In beiden Fällen muss die Gesamtwahrscheinlichkeit immer 1 ergeben, gleichbedeutend mit $V_X(\infty) = 1$. Beispiele für Verteilungsfunktionen sind Abb. 4.5d, S. 67 und Abb. 5.14d, S. 130.

4.2.1 Mehrdimensionale Verallgemeinerung und Erwartungswert

Die Zufallsvariable kann zu einer mehrkomponentigen, m-dimensionalen Vektorgröße, einem **Zufallsvektor** $\mathbf{X}$, verallgemeinert werden. Die Verteilungsfunktion des Zufallsvektors $\mathbf{X}$

$$V_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_m,) \tag{4.14}$$

ist dann das d-dimensionale Integral über die gemeinsame Dichtefunktion der einzelnen Zufallsvariablen X_i:

$$V_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_m} p_{\mathbf{X}}(x_1, x_2, \ldots x_d)dx_1dx_2 \cdots dx_m \tag{4.15}$$

Die Dichte einer einzelnen Zufallsvariablen X_i berechnet sich als Integral über alle anderen Komponenten der gemeinsamen Dichte

$$p_{X_i}(x_i) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_{i-1}} \int_{-\infty}^{x_{i+1}} \cdots \int_{-\infty}^{x_m} p_{\mathbf{X}}(\mathbf{x}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_m \tag{4.16}$$

Das Konzept der bedingten Wahrscheinlichen (vgl. Gl. 4.4) lässt sich auch auf die bedingte Dichte übertragen. Hier ein Beispiel mit zweikomponentigem Zufallsvektor:

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{X}}(x_1, x_2) = p_{X_2|X_1}(x_2|x_1) p_{X_1}(x_1) = p_{X_1|X_2}(x_1|x_2) p_{X_2}(x_2) \tag{4.17}$$

Wenn alle einzelnen Zufallskomponenten statistisch unabhängig sind, vereinfacht sich die gemeinsame Dichtefunktion zum Produkt der Einzelkomponentendichten:

$$p_{\mathbf{X}}(\mathbf{x}) = p_(x_1, x_2, \ldots x_d) = \prod_{i=1}^{m} p_{X_i}(x_i) \tag{4.18}$$

Mit Hilfe von Dichten lässt sich der Begriff **Erwartungswerte** kompakt einführen. Der Erwartungswert $E\{\cdot\}$ eines Ausdrucks $f(\mathbf{X})$ ist definiert als Integral über den gesamten Wertebereich ($\mathbb{R}^m$)

$$E\{f(\mathbf{X})\} = \int p_{\mathbf{X}}(\mathbf{x})\, f(\mathbf{X})\, d\mathbf{x} \tag{4.19}$$

oder, im Falle diskreter Verteilungen, als Summe:

$$E\{f(X)\} = \sum_{\forall i} p_i\, f(x_i). \tag{4.20}$$

4.2.2 Deskriptive Statistik für metrische Variablen

Sowohl theoretische Wahrscheinlichkeitsverteilungen als auch empirische Datenstichproben möchte man kompakt charakterisieren und beschreiben. Hier sind eine Reihe von berechenbaren kompakten Größen von Bedeutung. Einige sind eng mit den **Momenten** einer Datenverteilung verknüpft.

Das bekannteste Lagemaß ist der **Mittelwert** $\bar{x}$, das erste Moment (engl.: *mean, average*, öfter auch als $\langle X \rangle$ notiert). Er ist definiert als der

Erwartungswert $\bar{x} = E\{x\}$ und markiert den Schwerpunkt einer Verteilung. Der Mittelwert von N Werten $x_1, \ldots, x_N$ ist

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j. \tag{4.21}$$

Der Mittelwert ist aber nicht die einzige Größe, die die Lage einer Verteilung beschreibt. Insbesondere wenn starke Randausläufer auftreten, kann die Beschreibung durch den **Median** (siehe 4.2.3) oder mit der **Mode** sinnvoller sein.

Die Mode ist für kontinuierliche Variablen der Ort der größten Dichte x_{mode}, andernfalls die Wertausprägung j mit der maximalen Häufigkeit. Eine Verteilung heißt **unimodal**, wenn ihre Dichte nur ein globales Maximum besitzt, **bimodal**, wenn sie zwei verschiedene lokale Maxima besitzt, andernfalls heißt sie **multimodal**.

Als nächstes gilt es, die Ausgedehntheit, Variabilität, Streuung oder die „Breite“ einer Verteilung zu beschreiben. Am gebräuchlichsten ist die **Varianz**, die dem zweiten Moment einer Verteilung entspricht:

$$\text{Var}\{X\} = E\{(X - E\{X\})^2\} = \int_{-\infty}^{\infty} (x - E\{X\})^2 p_X(x) dx = \sigma^2. \tag{4.22}$$

Deren Quadratwurzel nennt man die **Standardabweichung** (engl: *standard deviation* kurz **s.d.**) $\sigma = s.d. = \sqrt{Var\{X\}}$.

Die **empirische Standardabweichung** einer Datenreihe wird, statt mit dem wahren Mittelwert, mit Hilfe des empirischen, also geschätzten Mittelwertes $\bar{x}$ berechnet. Die Varianzschätzung fällt in diesem Fall etwas zu klein aus, denn die quadratischen Datenabstände zum empirischen Mittelwert sind im Mittel kleiner als zum wahren Mittelwert, was sich für kleine Stichprobengrößen N bemerkbar macht. Die Korrektur $N - 1$ statt N im Nenner macht die Schätzung erwartungstreu (*unbiased estimator*):

$$\text{Var}(x_1, x_2, \ldots, x_N) = \sigma^2 = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \bar{x})^2. \tag{4.23}$$

Durch Klammerexpansion lässt sich die häufige Darstellung

$$\text{Var}(x_1, x_2, \ldots, x_N) = \sigma^2 = \frac{1}{N-1} \left[\left(\sum_{j=1}^{N} x_j^2 \right) - N\bar{x}^2 \right] \approx \overline{x^2} - \bar{x}^2 \tag{4.24}$$

herleiten. Sie erlaubt σ und $\bar{x}$ in einem Datendurchlauf zu berechnen, birgt aber die Gefahr größerer Rundungsfehler.

Der **Standardfehler** (engl.: *standard error*), kurz **s.e.**, beschreibt die Unsicherheit einer empirischen Parameterschätzung, z.B. der $\bar{x}$-Schätzung. Er ist von der Unsicherheit jeder einzelnen Messung geprägt und wird kleiner, je mehr Messungen vorliegen:

$$s.e. = \frac{\sigma}{\sqrt{N}}. \tag{4.25}$$

Die Standardabweichung sollte nicht mit dem Standardfehler verwechselt werden. Abb. 4.2 illustriert den obigen Zusammenhang anhand der Wahrscheinlichkeitsverteilungen einer Mittelwertschätzung, die auf unterschiedlichen Stichprobengrößen N beruht.

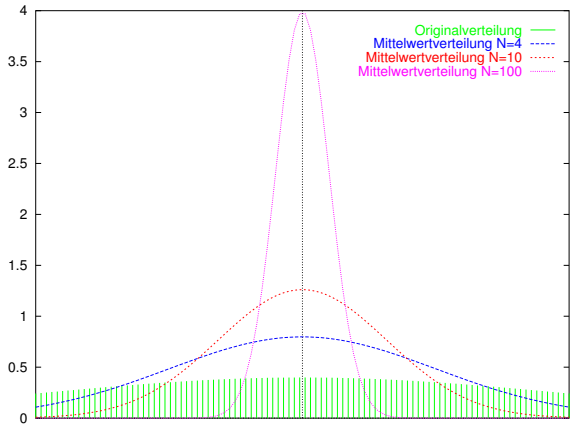


Abbildung 4.2: Vier Normalverteilungen mit gleichem Mittelwert und dem σ-Breitenverhältnis $10 : \sqrt{10} : 2 : 1$. Dieses Verhältnis entspricht der Reduktion des Standardfehlers (s.e.) durch eine Stichprobenvergrößerung $1 : 4 : 10 : 100$, siehe Gl. 4.25.

Im mehrdimensionalen Fall können sowohl die Varianzen jeder einzelnen Komponente als auch die kombinierten **Kovarianzen** kompakt in einer **Kovarianzmatrix** $\mathbf{C}$

$$\mathbf{C} = \mathrm{Var}\{\mathbf{X}\} = E\{(\mathbf{X} - E\{\mathbf{X}\})(\mathbf{X} - E\{\mathbf{X}\})^T\} \tag{4.26}$$

durch das äußere Vektorprodukt der zentrierten Verteilung dargestellt werden.

Zu bemerken ist, dass für einige Verteilungsfunktionen das zweite Moment nicht endlich ist, existiert – z.B. für die Cauchy-Verteilung – keine

Varianz. Ferner sind die höheren Momente sehr empfindlich gegen Ausreißer. Ein robusteres Breitenmaß ist die **mittlere Abweichung** (engl.: *average deviation, mean absolute deviation,* ***MAD***), die als

$$\text{ADev}(x_1, \ldots x_N) = \frac{1}{N} \sum_{j=1}^{N} |x_j - \bar{x}| \tag{4.27}$$

definiert ist. Die Betragsbildung macht diese Größe für analytische Betrachtungen allerdings unattraktiv.

Die **Schiefe** einer Verteilung wird durch das dritte Moment, englisch ***skewness***, berechnet:

$$\text{Skew}(x_1, \ldots x_N) = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{x_j - \bar{x}}{\sigma} \right)^3 \tag{4.28}$$

Ein positiver Wert markiert eine Asymmetrie mit dominant rechtsseitigen Ausläufern, auch „rechtsschief"´ genannt. Umgekehrt führen entlegene Ausläufer links vom Mittelwert zu negativen Werten. Die Normierung in Gl. 4.28 bewirkt für eine Normalverteilung eine Schiefe 0, die mit der Standardabweichung $\sqrt{6/N}$ normalverteilt ist.

Basierend auf dem vierten Moment wird mit der **Kurtosis** die „Spitzheit"´ einer Verteilung charakterisiert:

$$\text{Kurt}(x_1, \ldots x_N) = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left(\frac{x_j - \bar{x}}{\sigma} \right)^4 \right\} - 3 \tag{4.29}$$

Der Term -3 kommt durch das Referenzprofil zustande: Eine positive Kurtosis ist spitzer als die Normalverteilung (und hat damit längere Ausläufer), eine negative ist kompakter und randärmer. Die Kurtosis einer Normalverteilung ist selbst normalverteilt mit Mittelwert 0 mit der Standardabweichung $\sqrt{24/N}$.

4.2.3 Quantile, Median und Ordnungsstatistiken

Eine gegenüber Ausreißern wesentlich robustere Beschreibung basiert auf der Verteilungsfunktion. Sie steht nicht nur den metrischen, sondern auch den ordinal skalierten Variablen offen, denn es genügt hierbei die Ordnungsrelation $(<, =, >)$ zwischen Wertepaaren. Die Stichprobenwerte

werden geordnet $x_{[1]} \leq x_{[2]} \leq \cdots x_{[N]}$, indem man sie der Größe nach sortiert.

Eine α-**Quantile** Q_α ($\alpha \in [0,1]$) ist der x-Wert der geordneten Reihe $x_{[1]} \leq x_{[2]} \leq \cdots x_{[N]}$, unterhalb dessen gerade αN – und oberhalb gerade $(1-\alpha)N$ Datenwerte liegen. Der Zusammenhang zur Verteilungsfunktion $V_X(x)$ ist

$$V_X(Q_\alpha) = \alpha \,, \tag{4.30}$$

sofern der α-Quantilwert nicht mehrfach vorliegt und durch diese Bindung V_X (siehe Abs. 4.12.2) lokal angehoben ist.

Einige Schreibweisen für besondere Quantilen haben sich eingebürgert:

- Eine **Perzentile** ist ein Quantile, bei der α als Prozentwert angegeben ist.
- **Median** $Med = Q_{0.5} = Q_{1/2} = Q_{50\%}$. Er teilt die geordnete Reihe in zwei gleich große Teile, d.h. unterhalb und oberhalb liegen genau 50% der Beobachtungen. Ist N ungerade, gibt es einen eindeutigen Median-Datenwert $x_{[(N-1)/2]}$, andernfalls ist für metrische Größen die Mittelwertsbildung $(x_{[N/2-1]} + x_{[N/2+1]})/2$ vereinbart.
- Die erste **Quartile** $Q_{0.25} = Q_{1/4} = Q_{25\%}$ und dritte Quartile $Q_{0.75} = Q_{3/4} = Q_{75\%}$ markieren je ein Viertel. Gemeinsam mit dem Median teilen sie die Spannweite in vier gleichstark besetzte Bereiche.
- Minimum = Q_0 und Maximum = Q_1. Daraus ergibt sich auch die **Spannweite** der Verteilung $R = Q_1 - Q_0$.

Aus den 25% und 75% Quartilen werden der **Interquartilenabstand** (*interquartile range* IQR) oder auch die **F-Spanne** abgeleitet

$$s_F = Q_{75\,\%} - Q_{25\,\%} \,. \tag{4.31}$$

Diese enthält 50 % der Daten und wird in den Boxplots besonders hervorgehoben. Der Boxplot präsentiert, wie in Abb. 3.3, S. 34 zu sehen, in kondensierter Form die wichtigsten Verteilungsparameter: Min und Max durch die Extremwerte, den Median durch den Mittelstrich und die F-Spanne durch den Kasten. Beim Zeichnen des „T“-förmigen Ausläufers, dem *wisker*, wird es uneinheitlich: Die einfachste Art ist die Markierung der Extremwerte $Q_{0\%}$, $Q_{100\%}$. Der informationsreichere Weg nach Tuckey (1977)

begrenzt die Wisker-Länge auf den 1.5 fachen Interquartilenabstand s_F. D.h. der Wisker wird bis zum extremsten Datenpunkt in dieser Spanne gezeichnet, noch extremere werden durch Linien oder Punkte als Ausreißer gezeichnet. Manche differenzieren noch weiter zwischen milden und schweren Ausreißern.

4.3 Die Gauß'sche Normalverteilung

Die Glockenkurve nach Friedrich Gauß ist nicht nur wegen ihrer Abbildung auf der letzten Generation der 10DM-Geldscheine berühmt. Ihr anderer Name, **Dichtefunktion der Normalverteilung**, weist auf die fundamentale Bedeutung hin. Bei großen Stichproben verhält sich gemäß dem **zentralen Grenzwertsatz** der Mittelwert einer kontinuierlichen Zufallsvariable wie die Normalverteilung. Dabei spielen Detailzusammenhänge des Zufallsprozesses keine Rolle.

Die Normalverteilung ist mit genau zwei Parametern vollständig beschrieben, dem Zentrum (und Maximum) μ und der Varianz σ^2:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{4.32}$$

In Abb. 4.2 sind mehrere mittelwertgleiche Normalverteilungen gezeigt. Der Wendepunkte der Glockenkurven liegen jeweils genau bei $\mu \pm \sigma$. Das Maximum der breiten Verteilungen (mit größerem σ) ist kleiner, da die Gesamtfläche unter den Kurven als Gesamtwahrscheinlichkeit natürlich 1 ergeben muss.

Die Standardabweichung hat noch eine weitere Bedeutung. Bei gegebener Normalverteilung liegt im markierten σ-Intervall $[\mu-\sigma, \mu+\sigma]$ stets eine konstante, wohlbekannte Fläche, die 68.27% der Beobachtungen enthält (siehe auch Abb. 4.3). Dies führt uns im nächsten Abschnitt zu den Konfidenzintervallen.

Zunächst wenden wir uns der Verallgemeinerung der Normalverteilung für mehrdimensionale Zufallsvektoren in Matrixschreibweise zu:

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{C}) = \frac{1}{\sqrt{|2\pi\mathbf{C}|}}\ \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{C^{-1}}(\mathbf{x}-\mu)\right]. \tag{4.33}$$

Dabei bezeichnet μ den Mittelpunktvektor, der das Zentrum definiert, und die Kovarianzmatrix $\mathbf{C}$ beschreibt die *unimodale, elliptische* Ausdehnung

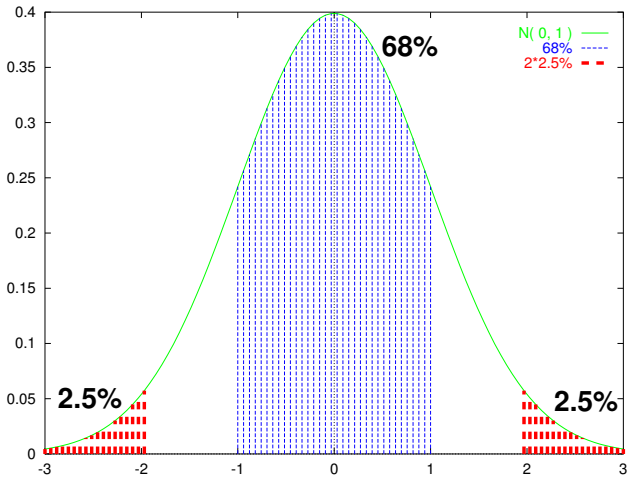


Abbildung 4.3: Die Standardnormalverteilung $\mathcal{N}$ (x)=$\mathcal{N}$ (x|0,1) hat eine Gesamtfläche von 1. Das Flächenintegral im Intervall [-1,1] beträgt 0.68, im Intervall [-1.96,1.96] 0.95 und 0.99 im Intervall [-2.57,2.57] ($Q^{\mathcal{N}}_{1-0.05/2} = 1.96$, $Q^{\mathcal{N}}_{1-0.01/2} = 2.57$, vgl. Fig. 4.5). Die Randschraffur deutet an, dass eine standardnormalverteilte Variable mit einer Wahrscheinlichkeit von 2.5 % einen Wert im Flügelbereich $x \geq 1.96$ annimmt.

der Verteilung. $\mathbf{C}^{-1}$ bezeichnet die Inverse von $\mathbf{C}$ und $|2\pi\mathbf{C}|$ die Determinate der mit 2π skalierten Kovarianzmatrix.

Zur Darstellung von realen Datenverteilungen genügt nicht immer eine unimodale, ellipsenförmige Dichteannahme. Häufig verwendet man den **Mischungsmodellansatz**

$$p(\mathbf{x}) = \sum_i p_i \mathcal{N}(\mathbf{x}|\mu_\mathbf{i}, \mathbf{C_i}), \tag{4.34}$$

eine p_i-gewichtete Linearkombination mehrerer Gaußverteilungen im hochdimensionalen Raum.

Der Begriff **Standardnormalverteilung** bezeichnet die Normalverteilung $\mathcal{N}(z) = \mathcal{N}(z|0,1)$ mit Erwartungswert $\mu = 0$ und Einheitsvarianz $\sigma^2 = 1$. Durch die **z**-Transformation $x \mapsto z = (x-\mu)/\sigma$ lässt sich jede Normalverteilung einfach auf die Standardnormalverteilung abbilden:

$$\mathcal{N}(z) = \mathcal{N}(\ z\ |0,1) = \mathcal{N}\left(\ \frac{x-\mu}{\sigma}\ |0,1\right) = \mathcal{N}(\ x\ |\mu,\sigma^2) \tag{4.35}$$

4.4 Konfidenzintervalle und Signifikanz

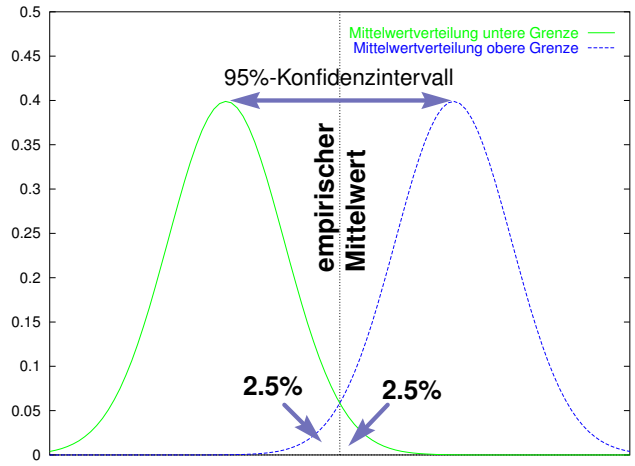


Abbildung 4.4: Das Konfidenzintervall der Mittelwertschätzung einer normalverteilten Variable mit angedeuteter Standardfehlerverteilung. Der empirische Mittelwert liegt hier in der Mitte. Die Grenzen UG und OG bestimmen sich symmetrisch durch die Überlegung einer statistischen Zufallsmessung. Die beiden Kurven deuten die beiden Extrempositionen der mutmaßlich wahren Verteilung an, die mit einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ noch verträglich sind.

Im vorigen Abschnitt wurde eine Reihe von deskriptiven Kennzahlen eingeführt. Sie beschreibt z.B. ein datenemittierendes System oder eine Population. Kennt man die wahren Kennzahlen nicht, schätzt man sie anhand von Stichproben. Nun stellt sich die Frage, mit welcher *Sicherheit* man einen Parameter bestimmen kann und wieviel *Vertrauen (confidence*) in diese Schätzung gelegt werden soll. Eine quantitative Erfassung ergibt sich aus dem Konzept des Konfidenzintervalls und der Irrtumswahrscheinlichkeit, die sich anhand zweier typischer Fragestellungen erläutern lassen:

Ist μ der wahre Wert eines Parameters (z.B. der Mittelwert einer Normalverteilung) einer Grundgesamtheit, wie wahrscheinlich ist dann die Richtigkeit der Aussage

$$UG \leq \mu \leq OG \tag{4.36}$$

mit gegebene Untergrenze UG und Obergrenze OG? Oder anders herum gefragt: Wie sind die Grenzen UG und OG beschaffen, damit die obige Aussage mit einer vorgegebenen Wahrscheinlichkeit $\Theta = 1 - \alpha$ richtig ist? Intervalle $[UG, OG]$, die die Bedingung

$$P(UG \leq \mu \leq OG) = \Theta = 1 - \alpha \tag{4.37}$$

erfüllen, nennt man **Konfidenzintervalle** (engl: *confidence interval*), kurz **CI**, zum sogennannten **Konfidenzkoeffizienten** Θ und dem **Signifikanzniveau** α.

Die Angabe eines Konfidenzintervalls bei der Schätzung des (wahren) Parameters ist kompakt aussagekräftig. Ist es sehr schmal, ist die μ-Schätzung präzise und umgekehrt wird eine ungenaue Schätzung an der großen Breite des Konfidenzintervalls deutlich. Sowohl eine eventuelle systematische Abweichung, auch **bias** genannt, als auch zufällige Abweichungen sind in $[UG, OG]$ enthalten und mit einer Sicherheitswahrscheinlichkeit Θ des Schätzers versehen.

Häufig wählt man die Sicherheitswahrscheinlichkeit Θ =95 % und damit das Signifikanzniveau α =5 %. Dies besagt, mit 95 % statistischer Wahrscheinlichkeit ist die gegebene Aussage richtig. Das Signifikanzniveau 5 % ist keine fundamentale Größe – es kann prinzipiell frei gewählt werden, aber es besitzt in vielen Wissenschaftsbereichen den Charakter einer *Konvention*. D.h., man erkennt eine Aussage als „**signifikant**“ an, wenn ihre statistische *Irrtumswahrscheinlichkeit* α =5 % nicht überschreitet („*very significant*“ bezeichnet per Konvention eine Aussage zum Signifikanzniveau $\alpha = 1\,\%$).

Liegt ein Ergebnis in Parameterform vor, ist die Angabe eines Konfidenzintervalls, z.B. das 95%-CI, die beste Weise, die Aussagekraft mit einer Genauigkeitsangabe zu flankieren.

Das Prinzip der Signifikanzbeurteilung mittels Irrtumswahrscheinlichkeit kann noch allgemeiner verwendet werden, um Aussagen mit einem kompakten Indikator zu qualifizieren – auch Aussagen, die keine CI-Angabe erlauben. Dies soll im Folgenden erläutert werden.

4.5 Nullhypothesen und p-Wert

Um praktische Fragestellungen einer statistischen Beurteilung zu unterziehen, wird zunächst eine Hypothese H_A formuliert, die man zu bestätigen sucht. Eine fundamentale Einsicht aus der Erkenntnistheorie lautet, dass man Aussagen meist nicht verifizieren, sondern nur falsifizieren kann (Karl Popper). Dies wird durch die Formulierung einer Gegenthese, der so genannten **Nullhypothese** H_0 realisiert. Die Nullhypothese sagt, dass die Aussage (Alternative) nicht gelte und das empirische Resultat ein zu-

fälliges, stochastisches Ergebnis sei. Dann sucht man Evidenz *gegen* die Nullhypothese, um die Alternative zu bestätigen.

Der Weg geht über den so genannten **p-Wert**: Man berechnet die statistische Wahrscheinlichkeit, dass – bei zutreffender Nullhypothese H_0 – ein Ergebnis *mindestens so extrem* wie das vorliegende eintritt. Ist dies hinreichend unwahrscheinlich, wird die Nullhypothese abgelehnt und die Alternativhypothese als *signifikant* bezeichnet. Die Testschwelle ist natürlich das Signifikanzniveau α. Ist $p > \alpha$, wird die Aussage als *nicht signifikant* behandelt, da nicht sicher genug ist, dass die Stichprobe ein stochastisches Ereignis bei Vorliegen von H_0 war.

4.6 Hypothesentest und Fehlerarten

Der grundsätzliche Ablauf beim statistischen Testen enthält

- das Formulieren der Hypothesen,
- die Auswahl des statistischen Testverfahrens,
- das Festsetzen der Irrtumswahrscheinlichskeitsgrenzen und des Stichprobenumfanges,
- das Ausführen des Tests und die Entscheidung.

Aufgrund der begrenzten Datenmenge kann man sich bei der Entscheidung über Annahme oder Ablehnung der statistischen Hypothese durchaus irren. Dabei sind zwei grundsätzliche Arten von Fehlentscheidungen möglich:

Fehler 1. Art ist die unberechtigte Ablehnung der Nullhypothese. Die Nullhypothese ist wahr (kein Effekt also), aber der Test zeigt die Ablehnung von H_0 an. Diese Irrtumswahrscheinlichkeit (*Risiko 1. Art*) ist

$$P(\text{Fehler 1. Art}) = \alpha.$$

Fehler 2. Art ist die unberechtigte Annahme der Nullhypothese. Die Alternativhypothese H_A gilt, wird aber irrtümlich verworfen. Diese Irrtumswahrscheinlichkeit des Fehlers 2. Art wird notiert als

$$P(\text{Fehler 2. Art}) = \beta.$$

Im Umfeld von Waren- oder Wirksamkeitstests, z.B. von Medikamenten, spricht man beim Risiko 2. Art auch vom **Produzentenrisiko** als dem Risiko der irrtümlichen Verwerfung eines wirksamen Stoffes (einer Therapie, eines Verfahrens etc.). Das **Konsumentenrisiko** beschreibt umgekehrt den Fehler 1. Art, nämlich die Verfahrensakzeptanz und den Kauf (oder die Anwendung) trotz eigentlicher Unwirksamkeit.

Test $\setminus$ **Realität**	H_0 **wahr** (H_A **falsch**)	H_A **wahr** (H_0 **falsch**)
H_A **angenommen**	P(Fehler 1. Art)	(richtige Entscheidung)
H_0 abgelehnt	α	$1-\beta$
H_0 **angenommen**	(richtige Entscheidung)	P(Fehler 2. Art)
H_A abgelehnt	$1-\alpha$	β

Tabelle 4.1: Mögliche Entscheidungen beim statistischen Hypothesentest

Die vier sich ergebenden Möglichkeiten werden in Tab. 4.1 sichtbar. Natürlich möchte man beide Fehlerarten klein halten, was (theoretisch) durch eine geeignet große Stichprobe erreichbar ist. Verändert man das Testverfahren, verhalten sich die beiden Fehlerarten antagonistisch: Eine Verkleinerung von α hat eine Vergrößerung von β zur Folge (und umgekehrt).

Als die **Schärfe** oder ***power*** $\pi = 1-\beta$ eines Tests wird die Wahrscheinlichkeit bezeichnet, mit der ein Test einen tatsächlich vorhandenen Wirksamkeitsunterschied zu identifizieren vermag (also H_0 ablehnt). Die Schärfe wird insbesondere durch das Testverfahren, die Stichprobenvarianz und die Stichprobengröße beeinflusst. Insbesondere, wenn die Beschaffung der Stichproben mit erheblichem Aufwand verbunden (d.h. teuer) ist, wird meist das gewünschte Signifikanzniveau und die Schärfe vorgegeben und die nötige Stichprobengröße in der Versuchsplanung minimiert (*experimental design*).

4.7 Ausgewählte statistische Tests

Statistische Tests werden je nach Grundannahmen in zwei Hauptklassen eingeteilt:

- parametrische Tests unterstellen eine Datenverteilung, z.B. die Normalverteilung eines Parameters μ;

- nichtparametrische Tests verzichten auf eine solche Annahme und sind daher allgemeiner einsetzbar. Dies kann allerdings mit einer verminderten Schärfe einhergehen.

Der einfachste parametrische Fall ist der Test auf Mittelwerte und Differenzen von Mittelwerten unter der Annahme der Normalverteilung.

4.7.1 Mittelwert einer Stichprobe

Wie oben beschrieben, sorgt der zentrale Grenzwertsatz generell dafür, dass sich bei sehr großen Stichproben die Mittelwertschätzung μ (eines Parameters) wie eine Normalverteilung verhält. Die Breite der Verteilung ist der bereits eingeführte Standardfehler $\frac{\sigma}{\sqrt{N}}$, der durch die Standardabweichung von μ und dem Umfang der Stichprobe N bestimmt ist (siehe Gl. 4.25). Das Konfidenzintervall für den tatsächlichen Mittelwert μ ist symmetrisch um das algebraische Mittel $\bar{x}$ verteilt:

$$\bar{x} - Q^{\mathcal{N}}_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \;\leq\; \mu \;\leq\; \bar{x} + Q^{\mathcal{N}}_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}. \tag{4.38}$$

Dabei ist $Q^{\mathcal{N}}_{1-\alpha/2}$ die Quantile der Standardnormalverteilung, die dem gegebenen Signifikanzniveau α entspricht. σ ist die Standardabweichung der Verteilung – sofern bekannt –, andernfalls wird die empirische Standardabweichung s verwendet.

4.7.2 Kleine Stichproben und die emphStudent-t-Verteilung

Sind die Stichproben nicht sehr groß, gibt es zusätzliche Ungenauigkeiten. Zum einen ist die empirische Standardabweichung s ungenau, zum anderen gilt der zentrale Grenzwertsatz nicht und die Verteilung der Mittelwerte kann nicht-normal werden. Das Problem wurde von William S. Gossett (1876-1937) durch die Einführung der t-Verteilung gelöst. Als Angestellter der Brauerei Guiness in Dublin konnte er seine Entdeckung nur unter Pseudonym veröffentlichen. Er wählte „Student“ und noch heute wird sein Test als **Student's-t-Test** bezeichnet.

Für große Stichproben sind σ und s austauschbar, für kleine nicht. Wo sich die Größe $(\bar{x}-\mu)/(\sigma/\sqrt{N})$ standardnormalverteilt verhält, verhält sich

die Größe $(\bar{x}-\mu)/(s/\sqrt{N})$ einer annähernd normalverteilten Zufallsvariable x nach der **Student-t-Verteilung** mit $d.f.=(N-1)$ **Freiheitsgraden** (*degrees of freedom*) und Gl. 4.38 wird zu:

$$\bar{x}-Q^{\mathcal{T}_{N-1}}_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}\leq\mu\leq\bar{x}+Q^{\mathcal{T}_{N-1}}_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}. \tag{4.39}$$

Ebenso wie die glockenförmige Standardnormalverteiltung $\mathcal{N}$ ist die t-Verteilung symmetrisch mit Mittelwert Null – aber mit etwas stärkerer Randverteilung. Mit zunehmenden Freiheitsgraden ν nähert sie sich $\mathcal{N}$ an. Die t-Verteilungsfunktion $V_t^{\mathcal{T}_\nu}$ ist darstellbar als

$$V_t^{\mathcal{T}_\nu}=\left[\nu^{\frac{1}{2}}\int_0^1\frac{(1-x)^{\frac{\nu}{2}-1}}{x^{\frac{1}{2}}}dx\right]^{-1}\int_{-\infty}^{t}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}dx \tag{4.40}$$

(Press et al. 1988; Bulirsch 1965) und ist für kleine, ganzzahlige ν häufig tabelliert, z.B. Zar (1996). Abb. 4.5 zeigt $V_t^{\mathcal{T}_\nu}$ für verschiedene ν im Vergleich zu $V_z^{\mathcal{N}}$.

Das Konfidenzintervall einer Parameterschätzung ist sehr eng mit dem Signifikanztest für Veränderung des Parameters verknüpft. Zum Beispiel möchte man anhand einer Stichprobe die Maßhaltigkeit bei einem Produktparameter (z.B. μ) feststellen. Gemäß Gl. 4.40 berechnet man dann das Konfidenzintervall für die gewünschte Stichprobengröße. Liegt der Sollwert nicht im Intervall, muss man die Nullhypothese verwerfen und eine Abweichung annehmen.

Will man ein Konfidenzintervall für einen neuen Datenpunkt abschätzen, erweitern sich die Intervallgrenzen der Parameterschätzung

$$\alpha\text{-CI der Datenschätzung:}\qquad \bar{x}\pm Q^{\mathcal{T}_{N-1}}_{1-\alpha/2}\left(1+\sqrt{\frac{1}{N}}\right)\sigma\ , \tag{4.41}$$

indem zum Standardfehler der Parameterschätzung in Gl. 4.40 die Standardabweichung der Verteilung hinzugefügt wird.

4.7.3 Ein- und zweiseitige Fragestellungen

Die Betrachtung im vorigen Beispiel ist mittensymmetrisch und ungerichtet, d.h. es liegt keine Information vor, in welche Richtung sich die Produktion verändern könnte. Man spricht dann auch von **zweiseitiger Fragestellung** (***two-sided***).

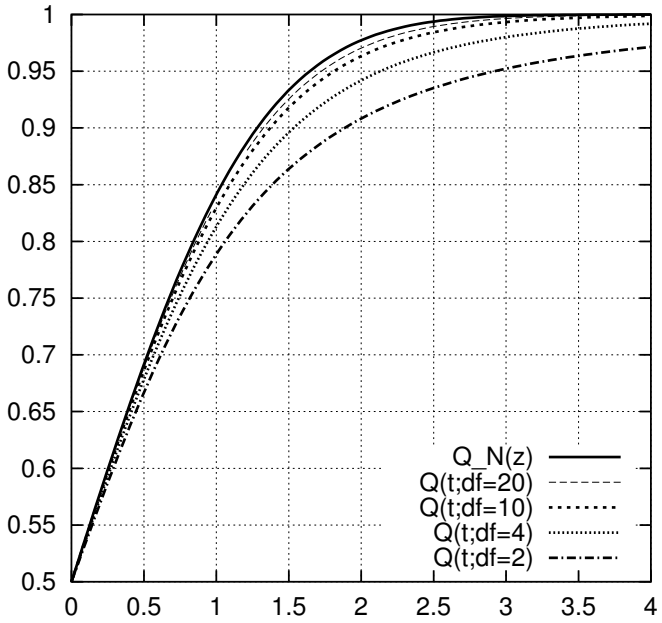


Abbildung 4.5: Verteilungsfunktionen der Student's-t-Statistik $V^{\mathcal{T}}(t;\,\nu)$. Wenige Beobachtungen (ν=df klein) können zu erheblichen Schwankungen der Mittelwerte führen. Für steigende Anzahl Freiheitsgrade ν nähert sich $V^{\mathcal{T}}$ rasch der Standardnormalverteilung $F^{\mathcal{N}}$ an. Negative t-Werte ergeben sich aus der Punktsymmetrie zu $(0,\frac{1}{2})$. Kritische Quantilenpunkte sind als Schnittpunkte zur gewünschten Horizontale ablesbar, z.B. vgl. $Q^{\mathcal{N}}_{0.975} \approx 1.96$ vs. $Q^{\mathcal{T}_{(\nu=4)}}_{0.975} \approx 2.78$ für $\alpha = 5\,\%$.

Kann man von Anfang an z.B. eine Vergrößerung (Verbesserung) des Parameters annehmen, formuliet man die Nullhypothese als „der Mittelwert ist nicht größer als μ". Diese **einseitige Fragestellung** (***one-sided***) bewirkt, dass nur eine Grenze berechnet wird und man das Irrtumsflächenintegral auf eine Seite schlägt, hier z.B. $OG = \bar{x} + Q^{\mathcal{T}_{N-1}}_{1-\alpha} \frac{\sigma}{\sqrt{N}}$ (beachte α nicht $\alpha/2$, vgl. Gl. 4.39).

4.7.4 Vergleich zweier (un-)abhängiger Stichproben anhand ihrer Mittelwerte: (*unpaired*/*paired*) t-Test

Liegen zwei unabhängige Stichproben mit den Größen N_1 und N_2 vor, möchte man beurteilen, ob sie von derselben Datenverteilung stammen

(können). Auch hier kann man dies nicht positiv beweisen, sondern testet wieder die Nullhypothese. Mittels dem **unabhängigem (*unpaired*) t-Test** kann die signifikante Verschiedenheit der beiden Mittelwerte beurteilt werden. Hierbei setzt man näherungsweise Varianzgleichheit und Normalverteilung voraus. Als Vergleichstatistik berechnet man die Differenz

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_D}\sqrt{\frac{N_1 N_2}{N_1 + N_2}} \quad \text{mit} \quad s_D^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \tag{4.42}$$

der beiden Mittelwerte $\bar{x}_1, \bar{x}_2$ relativ zu der *gemischten Varianz* (***pooled variance***) s_D^2, die aus dem gewichteten Mittel der empirischen Varianzen s_1^2, s_2^2 gebildet wird.

Das Konfidenzintervall ergibt sich wieder symmetrisch um die empirische Differenz, skaliert mit dem Standardfehler $s.e.$ und der Perzentile der t-Verteilung mit $d.f. = \nu = N_1 + N_2 - 2$ Freiheitsgraden:

$$CI_\alpha = (\bar{x}_1 - \bar{x}_2) \pm (Q_{1-\alpha/2}^{\mathcal{T}_{N_1+N_2-2}} s.e.) \quad \text{mit} \quad s.e. = s_D\sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \tag{4.43}$$

Stammen die beiden Stichproben ($N_1 = N_2 = N$) vom gleichen Tupel von Testern, spricht man von **abhängigen** oder **verbundenen Stichproben**. Zum Beispiel wird die Wirksamkeit zweier Medikamente an der selben Patientengruppe erprobt oder umgekehrt, zwei Kandidaten werden vom selben Prüfungskomitee bewertet. Allein die Unterschiedlichkeit der Patienten bzw. der Tester kann in solchen Fällen eine so hohe Stichprobenvarianz ergeben, dass die tatsächliche Differenz im Mittelwert bei der Beurteilung als zu klein erscheint und zu einer irrtümlichen Gleichwertung (also zum Fehler 2. Art) führt.

Liegt eine solche eindeutige Paarung der Stichprobenwerte x_{1i}, x_{2i} vor, sollte daher der **t-Test für abhängige Stichproben** (***paired t-test***) durchgeführt werden. Dazu prüft man die Testgröße

$$t = \frac{\bar{d}}{s_D}\sqrt{N} \quad \text{mit} \quad s_D^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2 \quad \text{und} \quad \bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i, \tag{4.44}$$

die auf den Paardifferenzen $d_i = x_{1i} - x_{2i}$ der beiden Stichproben $\{x_{1i}\}$ $\{x_{2i}\}$ beruht.

4.7.5 F-Test: Varianzgleichheit zweier Stichproben

Wie kann man mit einem statistischen Test überprüfen, ob die Voraussetzung der Gleichheit der Varianzen σ_1^2, σ_2^2 erfüllt ist? Man stellt die Nullhypothese $H_0 : \sigma_1^2 = \sigma_2^2$ auf und beurteilt sie mit dem so genannten **F-Test**. O.B.d.A. wird angenommen, dass $\sigma_1^2 > \sigma_2^2$ ist und der Quotient

$$F = \frac{\sigma_1^2}{\sigma_2^2} \tag{4.45}$$

als Testgröße gegen die α-Quantile der F-Verteilung $Q_{1-\alpha}^{\mathcal{F}_{N_1-1,N_2-1}}$ verglichen wird (andernfalls wird 1 und 2 vertauscht). Die kumulative F-Verteilung wird durch zwei Freiheitsgrade $\nu_1 = N_1 - 1$, $\nu_2 = N_2 - 1$ parametrisiert:

$$V^{\mathcal{F}_{\nu_1,\nu_2}}(F) = \left[\int_0^1 t^{\frac{\nu_2}{2}-1}(1-t)^{\frac{\nu_1}{2}-1}dt\right]^{-1} \int_0^{\frac{\nu_2}{\nu_2+\nu_1 F}} t^{\frac{\nu_2}{2}-1}(1-t)^{\frac{\nu_1}{2}-1}dt \ . \tag{4.46}$$

Die Wahrscheinlichkeit für Werte mit $F \gg 1$ nimmt rasch ab. D.h. ein großer F-Wert macht die H_0 also immer unwahrscheinlicher; wie schnell, hängt von den Stichprobengrößen ab.

4.8 Vergleich mehrerer Stichproben: *ANOVA*-Test

Möchte man K Stichproben eines Parameters vergleichen, ist der paarweise t-Test nicht empfehlenswert. Er ist nicht nur mühsam ($K(K-1)/2$ Vergleiche), sondern auch problematisch, da die Irrtumswahrscheinlichkeit 1. Art mit K steigt (eine obere Schranke ist $\alpha K(K-1)/2$). Um sicher zu gehen, kann man die Einzeltests mit einer Adjustierung $\alpha' = \alpha/K$ durchführen.

Besser jedoch ist eine integrierte Analyse der Mittelwerte aller Stichproben mittels der ANOVA-Methode:

Die Analyse der Varianz (engl: *ANalyis Of VAriance*, ANOVA) vergleicht (trotz des Namens) die *Mittelwerte* mehrerer Gruppen (Stichproben) unter Berücksichtigung der Variabilität innerhalb und zwischen den Gruppen.

Hier sei die einfaktorielle Varianzanalyse (*One-way ANOVA*) erläutert: Sei x_{ki} das i-te Datum aus der k-ten Gruppe mit $k \in \{1, \ldots K\}$ und $i \in \{1, \ldots, N_k\}$, so lassen sich Gruppenmittelwert $\bar{x}_{k\cdot} = \frac{1}{N_k}\sum_{i=1}^{N_k} x_{ki}$ der k-ten

Gruppe und der Gesamtmittelwert $\bar{x}_{..} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{N_k} x_{ki}$ mit Gesamtzahl $N = \sum_{k=1}^{K} N_k$ bestimmen (die $\cdot$ Punktnotation im Index markiert hier die Mittelung). Damit lässt sich die Gesamtvarianz, d.h. hier die Quadratsumme

$$QS_{gesamt} = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_{ki} - \bar{x}_{..})^2 = QS_{intra} + QS_{inter} \tag{4.47}$$

zerlegen in die Intragruppen- und Intergruppenkomponente

$$QS_{intra} = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_{ki} - \bar{x}_{k.})^2 \quad \text{und} \quad QS_{inter} = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(\bar{x}_{k.} - \bar{x}_{..})^2. \tag{4.48}$$

Die Nullhypothese H_0 geht von der Gleichheit der zugrunde liegenden Datenverteilungen, d.h. $\mu_1 = \mu_2 = \ldots = \mu_K$, aus und führt zur Testfrage: Ist die Intergruppenvariabilität vergleichbar zur Intragruppenvariabilität – oder größer? Aus dieser Überlegung wird die Testgröße

$$F = \frac{QS_{inter}/(K-1)}{QS_{intra}/(N-K)} \tag{4.49}$$

abgeleitet. Anhand der V_F-Verteilung mit $K-1$ und $N-K$ Freiheitsgraden kann man den p-Wert ablesen (Gl. 4.46). Die Nullhypothese H_0 wird dann abgelehnt, wenn die α-Quantile der F-Verteilung überschritten wird, also bei $F > Q_{\alpha}^{\mathcal{F}_{K-1,N-K}}$.

Mit der zwei- oder mehrfaktoriellen Varianzanalyse (*two-way ANOVA*) kann man Daten analysieren, die nach zwei oder mehr (kategorialen) Parametern klassifiziert sind, siehe z.B. Zar (1996).

4.9 χ^2-Verteilung und *Goodness-of-fit*-Test

Die χ^2-Statistik betrachtet die Quadratsumme $\chi^2 = \sum_{i=1}^{\nu} x_i^2$ von ν standardnormalverteilten Zufallsvariablen $x_i = \mathcal{N}(0,1)$. Die χ^2-Verteilungsfunktion

$$V^{\chi^2{}_\nu}(\chi^2) = \left[\int_0^{\infty} e^{-x}t^{\nu/2-1}dx\right]^{-1} \int_0^{\chi^2/2} e^{-x}t^{\nu/2-1}dx \tag{4.50}$$

ist in Abb. 4.6 dargestellt. Unter der Annahme dieser Ausgangssituation gibt sie die Wahrscheinlichkeit an, dass höchstens der beobachtete χ^2-Wert eintritt.

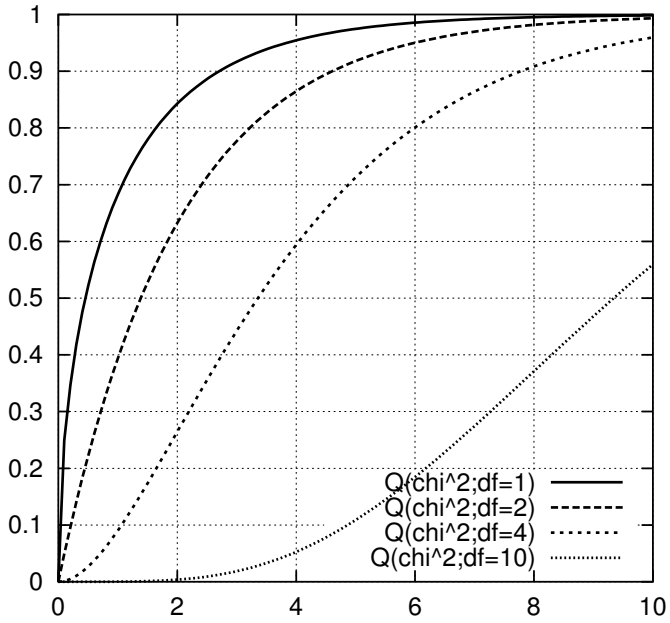


Abbildung 4.6: Die Verteilungsfunktion der χ^2-Statistik (Gl. 4.50). Mit steigender Zahl von Freiheitsgraden ν (=df) verschiebt sich die Verteilung zu größeren Werten, i.e. $E\{\chi^2\} = \nu$ und $var\{\chi^2\} = \sqrt{2\nu}$.

Die χ^2-Statistik wird u.a. zur Bewertung von Modellanpassungen eingesetzt (*goodness-of-fit*-Test). Hierbei wir die Quadratsumme

$$\chi^2 = \sum_{i=1}^{\nu} \left(\frac{x_i - \hat{x}_i}{\sigma_{(i)}} \right)^2 \tag{4.51}$$

der z-transformierten Residuen gebildet (d.h. die Differenzen zwischen beobachteten x_i und erwarteten Werten $\hat{x}_i$ geteilt durch die jeweilige Standardabweichung $\sigma_{(i)}$, vgl. Gl. 4.35). Die Nullhypothese besagt, dass die beobachteten Werte allein aufgrund zufälliger Schwankungen von der korrekten Erwartung abweichen. Der Zusammenhang zwischen Signifikanzen α oder p-Werten und einer empirischen χ^2-Messung stellt dann Gl. 4.50 her.

4.10 Kolmogorov-Smirnov Test für zwei Verteilungsfunktionen

Der Kolmogorov-Smirnov-Test macht – im Gegensatz zu den bisherigen Tests – keine Annahmen über Datenverteilungen, sondern vergleicht direkt zwei Verteilungsfunktionen: entweder *(i)* mit einer empirischen $V_N(x)$ und einer bekannten Verteilungsfunktion $V^{Dist}(x)$ oder *(ii)* mit zwei empirischen Verteilungsfunktionen $V^1_{N_1}(x)$, $V^2_{N_2}(x)$. Damit eignet er sich auch für ordinal-skalierte Daten. Der Test ist konzeptionell sehr einfach, man sucht den größtmöglichen vertikalen Abstand D zwischen den beiden kumulativen Verteilungskurven:

$$\textit{(i):} \quad D = \max_{-\infty<x<\infty} |V_N(x) - V^{Dist}(x)| \tag{4.52}$$

oder

$$\textit{(ii):} \quad D = \max_{-\infty<x<\infty} |V^1_{N_1}(x) - V^2_{N_2}(x)|. \tag{4.53}$$

Die Nullhypothese H_0 besagt, dass beide Verteilungen gleich sind, also D klein sein sollte. Je größer $D \in [0,1]$, desto unwahrscheinlicher ist H_0. Die $KS(\lambda)$-Statistik beschreibt den p-Wert in Abhängigkeit von D:

$$P(D' > D) = KS(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}, \tag{4.54}$$

wobei je nach Fall

$$\textit{(i):} \quad \lambda = \sqrt{\frac{N_1 N_2}{N_1+N_2}}D \quad \text{bzw. } \textit{(ii):} \quad \lambda = \sqrt{N}D \tag{4.55}$$

gilt.

Die bisher beschriebenen Tests sind für allgemeine, kontinuierliche Grundgesamtheiten geeignet. Nun sollten Tests im Vordergrund stehen, die Häufigkeitsverteilungen von dichotomen und kategorialen Daten vergleichen. Hierbei treten ganzzahlige Werte, absolute Häufigkeiten und Verhältnisse von solchen auf. Sie lassen sich bei großen Stichproben wieder durch die Normalverteilung approximieren – kleine Stichproben bedürfen aber einer besonderen Betrachtung. Für die Analyse einer dichtomen Variablen ist das Bernoulli-Experiment fundamental. Möchte man die Korreliertheit von kategorialen Variablen untersuchen, führt dies über die Analyse von Kontingenztabellen, insbesondere mit dem χ^2-Test.

4.11 Bernoulli-Experiment, Binomialverteilung und Zahlverhältnisse

Als **Bernoulli-Experiment** bezeichnet man ein Zufallsexperiment, das N-mal wiederholt wird und bei dem jeweils das Ereignis (A) mit der unveränderten Wahrscheinlichkeit π eintreten kann. Damit ist die Wahrscheinlichkeit für (¬A) konstant $1-\pi$ und alle Einzelergebnisse sind unabhängig. Die Wahrscheinlichkeit dafür, dass bei einem solchen Bernoulli-Experiment in m von N Realisationen das Ereignis (A) eintritt, ist durch die **Binomial-Verteilung** gegeben

$$p_{\pi,N}(m) = \left(\begin{array}{c} N \\ m \end{array} \right) \pi^m (1-\pi)^{N-m} = \frac{N!}{m!(N-m)!} \pi^m (1-\pi)^{N-m}. \quad (4.56)$$

Daran lassen sich die statistischen Schwankungen der relativen Häufigkeit p, also dem beobachteten **Zahlenverhältnis** $p = m/N$ (eng.: *proportion*, s.a. Gl. 4.2) in einer gegebenen Stichprobe mit N Objekten erkennen. Der Erwartungswert der relativen Häufigkeit p, die Varianz und der Standardfehler sind

$$E\{p\} = \pi \quad (4.57)$$

$$\sigma^2 = N\pi(1-\pi) \quad (4.58)$$

$$s.e.(p) = \sqrt{\pi(1-\pi)/N}\,. \quad (4.59)$$

Die Berechnung von Konfidenzintervallen wird schwierig, sobald mit $\pi \neq 0.5$ die Verteilung unsymmetrisch ist. In Abs. 9.5 wird eine numerische Monte-Carlo-Schätzmethode vorgestellt. Die p-Wert-Signifikanzbestimmung ist wiederum einfach und erfolgt durch Summieren der (im zweiseitigen Fall) doppelten Wahrscheinlichkeiten der mindestens so extremen m-Werte, d.h. $0 \ldots m$ für $m < \pi N$, ansonsten $m \ldots M$.

Für wachsendes N wird es einfacher, denn dann wird die Binomialverteilung sehr gut durch die Normalverteilung approximiert. Gute Ergebnisse sind erwartbar, wenn als Faustregel beide erwarteten Häufigkeiten von (A) und (¬A) mindestens 5 sind, i.e. $N\pi \geq 5 \wedge N(1-\pi) \geq 5$. Die Vergleichsstatistik für ein Verhältnis p gegen das einer bekannten Verteilung π ist die Standardnormalverteilung mit $z = (p-\pi)/s.e.(p)$. Da m aber nur ganzzahlig werden kann, ist die Verteilungsfunktion treppenförmig. Hier empfiehlt sich die **Kontinuitätskorrektur** nach Yates

$$z = \frac{|p-\pi| - \frac{1}{2N}}{\sqrt{\pi(1-\pi)\frac{1}{N}}}\,, \quad (4.60)$$

bevor z im nächsten Schritt mit der Standardnormalverteilung verglichen wird („Yates-Korrektur“).

Zwei empirische Verhältnisse p_1, p_2 lassen sich mit

$$z = \frac{|\pi_1 - \pi_2| - \frac{1}{2N_1} - \frac{1}{2N_2}}{\sqrt{p(1-p)(\frac{1}{N_1} + \frac{1}{N_2})}} \quad \text{mit} \quad p = \frac{m_1 + m_2}{N_1 + N_2} \tag{4.61}$$

auf gleiche relative Häufigkeit testen (d.h. $\pi_1 = \pi_2 = \pi$), indem die relative Gesamthäufigkeit p beider Stichproben eingesetzt wird.

4.12 Analyse von Kontingenztabellen und Assoziation zweier Variablen

Die Beurteilung von Assoziationen ist vom Datentyp (Skalierung) der Merkmale abhängig. Für kategoriale Variablen eignet sich der χ^2-Test.

4.12.1 Der χ^2-Test für Kontingenztabellen

Bei Vorliegen eines multivariaten, kategorialen Datensatzes tritt die Frage auf, ob es eine feststellbare Korreliertheit oder Assoziation von Variablenpaaren gibt und wie ist die Sicherheit oder Unsicherheit dieser Aussage zu bewerten ist.

	Normal	**Hyper-tonie**	$\sum$
m	**250**	**455**	705
w	**80**	**215**	295
$\sum$	330	670	1000

	X=1	**X=2**	**X=3**	**...**	
Y=1	N_{11}	N_{12}	N_{13}	...	$N_{1\cdot}$
Y=2	N_{21}	N_{22}	N_{23}	...	$N_{2\cdot}$
...	...	...	...	...	...
	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	...	$N_{\cdot\cdot}$

Tabelle 4.2: Beispiele einer Kontingenztabelle für zwei kategoriale Variablen: *(a, links)* konkretes Zahlenbeispiel mit der Merkmalskombination Geschlecht $\times$ Hypertonie und *(b, rechts)* Darstellung der Notation. Eingetragen werden die absoluten Häufigkeiten der Merkmalskombinationen, sowie die Marginalsummen, i.e. die jeweiligen Spaltensummen $N_{\cdot j}$, Zeilensummen $N_{i\cdot}$ und Gesamtzahl $N_{\cdot\cdot} = N$.

Zum Beispiel ist im Kontext einer herzchirurgischen Operationen von Interesse, ob bestimmte Erkrankungen bei Männern und Frauen gleich

häufig auftreten oder ob es signifikante Unterschiede gibt. Hierzu werden die Daten in eine so genannte **Kontingenztabelle** eingetragen. Wie in Tabelle 4.2 dargestellt, definieren die Ausprägungen j der einen Variable X die Spalten und die i der anderen Y die Zeilen. Alle Auftretenskombinationen werden gezählt und die Anzahl N_{ij} in die Zellen ij notiert. Am Rand werden die marginalen Häufigkeiten durch Summieren in den Spalten $N_{\cdot j}$ bzw. in den Zeilen $N_{i\cdot}$ ermittelt. Sie müssen sich beide zur Gesamtzahl der Fälle N addieren:

$$N_{\cdot j} = \sum_{\forall j} N_{ij} \ , \qquad N_{i\cdot} = \sum_{\forall i} N_{ij} \ , \qquad N = \sum_{\forall j} N_{i\cdot} = \sum_{\forall i} N_{\cdot j} \tag{4.62}$$

Die Nullhypothese H_0 besagt standardmäßig, dass die beiden Variablen x und y keine Assoziation haben. In diesem Falle sollten die Wahrscheinlichkeiten der x-Werte von den y-Werten unabhängig sein. Damit ergeben sich die **erwarteten Häufigkeiten** n_{ij} (in den Zellen ij) allein aus dem Produkt der beiden (empirischen, relativen) marginalen Häufigkeiten ($N_{i\cdot}/N$ und $N_{\cdot j}/N$), die aus den Randsummen der Kontingenztabelle bestimmt werden:

$$n_{ij} = N \, \frac{N_{i\cdot}}{N} \, \frac{N_{\cdot j}}{N} = \frac{N_{i\cdot} N_{\cdot j}}{N}. \tag{4.63}$$

Die erwarteten Häufigkeiten n_{ij} werden mit den beobachteten Häufigkeiten N_{ij} verglichen und zusammengefasst durch

$$\chi^2 = \sum_{\forall i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}}. \tag{4.64}$$

Ein Wert $\chi^2 \approx 0$ ist perfekt, kommt aber wegen Zufallsschwankung und der (i.A.) Nichtganzzahligkeit der n_{ij} kaum vor. Wieviel Abweichung erwartet werden kann, wird durch einen Vergleichstest mit der χ^2**-Statistik** ermittelt. Übersteigt $V^{\chi^2{}_\nu}(\chi^2) = 1 - p > \theta = 1 - \alpha$, ist die Assoziation als signifikant zu bezeichnen (Gl. 4.50 und Abb. 4.6, S. 71).

Die Zahl der sogenannten Freiheitsgrade ν (engl. *degrees of freedom, d.f.*) ist hier nicht gleich der Anzahl der Zellen, da die Terme über die Marginalsummen gekoppelt sind (es sei denn, die n_{ij} wurden anderweitig bestimmt). Die korrekte Zahl der Freiheitsgrade ist kleiner, i.e. $\nu = (I-1)(J-1)$, wobei I, J die Anzahl von Spalten bzw. Zeilen der Kontingenztabelle bezeichnet. Merkmalsausprägungen, die in der Datensatzauswahl nicht vorkommen (d.h. $N_{\cdot j} = 0$ oder $N_{i\cdot} = 0$), müssen vor der Berechnung aus der Tabelle durch Zeilen- oder Spaltenstreichung entfernt werden.

Die Einzelterme von Gl. 4.64 sind nicht unbedingt standardnormalverteilt. Die Näherung wird gut, wenn es genügend viele Terme gibt oder alle Zellen zahlreich besetzt sind. Eine präzisere Schätzung wurde von Cochran (1954) generell empfohlen, wenn die Gesamtzahl mit $N < 20$ sehr gering oder $20 < N < 40$ und $n_{ij} < 5$ ist. Für einen Vergleich zweier dichotomer Variablen, der eine 2×2-Tabelle ergibt, empfiehlt sich dann der exakte **Fischer-Test**. In diesen Zahlenbereichen kann die Wahrscheinlichkeit einer Tabellenkonstellation exakt und mit vertretbarem Aufwand berechnet werden:

$$P = \frac{e!\ f!\ g!\ h!}{n!\ a!\ b!\ c!\ d!} \tag{4.65}$$

	X=0	**X=1**	
Y=0	a	b	e
Y=1	c	d	f
	g	h	n

$$\chi^2 = \frac{n(ad-bc)^2}{efgh}, \qquad \nu = 1 \quad \textit{d.f.}$$

Tabelle 4.3: Verallgemeinerte Notation für eine 2×2 Kontingenztabelle für zwei dichotome Variablen. *(Rechts:)* Dabei vereinfacht sich Gl. 4.64.

Durch simultanes Inkrementieren von a, c und Dekrementieren von b, d (und umgekehrt) können alle möglichen (a, b, c, d)-Konstellationen mit gegebener gewünschter Randverteilung (e, f, g, h, n) erzeugt werden. Addiert man die Wahrscheinlichkeiten dieser und noch extremerer Konstellationen, ergibt sich das Signifikanzniveau der Alternativhypothese.

Tabellenassoziationsmaße

Der χ^2-Wert ist abhängig von der Stichprobengröße N, was ihn als direkten Assoziationsindikator ungeeignet macht. Gewünscht wird ein komfortabler Gesamtindikator im Intervall [0,1], der das Ausmaß der Assoziation direkt beschreibt. Von vielen Möglichkeiten sind zwei Varianten verbreitet:

Cramers V

$$V = \sqrt{\frac{\chi^2}{N \min(I-1, J-1)}} \in [0, 1], \tag{4.66}$$

das für $I = J = 2$ auch als **phi**-Statistik bezeichnet wird, und der **Kontingenzkoeffizient C**

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \in [0, 1[, \tag{4.67}$$

dessen oberes Limit leider von I und J abhängt. Beide Indikatoren lassen Wünsche offen, da mittlere Werte keine nützliche Interpretation zulassen, die über einen Größer-kleiner-Vergleich weit hinausgeht.

Begriffe: *relative risk* und *odds ratio*

Insbesondere im medizinischen Umfeld werden vielfach die Begriffe *relative risk* (RR) und *odds ratio* (OR) verwandt. Das **relative Risiko** RR, zum Beispiel einer Krankheit in Abhängigheit zu einem Merkmal (Eigenschaft, Substanzexposition oder Therapie etc.), ist der Quotient der Krankheitshäufigkeit unter den Merkmalsträgern relativ zu den Nichtmerkmalsträgern. Mit der Notation von Tab. 4.3 ergibt sich das

$$\text{relative Risiko} \quad RR = \frac{P(Y=1|X=1)}{P(Y=1|X=0)} = \frac{d/h}{c/g} = \frac{d\,g}{c\,h}\,, \tag{4.68}$$

wenn X hier das Merkmal und Y die Krankheit codiert.

Ähnlich ist das **Quotenverhältnis** (engl. ***odds ratio***, OR) definiert: Statt der Häufigkeit werden die beiden Verhältnisse krank-gesund (Quoten) in Relation gesetzt, also

$$\textit{odds ratio} \quad OR = \frac{\frac{P(Y=1|X=1)}{P(Y=0|X=1)}}{\frac{P(Y=1|X=0)}{P(Y=0|X=0)}} = \frac{d/b}{c/a} = \frac{a\,d}{b\,c}\,. \tag{4.69}$$

Der Vorteil des Quotenverhältnisses liegt in der Symmetrie der Spalten und Zeilen, wodurch etwaige Mehrdeutigkeiten der Bezugsgröße (Randsummen von X oder Y?) entfallen – Mehrdeutigkeiten, die bei RR entstehen können. I.A. gilt $0 < RR < OR$, für seltene Ereignisse (kleine Risiken mit $a/c << 1$) nähern $OR \approx RR$ sich an.

Die Berechnung von Konfidenzintervallen ist schwierig. Eine Näherung wurde von Miettinen und Nurminen (1985) als Basis des χ-Wertes vorgeschlagen: die 95%-CI-Schranken werden zu $OR^{1\pm.96/\chi}$ (bzw. $RR^{1\pm.96/\chi}$) abgeschätzt. Auf einer logarithmischen Skala ist die Verbreiterung symmetrisch und invers zu χ (nicht χ^2). Andere α-Niveaus ergeben sich durch

Ersetzen von 1.96 durch die entsprechende Standardnormalverteilungsquantile $Q^{\mathcal{N}}_{1-\alpha/2}$.

Tabelle 4.4 komplettiert das 2×2-Kontingenztabellenbeispiel aus Tab. 4.2a mit den daraus ableitbaren Beschreibungsgrößen.

	Normal	**Hypertonie**	
M	250 (232.7)	455 (472.4)	705
W	80 (97.4)	215 (197.6)	295
	330	670	1000

$$\chi^2 = 6.546 \Rightarrow\ p = 0.011$$

Tabelle 4.4: Beispiel Kontingenztabelle Tab. 4.2a Geschlecht $\times$ Hypertonie mit Angabe der erwartbaren Häufigkeiten (n_{ij}) in Klammer. Daraus ergibt sich $\chi^2 = 6.546$. Mit $\nu = 1$ entspricht dies einem $p = 0.011$, d.h. die Abweichung wird als signifikant betrachtet. Das Quotenverhältnis OR=1.477 (*odds ratio*) mit dem Konfidenzintervall CI-95% [1.094, 1.992] und dem relative Risiko RR=1.128. Cramers-V ist 0.081, und C=0.081.

Verbundene Daten (paired data): Ganz analog zum t-Test muss beachtet werden, dass eine Statistikanpassung erfolgen muss, wenn eine Datenpaarung vorliegt. Zum Beispiel: Das Vorliegen von Beschwerden wird für alle Untersuchungsteilnehmer zu zwei Zeitpunkten, vor und nach einer bestimmten Behandlung untersucht. Die Ergebnisanalyse mit dem zu verwendenden **McNemar-χ^2-Test** fokussiert sich auf die Zahl der **diskordanten Paare**, d.h. die Datenpaare, die eine Beschwerdenänderung verzeichne. Sei hier k die Zahl der Verbesserungen und l die Zahl der Verschlechterungen, dann berechnet sich der angepasste McNemar-χ^2-Test mit $\nu = 1$

$$\chi^2_{paired} = \frac{(k-l)^2}{k+l}, \qquad OR_{paired} = \frac{l}{k}. \tag{4.70}$$

McNemar's-Test gilt als valide, wenn die Gesamtzahl der diskordanten Paare $k + l \geq 10$ ist.

Was ist zu tun, wenn die Beziehung zwischen drei oder mehr Variablen untersucht werden soll? Eine Möglichkeit ist 2×2-Kontingenztabellen χ^2-Tests in disjunkten Merkmalsuntergruppen zu berechnen und mit der **Mantel-Haenszel**-Testvariante zusammenzuführen. Die meist bessere Alternative ist die multivariate (logistische) Regression, sie wird in Abs. 5.7.2 erläutert.

4.12.2 Nichtparametrische Tests

Nichtparametrisch oder parameterfrei heißen alle statistischen Tests, die nicht an die Voraussetzung einer bestimmten Verteilung mit entsprechenden Parametern gebunden sind. Sie sind also immer dann von Bedeutung, wenn z.B. die Normalverteilung der Daten nicht angenommen werden kann. Einen solcher Test, der Kolmogorov-Smirnov-Test wurde in Abs. 4.10 bereits erläutert. Weitere wichtige Tests sollen hier kurz beschrieben werden: der Mann-Whitney-U-Test, Kendall's τ-Test, Spearman's Rangkorrelationstest und der Wilcoxon-Test.

Rangzahlen und Bindungen

Die meisten nichtparametrischen Verfahren bauen auf dem Konzept der Rangzahlen auf. Dies setzt voraus, dass eine Ordnungsrelation auf den Daten definiert ist, d.h. mindestens eine Ordinalskalierung vorliegt. Liegen kontinuierliche Daten vor, bedeutet dies einen Informationsverlust, der aber verbunden ist mit einem Gewinn an Robustheit gegen *Outliers*.

Zunächst werden die N Werte einer Stichprobe $\{x_i\}$ aufsteigend sortiert und es wird die **Rangzahl** $rank(x_i)$ als natürliche Zahl $1, 2, \ldots, N$ gemäß der resultierenden Position vergeben.

Eine **Bindung** der Länge l nennt man das l-fache Auftreten eines Wertes. Alle Rangzahlen einer jeden Bindung werden im letzten Schritt durch ihren jeweiligen Positionsmittelwert ersetzt. Diese so genannten *midranks* können also halb- oder ganzzahlig sein. Die Summe aller Rangzahlen bleibt dabei unverändert $N(N-1)/2$. Im folgenden Beispiel

x_i	2	4	6	7	4	7	1	8	7
Rangzahl$_i$	2	3.5	5	8	3.5	8	1	9	8

liegen zwei Bindungen mit Länge $l = 2$ (4,4) und $l = 3$ (7,7,7) vor, wie man an der sortierten Reihe {1, 2, 4,4, 6, 7,7,7, 8} unmittelbar erkennen kann.

Der Mann-Whitney-U-Test

Ähnlich dem Student's t und dem Kolmogorov-Smirnov-Test ist der Mann-Whitney-U-Test für den Vergleich zweier unabhängiger Stichproben an-

wendbar. Er braucht aber keine Verteilungsannahmen und zeigt meist eine höhere Schärfe (*power*) als der Kolmogorov-Smirnov Test.

Die Nullhypothese H_0 besagt, dass die Stichproben N_1 und N_2 von derselben Verteilung stammen. Zunächst werden Rangzahlen für die *gemeinsame* Datenstichprobe vergeben, dann wird die Rangsumme R_1 für die erste Stichprobe ermittelt. Wie wahrscheinlich ist es nun, dass H_0 zutrifft und R_1 zufällig ist? Im Gedankenexperiment und in einer Monte-Carlo-Simulation können nun die Daten-Stichproben-Zuordnungen zufällig vertauscht werden. Wiederholt man dies (etwa 1000–10000 mal) und zeichnet die aggregierten Verteilung R_1' auf, ist der p-Wert ablesbar an der Perzentile von R_1 in $V_{R_1'}$.

Ein anderer Weg ist die **U-Statistik**: Die Stichproben N_1 und N_2 werden aufsteigend geordnet. O.B.d.A., sei $|N_1| <= |N_2|$, so werden für jedes Element von N_1 die Anzahl Elemente von N_2 aufsummiert, die links davon in der Liste stehen. Für große Stichproben ist die Summe U hinreichend normalverteilt mit dem Erwartungwert und der Varianz

$$E\{U\} = \frac{N_1 N_2}{2} \qquad Var\{U\} = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \,. \tag{4.71}$$

Für kleine Stichproben gibt es Vergleichstabellen für den kleineren Randabstand $U' = \min(U, N_1 N_2 - U')$ im Intervall möglicher Werte $U \in [0, N_1 N_2]$ (Mann und Whitney 1947).

4.12.3 Nichtparametrische Tests für abhängige Stichproben

Für verbundene, abhängige Stichproben ist neben dem McNemar-Test der **Vorzeichentest** (*sign test*) konzeptionell sehr einfach. Die Datenpaarungen werden nach Verbesserung „+“, Verschlechterung „-“ und Unverändert „=“ eingestuft (was einer Ordinalskalierung bedarf). Die „=“-Gruppe wird ausgeschlossen und ein Bernoulli-Experiment mit einer 50% Eintrittswahrscheinlichkeit für „+“ (versus „-“) angenommen und getestet, s. Abs. 4.11.

Der Wilcoxon-Test (*Wilcoxon Matched-Pairs Signed-Ranks-Test*)

Der Wilcoxon-Test ist schärfer und anspruchsvoller. Er berücksichtigt die Größe der Paardifferenzen $d_i = x_i - y_i$ der verbundenen Beobachtungen

x_i, y_i und erfordert dafür ein kontinuierliches Skalenniveau. Nachdem alle $d_i = 0$ ausgeschlossen sind, werden für die verbleibenden N' Absolutbeträge $\{|d_i|\}$ die Randzahlen bestimmt. Nun bildet man zwei Summen, T_+ für die Rangsumme aller ursprünglich positiven Differenzen $\{d_i|d_i > 0\}$ und T_- für $\{d_i|d_i < 0\}$. Sind x und y unabhängig, sollten auch die Differenzen in beide Richtungen „bunt" gemischt sein, also $T_+ \approx T_-$. Da die Summe konstant ist ($T_+ + T_- \frac{1}{2}N'(N'+1)$), wählt man ein T, z.B. $T = \min(T_+, T_-)$. Bei Gültigkeit der Nullhypothese verhält sich T bei großer Stichprobe ($N' > 100$) normalverteilt mit Erwartungswert und Varianz:

$$E\{T\} = \frac{1}{4}N'(N'+1) \qquad var\{T\} = \frac{N'(N'+1)(2N'+1)}{24}. \tag{4.72}$$

Für kleine Stichproben gibt es Tabellen für T, s. z. B. Zar (1996).

4.13 Weitere Assoziationsmaße für zwei Verteilungen

Im letzten Abschnitt wurde der χ^2-Test erläutert, der die Signifikanz einer allgemeinen Assoziation zwischen zwei Variablen beurteilen kann. Der Kontingenzkoeffizient und Cramer's V waren zwei abgeleitete Maße, um die Stärke dieser Gekoppeltheit auszudrücken. Leider geben diese Zahlen nur relative, aber kaum absolute Interpretationsmöglichkeiten.

Im folgenden wird *(i)* ein informationstheoretischer, entropie-basierter Ansatz vorgestellt, der nichts über die Signifikanz der Assoziation aussagt, dafür aber deren Stärke elegant quantifiziert. Ferner *(ii)* der lineare Pearson's Korrelationskoeffizient und *(iii)* Spearman's Rangkorrelationskoeffizient, ein nichtparametrisches Maß.

4.13.1 Entropie-basierte Assoziationsmaße

Betrachten wir das Spiel „Heiteres Beruferaten", in dem eine Sequenz von Fragen gestellt wird, um die gesuchte richtige Antwort aus sehr vielen möglichen Antworten herauszufinden. Das Ziel, möglichst wenige Fragen stellen zu müssen, wird erreicht, indem man geschickte Fragen stellt. Hier verallgemeinert man die ursprüngliche Spielregel mit ihrer Einschränkung

auf dichotome Ja-Nein-Fragen und erlaubt kategoriale 1-aus-I Fragen. Informationsträchtig sind die Fragen, die möglichst viele andere Möglichkeiten ausschließen. Claude Shannon schlug (1948) ein **Informationsmaß** vor, das eine Frage mit I Alternativen und deren a-priori-Wahrscheinlichkeiten p_i quantitativ bewertet

$$H = -\sum_{i=1}^{I} p_i \log_2 p_i. \tag{4.73}$$

Es besitzt wichtige Eigenschaften:

- H ist additiv, so dass einer Frage, die alle Möglichkeiten bis auf 1/6 ausräumt, gleich viel Informationsgehalt zugeschrieben bekommt wie ein Fragenpaar, das zuerst 1/2 und dann 1/3 der Möglichkeiten ausräumt;
- der Wert von H liegt zwischen 0 und $\log_2 I$ und hat sein Maximum, wenn alle Wertalternativen gleich wahrscheinlich sind $p_i = 1/I$. Das Minimum $H = 0$ beschreibt eine „wertlose“ Frage, d.h., eine mit sicherem Ausgang $p_{i^*} = 1$, deren andere „Alternativen“ nicht auftreten (beachte $\lim_{p\to 0} p \log p = 0$);
- Die Maßeinheit von H ist das **bit**, wenn man den Logarithmus dualis log_2 zur Basis 2 verwendet.

H wird häufig auch die **Entropie** einer Verteilung genannt, ein Begriff, der aus der statistischen Physik entlehnt wurde. Er erinnert daran, dass Entropiemaximierung mit Gleichverteilung der Eintrittswahrscheinlichkeiten einhergeht.

Die Assoziationsanalyse zweier kategorialer Variablen x und y prüft nun den Informationsgehalt der konkreten Antworten der Einzelfragen versus der kombinierten Antwort (x, y). Die Notation der Kontingenztabelle (Gl. 4.62, Abschnitt 4.12.1) wird ergänzt durch die empirisch relativen Häufigkeiten

$$\begin{aligned} p_{ij} &= \frac{N_{ij}}{N} \\ p_{i\cdot} &= \frac{N_{i\cdot}}{N} \qquad (I \text{ Antworten der Frage nach } x \text{ allein}) \\ p_{\cdot j} &= \frac{N_{\cdot j}}{N} \qquad (J \text{ Antworten der Frage nach } y \text{ allein}). \end{aligned} \tag{4.74}$$

Der Informationsgehalt der isolierten Fragen nach x und y ist damit

$$H(x) = -\sum_{i=1}^{I} p_{i\cdot} \log_2 p_{i\cdot} \ , \tag{4.75}$$
$$H(y) = -\sum_{j=1}^{J} p_{\cdot j} \log_2 p_{\cdot j}$$

und die der Kombinationsfrage

$$H(x,y) = -\sum_{i,j} p_{ij} \log_2 p_{ij} \tag{4.76}$$

Nun, was ist der Informationsgehalt der Frage nach y, wenn wir das Ergebnis der Frage x bereits kennen? Diese bedingte Entropie ist der mit $p_{i\cdot}$ gewichtete Erwartungswert der Entropieberechnungen in den jeweiligen Spalten i der Kontingenztabelle:

$$H(y|x) = \sum_i p_{i\cdot} \sum_j \frac{p_{ij}}{p_{i\cdot}} \log_2 \frac{p_{ij}}{p_{i\cdot}} = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_{i\cdot}} \tag{4.77}$$

Entsprechend ist die Entropie von x gegeben y:

$$H(x|y) = \sum_j p_{\cdot j} \sum_j \frac{p_{ij}}{p_{\cdot j}} \log_2 \frac{p_{ij}}{p_{\cdot j}} = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_{\cdot j}}. \tag{4.78}$$

Spielt y die Rolle der Zielvariablen, so wird $H(y|x)$ auch die **Vorwärtsentropie** und $H(x|y)$ auch die **Rückwärtsentropie** genannt (Pyle 1999). Mit Hilfe der Relation $\log_2 z \leq z-1$ kann man leicht zeigen, dass der Informationsgehalt der Frage nach y durch Vorwissen von x höchstens geschmälert werden kann, i.e.: $0 \leq H(y|x) \leq H(y)$. Genau damit lässt sich ein Maß des **Informationsgewinns** (***information gain***)

$$IG_y(x) = H(y) - H(y|x) \tag{4.79}$$

durch x auf dem Weg der Beantwortung der Frage nach y definieren. Es findet sowohl zur Attributcharakterisierung als auch zur Attributselektion Anwendung (s. Abs. 5.6.1). Die normierte Fassung wird gelegentlich auch **Unsicherheitskoeffizient (*uncertainty coefficient*)**

$$U(y|x) = \frac{H(y) - H(y|x)}{H(y)} \ \in [0,1] \tag{4.80}$$

oder **normierte Vorwärtsentropie** *normalized forward entropy* genannt und beschreibt die „Determiniertheit“ y von x präzise durch die relative Entropieschmälerung von $H(y)$ durch Vorwissen um x. Der Wert $U(y|x) = 0$ bedeutet, x und y sind völlig unkorreliert und x stellt keinen Informationsgewinn für y dar. $U(y|x) = 1$ besagt, x determiniert y vollständig (wobei $H(y|x) = 0$).

Natürlich kann auch die Umkehrung betrachtet werden:

$$U(x|y) = \frac{H(x) - H(x|y)}{H(x)} \in [0, 1]. \tag{4.81}$$

Der Umstand, dass die Assoziationskennzahl U von x und y nicht invariant gegen Variablentausch ist, mag zunächst irritieren: $U(x|y) \neq U(y|x)$. Betrachten wir zum Beispiel eine kontinuierliche Zufallsvariable z, die je zweimal aufgezeichnet wird: x notiert sie ganzzahlig abgerundet und y halbzahlig abgerundet. Weiß man y, ist x vollständig bestimmt: $U(x|y) = 1$. Hingegen ist $U(y|x)$ kleiner, da ein bekannter x-Wert hier zu je zwei möglichen Halbzahlen y assoziiert ist.

Wünscht man dennoch eine symmetrische Assoziationskennzahl $U(x, y)$, ist das entropiegewichtete Mittel die geeignete Kombination:

$$U(x, y) = \frac{H(x)U(x|y) + H(y)U(y|x)}{H(x) + H(y)} = 2\frac{MI(x, y)}{H(x) + H(y)} \tag{4.82}$$

$$MI(x, y) = H(x) + H(y) - H(x, y)\ . \tag{4.83}$$

$U(x, y) \in [0, 1]$ kann auch als die normierte Form der **wechselseitigen Information** $MI(x, y)$ (*mutual information*) der Verteilung x und y betrachtet werden. Auch hier wird das Minimum 0 bei perfekter Unabhängigkeit erreicht und das Maximum $U(x, y) = 1$, wenn x und y sich wechselseitig vollständig festlegen (wobei dann $H(x) = H(y) = H(x, y) = MI(x, y)$). Eine Anwendung der Assoziationskennzahlen findet sich in Abs. 9.6.

4.13.2 Lineare Korrelation

Ein wichtiges Assoziationsmaß für Paare von kontinuierlichen (oder ordinalen) Variablen ist der lineare Korrelationskoeffizient, oder **Pearson's** $\mathbf{r}$. Er basiert auf dem Konzept der Konstruktion eines linearen Modells (siehe lineare Regression in Abs. 5.7.1) und wird durch die Mittelwerte $\bar{x}, \bar{y}$ wie

folgt berechnet:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad \in [-1, 1]. \tag{4.84}$$

Der Maximalwert $r = 1$ wird für vollständige, „positive Korrelation“ erreicht, indem alle Punkte exakt auf einer ansteigenden Geraden $y(x)$ liegen. Liegen alle Punkte entlang einer fallenden Geraden (mit negativer Steigung), sind x und y „negativ korreliert“ und $r = -1$. Je dichter die Punkte an der Regressionsgeraden liegen, umso näher liegt $|r|$ bei 1. Ein $r \approx 0$ bedeutet, dass keine lineare Korrelation vorliegt. Für zwei unkorrelierte, unabhängige (ungefähr standardnormale) Verteilungen sind die zufällig erwartbaren r-Werte stark von der Stichprobengröße abhängig. Ein Signifikanztest für die Ablehnung der Nullhypothese ist mit der Student's-t-Statistik

$$t = r\sqrt{\frac{N-2}{1-r^2}} \qquad \text{mit d.f.} = \nu = N - 2. \tag{4.85}$$

möglich (Abs. 4.7.4).

4.13.3 Nichtparametrische Korrelationsmaße

Sind x und y perfekt nichtlinear korreliert, z.B. $y = x^2$, wird Pearson's r irrtümlicherweise Assoziationsdefizite anzeigen. Ferner ist r störanfällig für Ausreißer, da die Differenzterme in Gl. 4.84 quadratisch gewichtet sind.

Spearman schlug eine nichtparametrische Variante vor, die allein auf Rangzahlen der originalen Werte beruht. Zunächst werden sie getrennt für x und y – als $R_i = rank(x_i)$ und $S_i = rank(y_i)$ – bestimmt (*midranks*, siehe Abs. 4.12.2). Ein streng monotoner Zusammenhang wird sich nun in der perfekten Korrelation der Ränge widerspiegeln. Die Berechnung von Pearson's r (vgl. Gl. 4.84) ergibt den so genannten **Spearman's Rangkorrelationskoeffizient**

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}} \quad \in [-1, 1]. \tag{4.86}$$

Als Signifikanztest für r_s wird der Student's-t-Test analog zu Gl. 4.85 verwendet.

Nimmt man die Quadratsumme der Rangdifferenzen

$$D = \sum_{i=0}^{N} (R_i - S_i)^2 \,, \tag{4.87}$$

stellt sich ein exakter Zusammenhang von D zu Spearman's-r_s heraus:

$$r_s = k_1 \left(1 - \frac{6(D + k_0)}{N^3 - N}\right). \tag{4.88}$$

k_0 und k_1 dienen als Korrekturterme, wobei $k_0 = 0$, $k_1 = 1$, wenn keine Bindungen vorliegen. Andernfalls verändert dies das erwartbare D und die Korrekturterme berechnen sich aus den Bindungshäufigkeiten. Die k-te x-Gruppe enthalte f_k gleiche x-Werte (mit gleichem *midrank*), die m-te Gruppe enthalte g_m gleiche y-Werte. Dann verändern sich die beiden Korrekturterme zu

$$\begin{aligned} k_0 &= \frac{1}{2}\sum_k (f_k^3 - f_k) + \frac{1}{2}\sum_m (g_m^3 - g_m) \\ k_1 &= \left[1 - \frac{\sum_k (f_k^3 - f_k)}{N^3 - N}\right]^{-1} \left[1 - \frac{\sum_m (g_m^3 - g_m)}{N^3 - N}\right]^{-1}. \end{aligned} \tag{4.89}$$

Ein alternativer direkter Signifikanztest für die Nullhypothese zweier unkorrellierter Variablen ist die erwartete Normalverteilung von D. Die Quadratsumme der Rangdifferenzen hat den Erwartungswert und die Varianz

$$E\{D\} = \frac{1}{6}(N^3 - N - k_0) \qquad Var\{D\} = \frac{(N-1)N^2(N+1)^2}{36k_1}. \tag{4.90}$$

Spearman's $|r_s|$ ist ein relativ robustes Korrelationsmaß für zwei Variablen, das keinen streng linearen, aber monotonen Zusammenhang vergleicht. Einen Assoziationszusammenhang, etwa eine sinusförmige ($y = \sin x$) oder kreisartige ($x^2 + y^2 = 1$) Relation, wird man mit Spearman's r_s nicht klar erkennen. Der Berechnungsaufwand wird durch die beiden Rangzahlsortierungen mit $N \log N$ dominiert.

Durch Gruppieren von Werten in Intervallen (engl. *binning*) ist die Methode auch auf alle anderen Datentypen anwendbar, was aber in der Regel mit einem Informationsverlust einhergeht.

Kapitel 5

Modellbildung

Ein wichtiger Schritt im *Datamining*-Prozess ist die Modellbildung. Hier vollzieht sich ein essentieller Repräsentationswechsel, weg von den Rohdaten, hin zu einer kompakteren Modellrepräsentation. Diese versucht, nützliche, kompakte Ausschnitte der Welt (oder auch einer virtuellen Realität), die in den Daten enthalten ist, zu verschiedenen Zwecken für den Computer bzw. den Menschen nachvollziehbar zu machen.

Heute gibt es eine Fülle von Datamining-Algorithmen, die für die Lösung dieser Aufgabe entwickelt wurden. Dieses Kapitel gibt einen Überblick über die wichtigsten Algorithmen. Einige wichtige Vertreter, die auch in den folgenden Kapiteln Verwendung finden, werden detaillierter dargestellt. Eine vollständige Darstellung aller relevanten Arbeiten würde den Rahmen dieser Arbeit sprengen, somit ist an vielen Stellen auf weiterführende Quellen verwiesen.

Die Bildung eines Modells beinhaltet mehrere Kernaspekte, die das Modell charakterisieren und gleichzeitig zur Einordnung dienen:

1. Zunächst steht die Datamining-**Aufgabe** im Vordergrund. Was ist das Hauptziel des Modells? Dient es der Visualisierung, der unüberwachten Clusterbildung, der Klassifikation, der Approximation oder dem Finden von Mustern und Regeln;
2. Die **Struktur** des Modells definiert die Grenzen dessen, was das Modell lernen oder approximieren kann. Sie kann z.B. die funktionale Form einer Approximation, ein neuronales Netzwerk oder ein hierarchischer Clusteransatz sein. Die gegebenen Daten treiben die in Parametern kodierte Ausformung des Modelles;

3. Die **Bewertungsfunktion** (*scoring function*) wird benötigt, um die Qualität des Modelles anhand von beobachteten Daten zu messen. Sie heißt auch **Zielfunktion**, denn sie leitet die Bestimmung der Modellparameter, indem die *scoring*-Funktion minimiert (oder maximiert) wird. Sie misst die Modellgüte, z.B. durch Missklassifikations- oder Approximationsfehler. Sie ist entscheidend, sowohl für die Lern- als auch für die Generalisierungsleistung des Modelles, und sollte daher den tatsächlichen Gesamtnutzen des Modells so realitätsnah wie möglich widerspiegeln;

4. Die **Such- und Optimierungsverfahren** zum Finden der Modellparameter können z.B. iterative Gradientenverfahren sein, die die Bewertungsfunktion optimieren. Ist die Modellstruktur fix, werden nur Koeffizienten und Gewichte angepasst, ansonsten können zusätzlich auch Modellstrukturparameter gesucht werden;

5. Eine weitere Komponente können spezielle **Datenmanagementtechniken** sein, die Verwendung finden, um große Datenmengen zu speichern, zu indizieren und zu verarbeiten. Viele Algorithmen spezifizieren keine Datenmanagementtechniken. Sie implizieren oft, dass die Datenmengen nicht sehr groß sind und keine besonderen Kosten für beliebigen wahlfreien Datenzugriff (*random access*) auftreten. Passen nicht mehr alle Daten in den Hauptspeicher, werden die Verarbeitungskosten durch wiederholte Sekundärspeichertransfers (Plattenspeicher) erheblich vergrößert. Wird die zu verarbeitende Datenmenge sehr groß, so wird die Frage nach der geschickten Reduktion der insgesamt zu verarbeitenden Daten immer wichtiger.

Abhängig von der Aufgabenstellung kann die Verstehbarkeit des Modelles ein große Rolle spielen: Die Modelle können einfach und für den Menschen leicht nachvollziehbar oder eher undurchsichtig und *black-box*-artig sei. Insbesondere in sicherheitskritischen Bereichen haben undurchschaubar komplexe Modelle große Akzeptanzprobleme. Sind sie Dritten schwer vermittelbar, kommen sie selbst dann kaum zum Einsatz, wenn sie z.B. statistisch eindeutig bessere Prädiktionsleistungen zeigen.

5.1 Bayes'sche Modelle und Methoden

Sind empirische Merkmalsverteilungen rein zufällig – oder sind sie voneinander abhängig? Wenn ja, wie kann man diese Abhängigkeiten beschreiben? Diese Frage führt zunächst zu den Bayes'schen Methoden.

Zentral für die Bayes'sche Methoden ist das gleichnamige Theorem für bedingte Wahrscheinlichkeiten Gl. 4.5 (S. 51), das dem britischen Prediger Thomas Bayes (1701–1761) zugeschrieben wird. Zwei unveröffentlichte Essays wurden aus seinem Nachlass der Royal Society zugesandt, die aber erst wirklich Aufmerksamkeit erlangten, als der französische Mathematiker Pierre-Simon Laplace zur selben Erkenntnis gelangte.

Die Bayes'schen Modelle sind zum einen deskriptiv, zum anderen können sie zu probabilistischen Aussagen über Hypothesen und zu Rückschlüssen (Inferenzen) über die Auftretenswahrscheinlichkeit von Zuständen genutzt werden (s. Prädiktion und Diagnose in Abb. 5.1).

Die Bayes'schen Methoden erlauben Beobachtungen mit Vorwissen zu kombinieren, um die Wahrscheinlichkeit $p(h_i)$ einer Hypothese h_i zu bestimmen. Zunächst kann man generell nach der besten Hypothese fragen. Mit anderen Worten: Wenn eine Menge von Hypothesen vorliegt, welche Hypothese H ist dann die wahrscheinlichste bei Vorliegen aller Daten X?

$$\begin{aligned} h_{MAP} &= \arg\max_{h_i \in H} p(h_i|X) \\ &= \arg\max_{h_i \in H} \frac{p(X|h_i)p(h_i)}{p(X)} \\ &= \arg\max_{h_i \in H} p(X|h_i)p(h) \end{aligned} \tag{5.1}$$

Sind die Hypothesen h_i alle gleichwahrscheinlich, wird die **maximale aposteriori Hypothese** (**MAP**-Hypothese) $h_{MAP} = h_{ML}$ gleich der ***maximum likelihood*** **Hypothese** (**ML**-Hypothese)

$$h_{ML} = \arg\max_{h_i \in H} p(X|h_i) \ , \tag{5.2}$$

wobei der Ausdruck $p(X|h)$ als die ***Likelihood*** bezeichnet wird. Das Beispiel eines medizinischen Tests in Box 4.1 (Seite 52) zeigt, dass das MAP-Ergebnis sehr stark von den a-priori-Wahrscheinlichkeiten $p(h)$ beeinflusst werden kann. Das Prinzip der Maximierung der *Likelihood* wird häufig auch zur Bestimmung kontinuierlicher Parameter verwandt, siehe Abs. 5.7.2.

Der **Bayes'sche Klassifikator** fragt nach der wahrscheinlichsten Klassifizierung, wenn eine neue Beobachtung $\mathbf{x}$ gegeben ist. Zum Beispiel

(nach Mitchell 1997) wurden drei Hypothesen – eine positive h_1 und zwei negative h_2, h_3, mit a-posteriori Wahrscheinlichkeiten $p(h_1|X) = 0.4$ und $p(h_2|X) = p(h_3|X) = 0.3$ – bewertet, wobei für einen konkreten neuen Merkmalstupel $\mathbf{x}$ nur die negativen (-) Hypothesen h_2 und h_3 zutreffend sind. Dies zeigt, dass die h_{MAP} nicht die wahrscheinlichste Klassifikation sein muss (hier ist die negative $p(-) = 0.3 + 0.3 = 0.6$ versus die positive $p(+) = 0.4$).

Allgemein ist die **optimale Bayes'sche Klassifikation** v_j aus einer Menge von möglichen Klassifikationsergebnissen V durch Maximierung von

$$p(v_j|\mathbf{x}) = \sum_{h_i \in H} p(v_j|h_i)p(h_i|\mathbf{x}) \tag{5.3}$$

gegeben:

$$v_{Bayes} = \arg\max_{v_j \in V} \sum_{h_i \in H} p(v_j|h_i)p(h_i|\mathbf{x}). \tag{5.4}$$

Besonders einfach ist der **Naive Bayes'sche Klassifikator**. Er geht allerdings von der einschneidenden Annahme aus, dass alle Merkmale *unabhängig voneinander* die Wahrscheinlichkeitsfunktion

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_m) = \prod_{k=1}^{m} p(x_k) \tag{5.5}$$

beeinflussen und daher die Dichtefunktion $p(\mathbf{x})$ faktorisiert werden kann. Dies führt (analog zu Gl. 5.1) zum **N**aiven **B**ayes'schen Klassifikator

$$v_{NB} = \arg\max_{v_j \in V} p(v_j|\mathbf{x}) = \arg\max_{v_j \in V} p(v_j) \prod_{k=1}^{m} p(x_k|v_j). \tag{5.6}$$

Vorteilhaft ist, dass der Berechnungsaufwand relativ gering ist, da nur (Merkmalsanzahl) m viele univariate Dichteschätzer trainiert werden müssen. Das Naive Bayes-Verfahren wird zum Beispiel zur Kategorisierung von Nachrichtentexten oder zur Klassifikation von unerwünschter Email (*spam*) verwendet. Für jede Kategorie-Wort-Paarung (v_j, x_k) wird der bedingte Erwartungswert $p(x_k|v_j)$ anhand der Auftretenshäufigkeiten über eine Trainingstextmenge durch Abzählen ermittelt. Ein neues Dokument mit Worthäufigkeiten $\mathbf{x}$ wird gemäß Gl. 5.6 klassifiziert. Weitere Subtilitäten im Umgang mit verschwindenden $p(x_k|v_j) = 0$ sowie weitere Beispiele finden sich z.B. in Mitchell (1997).

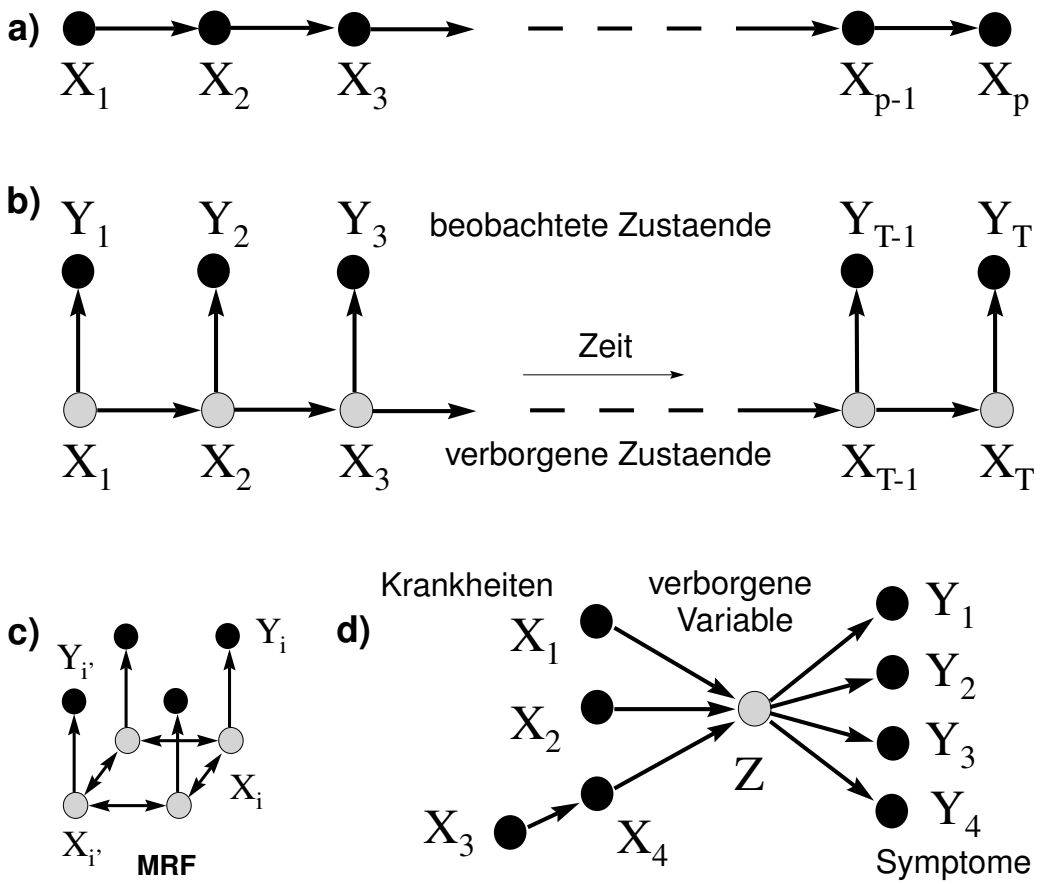


Abbildung 5.1: Graphische Modelle: *(a)* Struktur einer Markov-Kette erster Ordnung; *(b)* Struktur eines *Hidden Markov Model* (HMM) erster Ordnung, das z.B. in der Spracherkennung zeitliche Zustandsübergänge modelliert. Die internen (grau gezeigten) Zustände X_i sind verborgen und können nur indirekt über die Zustände Y_i beobachtet werden; *(c)* in *Markov Random Field* (MRF) beeinflussen sich die verborgenen Zustände X_i wechselseitig, z.B. auf einem Gitter von Bildpixeln; *(d)* in einem typischen *Belief Net* werden Zustandsabhängigkeiten in einem azyklischen Graph, ggf. auch mit verborgenen Zuständen, modelliert. Dies erlaubt Inferenzen: von links nach rechts (in Pfeilrichtung) *Prädiktionen* (hier von Symptomen) und umgekehrt *Diagnosen* (hier von Krankheiten).

Obwohl die Unabhängigkeitsannahme oft grob verletzt ist, führt dieser naive Ansatz mitunter zu erstaunlichen Erfolgen (s.o.). Weitere interessante Konzepte wurden entwickelt, um die Unabhängigkeitsannahme gezielt einzuschränken und um Wissen über Merkmalszusammenhänge in Strukturen zu fassen.

Zuerst seien die **Markov-Ketten** genannt, die die allgemein gültige Verbunddichtefunktion (s. Kettenregel aus Gl. 4.6)

$$p(\mathbf{x}) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2) \cdots p(x_m|x_1, x_2, \ldots x_{m-1}) \tag{5.7}$$

dahingehend einschränken, dass eine bedingte Wahrscheinlichkeit nurmehr von einem (oder r) Vorgänger(n) abhängen darf $p(x_k|x_1, \ldots, x_{k-1}) = p(x_k|x_{k-1})$. Diese Struktur wird Markov-Kette erster (bzw. r-ter) Ordnung genannt und ist in Abb. 5.1a skizziert.

Eine weitere konzeptionelle Zutat sind die verborgenen Zustände (*hidden states*). Dies sind interne Merkmale, die nicht direkt beobachtet werden können, sondern sich nur indirekt und probabilistisch über so genannte Observablen y_i bemerkbar machen. Abb. 5.1b zeigt die Struktur dieses Grundmusters, das **Hidden Markov Model (HMM)** erster Ordnung. Die HMMs werden in der Erkennung gesprochener Sprache sehr erfolgreich eingesetzt (s. z. B. Rabiner 1989; Jelinek 1997). Zum Trainieren der nicht-beobachtbaren internen Zustandsschätzer wurden speziell Verfahren, wie der Viterbi- und der Baum-Welsch-Algorithmus (äquivalent dem EM-Verfahren) entwickelt.

Eine Verallgemeinerung sind die **graphischen Modelle**, die Merkmalszusammenhänge in Graphen notieren. Zur Gruppe der ungerichteten Modelle zählen die **Markov-Random-Field**-Ansätze, die z.B. im Bereich des Computersehens Anwendung finden (Abb. 5.1c). Die gerichteten Modelle sind im Bereich des maschinellen Lernens unter den Namen **Bayes'sche Netze** oder ***Belief Nets*** verbreitet. Sie bieten eine präzise Formulierung der Merkmalsbeziehungen durch azyklische Graphen sowie Berechnungsvorschriften für bedingte Wahrscheinlichkeiten. Statt der vollständigen Faktorisierung wie in Gl. 5.7 wird die Verbunddichtefunktion

$$p(\mathbf{x}) = \prod_{j=1}^{m} p(x_j|pa(x_j)) \tag{5.8}$$

durch ein Produkt bedingter Wahrscheinlichkeitsfunktionen ersetzt. $pa(x_j)$ bezeichnet den Satz von Elternknoten der Variablen x_j. In Abb. 5.1d sind z. B. X_1, X_2, X_3, X_4, Z die Vorgänger von Y_1. Die Verbindungsstruktur des Graphen impliziert die bedingte Unabhängigkeit für $p(\mathbf{x})$ dahinhingehend, dass eine Variable x_j von x_i unabhängig ist, wenn x_i nicht zu den Vorgängern im azyklischen Graphen gehört. Ist die Elternknotenzahl viel kleiner als die Variablenanzahl, ist dies eine erhebliche Vereinfachung im Vergleich zum vollen Modell.

Vorwissen lässt sich in *Belief Nets* direkt in die graphische Struktur integrieren. Sind die Merkmale nur partiell beobachtbar, werden die verborgenen Zustandswahrscheinlichkeiten z.B. mittels des EM-Verfahrens bestimmt. Ist die Struktur unbekannt, können strukturelle EM-Verfahren eingesetzt werden. Wie viele nicht-triviale Strukturoptimierungsfragen ist dies ein NP-hartes Problem und Gegenstand aktueller Forschung (s. z. B. Buntime 1994; Jordan 1999).

Die Bezeichnung Bayes'sche Netze bezieht sich auf die Bedeutung des Bayes'schen Theorems und nicht auf die ausschließliche Verwendung des Bayes'schen Klassifikators an den Knoten. Die Bayes'schen Netze umfassen auch Knotenschätzer, die die bedingten Wahrscheinlichkeitsfunktionen durch andere Methoden, z.B. Mischungsmodellansätze (Gl. 4.34), oder durch neuronale Netze approximieren.

5.2 Approximationsmodelle

Viele Modellierungsaufgaben können als Approximationsproblem formuliert werden. Voraussetzung ist, dass die Daten als Merkmalsvektoren repräsentiert (s. Kap. 2) werden können und sich eine Frage nach einer kontinuierlichen Ausgangsgröße, in Abhängigkeit von einem Merkmalsvektor $\mathbf{x}$, stellen lässt. Gesucht ist dann die Abbildung

$$F(\mathbf{w}, \mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R} \ , \tag{5.9}$$

wobei $\mathbf{w}$ die Parametrisierung eines gewählten Modelltyps umschreibt, der hier, o.B.d.A., mit einer eindimensionalen Ausgabe notiert ist. Je nach Verfahrensherkunft wird die Suche nach $F()$ als Regression (Statistik), Approximation (Mathematik) oder auch als Lernen (Neuronale Netze) benannt. Im Folgenden sind typische Modellstrukturen aufgezeigt.

Die lineare Regression basiert auf der Darstellung des Eingabe-Ausgabe-Zusammenhangs als einem Skalarprodukt

$$F(\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x} \tag{5.10}$$

mit Eingabevektor $\mathbf{x}$, der, zur Vereinfachung der Schreibweise, um die Konstante 1 erweitert ist (s. Abs. 5.7.1).

Die logistische Regression erweitert Gl. 5.10 um eine streng monotone steigende, sigmoide Funktion

$$F(\mathbf{w},\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} \; . \tag{5.11}$$

Sie ist auf das Intervall [0,1] beschränkt und wird insbesondere zur Modellierung von Wahrscheinlichkeiten verwendet, s. Abs. 5.7.2.

Das lineare und das logistische Regressionsschema ist äquivalent dem **Perzeptron**, einer klassischen Form von neuronalen Netzen (s.u. „MLP“).

Das klassische Approximationsschema ist eine Linearkombination aus einem Satz geeigneter Basisfunktionen $\{B_i\}$ über dem Eingabevektor $\mathbf{x}$

$$F(\mathbf{w},\mathbf{x}) = \sum_{i=1}^{m} w_i \, B_i(\mathbf{x}) \tag{5.12}$$

Dies schließt auch Polynome, Splines und Fourierdarstellungen mit ein.

Projection Pursuit Regression verwendet Approximationsfunktionen, die sich als Summe univariater Funktionen U_i von einer Linearkombination der Eingabevariablen darstellen lassen:

$$F(\mathbf{w},\mathbf{x}) = \sum_{i=1}^{m} U_i(\mathbf{w}_i\cdot\mathbf{x}) \; . \tag{5.13}$$

Vorteilhaft ist, dass affine Transformationen der Eingabevektoren direkt kompensiert werden können (Friedman 1991).

Normalisierte Radiale Basisfunktionen (RBF): RBF-Netzwerke

$$F(\mathbf{w},\mathbf{x}) = \frac{\sum_i \rho(|\mathbf{x}-\mathbf{u}_i|)\; y_i}{\sum_i \rho(|\mathbf{x}-\mathbf{u}_i|)} \tag{5.14}$$

gestalten sich als gewichtete Summe von Basisfunktionen $\rho()$, die jeweils nur vom Abstand eines Referenzpunktes $\mathbf{u}_i$ im Merkmalsraum abhängen ($\mathbf{w}$ umfasst alle Parameter $\{\mathbf{u}_i, y_i\}$). Der Nenner sorgt für eine Normierung und eine flache Extrapolation. Dies wird in Abb. 5.2 anhand der am häufigsten verwendeten Basisfunktion, der Gauß'schen Glockenkurve $\rho(r) = e^{-(r/\sqrt{2}\sigma)^2}$, illustriert (Powell 1987). Generalisierte RBF (Hyper-RBF) erhöhen die Adaptationsfähigkeit durch Formulierung von individuellen elliptischen Basisfunktionen, die mittels der *Maximum-Likelihood*-Methode trainiert werden können (s.a. Abs. 5.7.3).

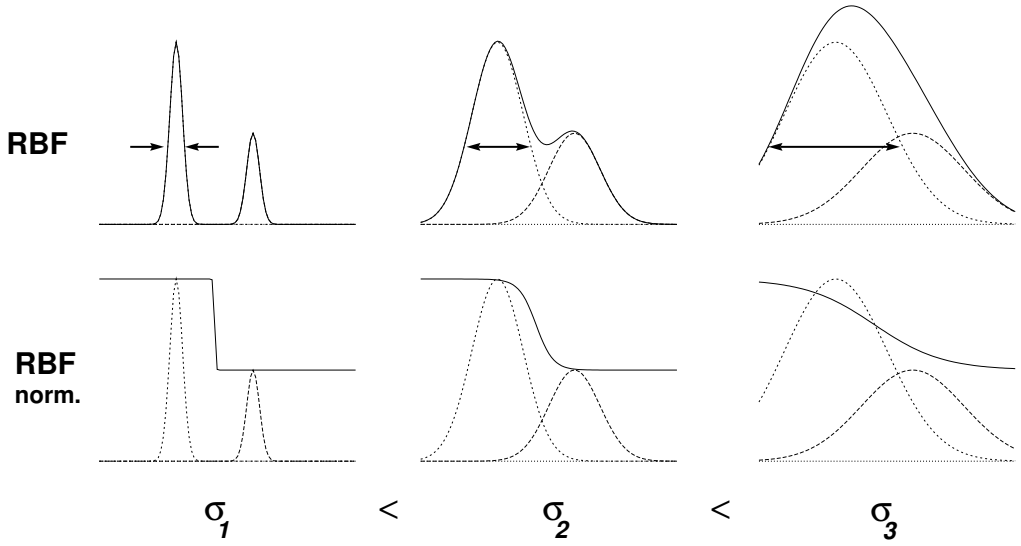


Abbildung 5.2: Approximation mit zwei RBFs (radiale Basisfunktionen). *Oben*, das direkte RBF-Verfahren und *unten* nach dem Normalisierungsschritt. Von *links* nach *rechts* illustrieren drei unterschiedliche Gaußglockenbreiten $\sigma_{1\cdots3}$ den Glättungseffekt eines wachsenden Breiten-Abstands-Verhältnisses.

Multilagen-Perzeptron (MLP) und verschachtelte Sigmoiden können als

$$F(\mathbf{w},\mathbf{x}) = g\left(\sum_k u_k\, g\left(\sum_j v_{kj}\, g\left(\cdots g\left(\sum_i w_{ji}\, x_i\right)\cdots\right)\right)\right) \quad (5.15)$$

geschrieben werden, wobei $g(\cdot)$ eine sigmoide Transferfunktion (z.B. von der Art Gl. 5.11) und $\mathbf{w} = (w_{ji}, v_{kj}, u_k, \cdots)$ die synaptischen Eingangsgewichte eines Neurons bezeichnet. Diese Darstellung war völlig neu für die klassische Theorie der Approximation (Poggio und Girosi 1990) und es konnte gezeigt werden, dass ein MLP als „universeller Approximator“ Funktionen beliebig genau modellieren kann (Hornik, Stinchcombe und White 1989). Neuronale Netze werden weiter in Abs. 5.8 beschrieben.

Topographische Karten sind dimensionsreduzierende Modelle, die von einem höherdimensionalen Merkmalsraum in eine niedrigdimensionale, zunächst diskrete, Menge abbilden. Sie werden später in Abs. 5.9 behandelt.

Regressionsbäume entstehen durch rekursive, achsparallele Partitionierung des Merkmalsraumes, z.B. das CART- (*Classification and Re-*

gression Tree, Breimann, Friedman, Olshen und Stone 1984)) und das MARS- (*Multivariate Adaptive Regression Splines* Friedman 1991) Verfahren (s. a. Klassifikationsbäume Abs. 5.3). Die rechteckigen Raumteilungen werden in Binärbäumen gespeichert und mit konstanten oder meist mit uni- oder bivariaten polynomialen Regressionen verküpft.

Scoring-funktion*:* Approximationsmodelle werden anhand einer *lack-of-fit*-Funktion $LOF(F)$ beurteilt , die als Erwartungswert

$$LOF_D(F) = \langle dist\left(f(\mathbf{x}), F(\mathbf{w}, \mathbf{x})\right)\rangle_D \tag{5.16}$$

des Approximationsfehlers in einer bestimmten Domäne D definiert ist. Meist ist die Distanzfunktion $dist(\cdot)$ einfach die euklidische Distanz zwischen dem Soll- $f(\mathbf{x})$ und dem Schätzwert des Modells $F(\mathbf{w}, \mathbf{x})$.

Ist das Modell reich an Freiheitsgraden (Parametern), tritt bei endlich großen, rauschbehafteten Trainingsstichproben das grundsätzliche Lernproblem des *over-fitting* auf. D.h. dem Modell gelingt es, sich zufälligen (nichtrepräsentativen) Feinheiten des Trainingsdatensatzes anzupassen, die der Generalisierungsleistung des Modells bei der Anwendung auf neuen Daten schadet. Gegenstrategien beinhalten die Reduktion der Freiheitsgrade (z.B. *optimal brain damage*) und Verkürzung der Lernzeit (*early stopping*), die auf dem Monitoring des Over-fitting-Effektes auf der Basis von Gl. 5.16 beruhen.

Um eine realistische Schätzung des Erwartungswertes von Gl. 5.16 zu erhalten, teilt man den zur Verfügung stehenden Datensatz in eine Trainings- und in eine Testdatenmenge. Sind die Datenmengen nicht sehr groß, kann die Technik der n**-fachen Kreuzvalidierung** (*cross validation*) nützlich sein. Man teilt den Datensatz in n vergleichbare Teile und wiederholt einen Training-Test-Zyklus, bei dem der i-te Teil für das Testen reserviert ist und der Rest dem Modelltraining dient. Nach n Wiederholungen war jedes Datum genau einmal an der Testmenge beteiligt und die Gesamtleistung kann zusammengefasst werden.

Such- und Optimierungsverfahren bedienen sich in einfachen Fällen linearer Verfahren, meist sind es jedoch iterative Gradientenabstiegsverfahren für Gl. 5.16. Später in diesem Kapitel werden u.a. das Levenberg-Marquardt- und das EM-Verfahren vorgestellt.

5.3 Klassifikation

Aufgabe der Klassifikation ist es, Objekte aufgrund ihrer Attributwerte einer der *vorgegebenen* Klassen zuzuordnen. Sind keine Klassen vorgegeben, stehen unüberwachte (*unsupervised*) Clusterverfahren im Vordergrund, die in Abs. 5.4 erläutert werden. Die Aufgabenstellung ist im Prinzip ähnlich der der Approximation, nur wird ein kontinuierlicher Ausgabewert durch eine diskrete Klassen- oder Kategoriezugehörigkeit ersetzt. Neben dem Bayes'schen Klassifikator (s. o.), können, wie erwähnt, prinzipiell alle obigen Approximationsfunktionen eingesetzt werden, um neue Objekte zu klassifizieren. Die theoretische Verknüpfung sind die Bayes'schen Methoden, siehe Abs. 5.1.

Die generelle Bewertungsfunktion von Gl. 5.16 misst für Klassifikatoren typischerweise den Erwartungswert der Fehlklassifikationskosten: z.B. im einfachsten Falle 0 für eine korrekte und Kostenbeitrag 1 für eine inkorrekte Klassenschätzung. Sie können aber auch asymmetrisch definiert werden. Schätzer für eine dichotome Variable sind in vielen Bereichen von großem Interesse. Um die Güte eines ordinalen oder probabilistischen Klassifikators zu messen, kann man ein integrales Maß auf der Basis der so genannten ROC-Kurve ermitteln, s, Abs. 5.7.4.

Ein verbreiteter Strukturansatz für Klassifikatoren sind die **Entscheidungsbäume (*Decision Trees*, DT)**, die ihre Beliebtheit aus einer sehr guten Nachvollziehbarkeit schöpfen. Abb. 5.3 zeigt einen solchen Entscheidungsbaum für die Aufgabe, die Herkunftsregion für Automodelle (Datensatz aus Kap. 3 s. Abb. 3.10) aufgrund von Fahrzeugmerkmalen zu schätzen. Für jedes Objekt ergibt sich genau eine Entscheidungskette: Ausgehend vom Wurzelknoten (in der Abb. oben) wird an jedem Knoten (i.d.R.) ein Attribut ausgewertet und entsprechend des konkreten Wertes weiterverzweigt, bis ein terminales „Blatt" erreicht ist. In Abb. 5.3 werden neben den Entscheidungskriterien inklusive der χ^2-Statistik auch die Kategoriehäufigkeiten an jedem Knoten tabelliert und durch Balken visualisiert. Die Entscheidungskriterien wurden mit dem Ziel ausgewählt, möglichst „sortenreine" Blätter zu erhalten. An den daraus resultierenden Entscheidungskriterien sind unmittelbar interpretierbare Regeln ablesbar, z.B. umfasst Knoten 5 die hubraumstärksten Modelle, die alle zur Gruppe USA (1 = rot) gehören (s.a. vergleichenden Darstellungen in Kap. 3). Die zweigrößte Hubraumgruppe teilt sich (an Knoten 4) in eine 1=USA dominierte Gruppe (Knoten 11) mit weniger PS und eine USA-Europa-gemischte Gruppe (1+2 = rot+grün) mit mehr PS.

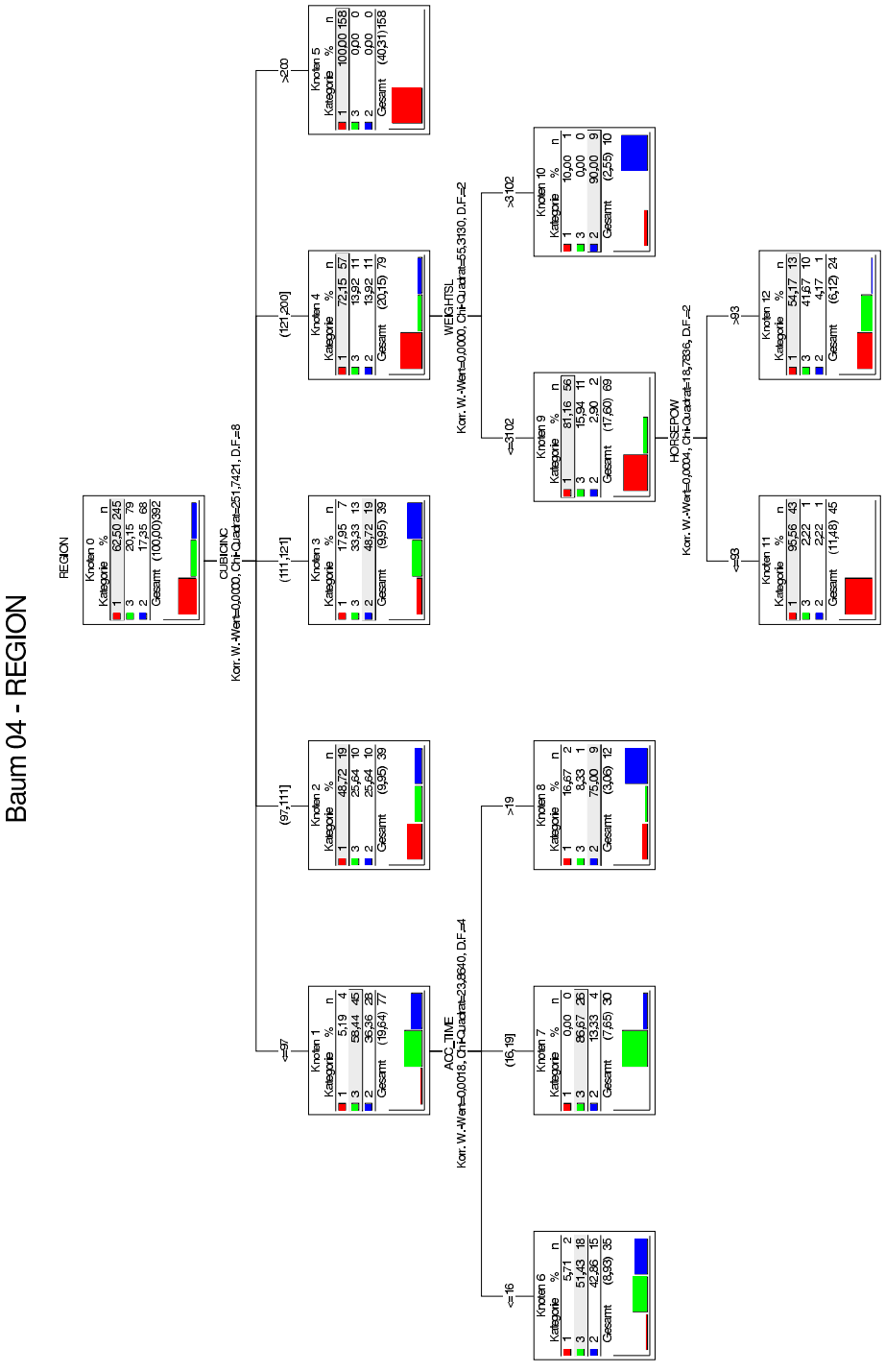


Abbildung 5.3: Entscheidungsbaum zur Klassifizierung der Herkunftsregion einer Automarke (Kodierung USA=1, Europa=2, Japan=3; Datensatz s. a. Abb. 5.3).

Wichtige Vertreter von Entscheidungsbaumverfahren sind der klassische C4.5 Algorithmus von Quinlan (1993), Quest von Loh und Shih (1997) und neuere, hochskalierbare Entwicklungen wie RainForest (Gehrke et al. 2000). Der Nachteil von Entscheidungsbaumverfahren liegt darin, dass zum einen „der optimale Baum“ i.A. nicht existiert, sondern im Gegenteil die Ergebnisse erheblich von den notwendigen Heuristiken und Parametern zum Finden bester Trennkriterien abhängen können. Zum anderen wird ihre Klassifikationsleistung von anderen Klassifikationsverfahren meist übertroffen, wie z.B. in der breit angelegten Vergleichsstudie STATLOG[1] (Michie et al. 1994) dargestellt und analysiert wird.

Zu den besten Klassifikationsverfahren gehören moderne ***Support Vector Machines*** **(SVM)**. Das Grundprinzip ist das Folgende: In einem hochdimensionalen Raum wird ein linear separables Zweiklassenproblem betrachtet und die Hyperebene (Gl. 5.10) gesucht, die alle Vertreter der einen von der anderen Klasse optimal trennt. Optimal heißt, die Trennfläche ist gleich weit und maximal von den nächsten Vertretern beider Klassen entfernt (*Optimal-margin*-Kriterium). Da nur diese letztlich entscheidend sind, nennt man sie *support vectors*. Interessant wurde die Methode aus zwei Gründen: Zum eine entdeckte man den so genannten „Kerneltrick“, der erlaubt, wichtige Familien von nichtlinearen Raumtransformationen sehr effektiv berechenbar zu gestalten. Damit wird die lineare Separation gleichwertig zur Trennung durch gekrümmte Trennflächen im untransformierten Raum. Zum Zweiten wurden große Fortschritte in der numerischen Lösung sehr großer Optimierungsprobleme errungen, die auch komplexere Probleme numerisch handhabbar machen (u.a. LMI-Verfahren, siehe z.B. Schölkopf und Smola 2001.

Die vorgenannten Klassifikationsstrukturen dienen neben der Klassifiktionsleistung auch der Generierung von Klassifikationswissen und sind ein gutes Beispiel für den Prozess der *knowledge discovery*, also Gewinnung von Erkenntnissen und explizitem Wissen aus den Daten.

Ein Gegenbeispiel ist ein rein speicherbasierter Klassifikator, der keinerlei Anstrengung der Wissenskondensation unternimmt: Der **k-nächste Nachbarn**-Klassifikator (*k-nearest neighbor*, **k-NN**) sucht für einen neuen Merkmalsvektor $\mathbf{x}_{neu}$ die k ähnlichsten Vertreter der Trainingsmenge und returniert das Klassenlabel, das darunter am häufigsten vorkommt. Das Verfahren gehört zu den so genannten instanz-basierten Lernmethoden.

Zur Steigerung der Klassifikationsleistung eines Klassifikators können

[1] s.a.: http://www.liacc.up.pt/ML/statlog

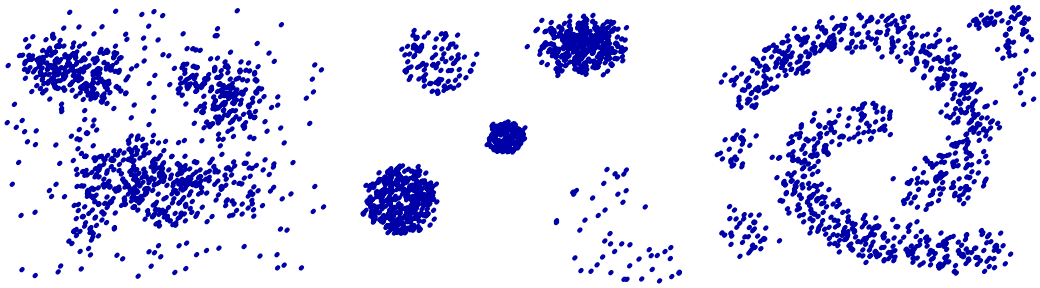


Abbildung 5.4: Beispiele für 2D-Clusterstrukturen mit verschiedenen Charakteristika, wie Größe, Form, Dichte. Sie stellen sehr unterschiedliche Herausforderungen an die jeweilige Herangehensweise. Die räumliche Nähe steht hier für die Ähnlichkeit der Datenobjekte.

mehrere Modelle parallel trainiert werden und das Gesamtergebnis z.B. durch Majoritätsvotum zusammengefasst werden (*voting by committee*). Die Modelle können unterschiedlicher Architektur sein oder auf verschiedenen Datenteilmengen traininert werden (*bagging*).

5.4 Clustermodelle

Clusteranalyse ist die Zerlegung eines Datensatzes in Gruppen (Cluster), so dass die Datenobjekte innerhalb einer Gruppe möglichst ähnlich sind und möglichst ungleich für Datenobjekte aus verschiedenen Gruppen. Dabei kann es um eine nützliche Trennung von kontinuierlich verstreuten Daten gehen (um z.B. aus Messdaten von Kragen-, Brust-, Bauchumfang und Armlänge eine kleine Anzahl von Schnittvorgaben zu bestimmen, vgl. T-Shirts in Konfektionsgrößen S, M, L, XL) oder um Segmentierung mit dem Ziel häufige Zusammenhänge zwischen Merkmalen zu entdecken und die Objekte zusammenzufassen (z.B. um Arten von typischen Kreditkartennutzungsverhalten anhand von Umsatz, Häufigkeit, marktbezogenen und geographischen Merkmalen etc. zu modellieren); Abb. 5.4 illustriert einige Beispiele und Herausforderungen an das Clustern. Es wurde hierfür eine sehr breite Palette von Techniken entwickelt, was damit zusammenhängt, das es keine allgemeingültige Definition von Clustern gibt. Im Gegenteil, die Nützlichkeit eines Clusterergebnisses ist eine durchaus subjektive Größe und hängt auch von den Intentionen des Betrachters ab.

Clusterverfahren lassen sich in folgende Hauptgruppen einteilen: Verfahren zur optimalen Partitionierung, hierarchische Ansätze und probabilis-

tische Beschreibung der Cluster.

Um den Schwerpunkt einer Objektmenge, auch **Centroid** genannt, bilden zu können, muss ein vektorieller Merkmalsraum vorliegen. Die meisten Verfahren setzen dies voraus, um damit einen Repräsentanten $\mathbf{r_k}$ eines Clusters k zu formulieren. Ähnlichkeit und Nähe zwischen Datenobjekten und zu Clusterrepräsentanten wird durch eine Distanzfunktion $dist(\cdot)$ beschrieben, siehe auch Abs. 2.2.

5.4.1 Partitionierende Verfahren

Partitionierende Verfahren zerlegen eine Datenmenge disjunkt in K Cluster. Ein Cluster ist nicht nicht-leer und die Anzahl K ist meist vorgegeben.

Schon eine moderate Anzahl von Daten macht eine vollständige Suche nach der optimalen Clustertrennung unmöglich: Möchte man nur 100 Datenobjekte in nur zwei Cluster aufteilen, gibt es $2^{100} \approx 10^{30}$ Zuordnungsmöglichkeiten. Folglich arbeiten alle praktischen Verfahren iterativ und bergen die übliche Gefahr, statt des globalen nur ein lokales Optimum zu finden.

Ein klassisches Verfahren ist das **k-means**-Clustering (MacQueen 1967). Ausgehend von einer Zufallszuordnung aller Datenobjekte zu den K Clustern werden die K Centroide berechnet. Wiederholt werden dann alle Datenobjekte jeweils demjenigen Cluster zugeordnet, dessen Centroid $\mathbf{r_k}$ sie am nächsten sind, und anschließend wird die r_k erneut berechnet. Das Endergebnis liegt vor, wenn keine Änderung der Clusterzugehörigkeiten mehr erfolgt.

Eine Alternative ist das **k-medoid**- oder *Partitioning-Around-Medoids*-(**PAM**) Verfahren (Kaufman und Rousseeuw 1990), das als Clusterrepräsentanten r_k dasjenige Datenobjekt („medoid“ oder Zentralobjekt) verwendet, dessen Distanzsumme zu den anderen Clustermitgliedern minimal ist (da keine Mittelung erfolgen muss, kann durch die Verfügbarkeit einer Paardistanzmatrix auf die Distanzfunktion $dist(\cdot)$ verzichtet werden). Ng und Han haben einen wesentlich effizienteren, aber auch weniger gründlichen Algorithmus CLARANS vorgeschlagen (z.B. in Han und Kamber 2001).

Cluster können auch als Gebiete im hochdimensionalen Raum angesehen werden, in denen die Objekte dicht beieinander liegen, getrennt

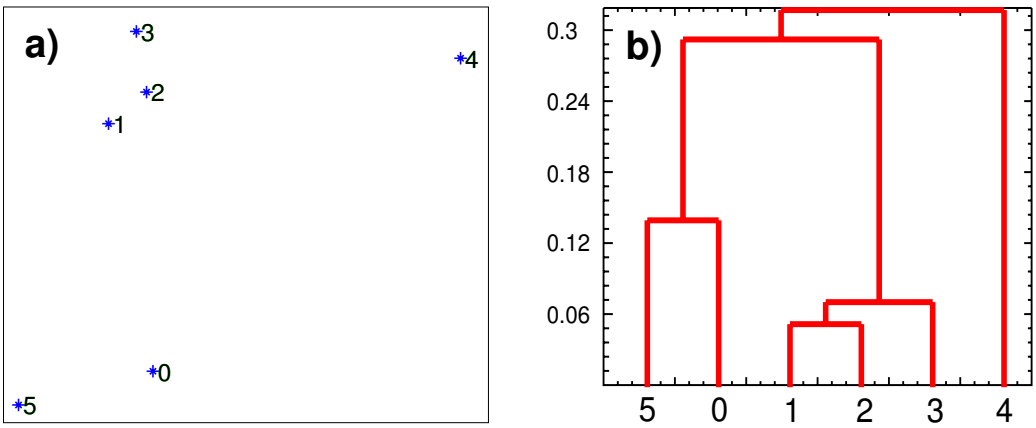


Abbildung 5.5: Beispiel für hierarchisches, agglomeratives Clustern mit der *average linkage* Distanzfunktion (Gl. 5.17) für sechs Punkte in 2D *(a, links)*. *(b, rechts)* Im zugehörigen Dendrogramm kann man die Distanzen auf der Vertikalen erkennen, an denen Subcluster zusammengeführt werden. Die Ziffern in *(a)* und *(b) unten* nummerieren die Objekte. Die Entscheidung über die letztendliche Clusteranzahl bleibt erhalten: Drei Cluster erscheinen hier naheliegend, da hierfür die größte vertikale „Lücke“ im Dendrogramm ersichtlich ist.

durch Gebiete, die von Objekten nur dünn besetzt sind. Die Grundidee für dichtbasierendes Clustern (Ester et al. 1996) gründet auf dem Konzept des Kernobjektes und der Dichte-Verbundenheit. Ein **Kernobjekt** besitzt mindestens M Nachbarn in einer ϵ-Umgebung. Zwei Objekte gelten dann als **dichte-verbunden**, wenn sie durch eine Kette von Kernobjekten verbunden sind, deren Abstand zum Nachbar $\leq \epsilon$ ist. Ein Cluster ist dann eine Menge von dichte-verbundenen Objekten. Im DBSCAN-Algorithmus wird mit den Parametern ϵ und M die Separation von Gebieten, die dünn verbunden sind, gesteuert und damit das Verhalten bei vorhandenem Rauschen bestimmt. Weitere Verfahren werden in Han und Kamber (2001) oder Ester und Sander (2000) erläutert.

5.4.2 Hierarchische Verfahren

Statt einer einzigen Zerlegung des Datensatzes in Cluster erzeugen hierarchische Verfahren eine Hierarchie von möglichen Clusterteilungen, indem sie die Daten in einen Baum, dem *Dendrogramm* repräsentieren, wie in Abb. 5.5 dargestellt. Die Wurzel oben enthält alle Objekte in einem Cluster, die Blätter repräsentieren Einzelobjektcluster (auf der Grundlinie) und

die gezeichnete Höhe der (binären) Verzweigungen repräsentiert den Abstand der beiden Subcluster.

Der Baumaufbau erfolgt *agglomerativ* durch sukzessive Verschmelzung von Clusterpaaren kleinsten Abstands oder umgekehrt, *divisiv*, durch sukzessive Teilung des Gesamtclusters. In beiden Fällen benötigt man die Definition einer Distanzfunktion $dist(C_i, C_j)$ für zwei Objektmengen C_i, C_j, z.B.:

$$\begin{aligned} dist_{single-linkage}(C_i, C_j) &= \min_{x \in C_i, y \in C_j} dist(x, y) \\ dist_{complete-linkage}(C_i, C_j) &= \max_{x \in C_i, y \in C_j} dist(x, y) \\ dist_{average-linkage}(C_i, C_j) &= \frac{1}{|x|\,|Y|} \sum_{x \in C_i, y \in C_j} dist(x, y)\ . \end{aligned} \tag{5.17}$$

Das ***Single-Linkage*-Clustering** entsteht z.B. durch Verwendung des Minimalabstands (*nearest neighbor*) aller möglichen Objektpaarungen und neigt zur Kettenbildung. Das ***Complete-Linkage***- verwendet den größtmöglichen Abstand (*furthest neighbor*) und ***Average-Linkage*-Clustering** nutzt den Mittelwert aller Paarabstände. Nach einer Verschmelzung werden die Abstände zu allen anderen aktuellen Clustern berechnet und im nächsten Schritt wird das Minimum aller ausgewählt.

Die Auswahl der letztlichen Clusterteilung erfolgt meist nach manueller Inspektion der Ergebnisse. Liegt Rauschen vor, können sehr willkürlich erscheinende Clusterverbindungen entstehen, die hinterher nicht einfach korrigierbar sind.

Ein Kombination mit dem oben erläuterten Dichte-Verbundenheits-Konzept liefert das OPTICS-Verfahren (s. z.B. Ester und Sander 2000; Han und Kamber 2001). Für sehr große Datensätze wurde das **BIRCH**-Verfahren entwickelt, das eine stark komprimierte, hierarchische Repräsentation der Daten in so genannten *Clustering Features* (CF) bildet (*Balanced Iterative Reducing and Clustering using Hierarchies* von Zhang et al. 1996).

Werden die Objektzahlen groß, entfällt die Möglichkeit, einige der Algorithmen durch Verarbeitung der Abstandsmatrix im Hauptspeicher zu beschleunigen (N^2 Speicherbedarf). Braucht man den direkten Datenzugriff über die Merkmale, ist die Verwendung von räumlichen (multivariaten) Indexstrukturen in Datenstrukturen oder Datenbanken (*spatial databases* aus der Familie der *R-trees* eine grundsätzliche Beschleunigungsmöglichkeit (s. z.B. Ester et al. 1996) .

5.4.3 Probabilistische modellbasierte Clusterverfahren

Probabilistische Clusterverfahren basieren auf Modellen, die die Wahrscheinlichkeitsdichte der Daten durch ein Mischungsmodell

$$f(\mathbf{x}) = \sum_{i=1}^{K} \pi_k \, f_k(\mathbf{x}, \theta_k) \; , \tag{5.18}$$

bestehend aus K Komponenten anpassen. Zunächst muss K herausgefunden, dann müssen die Einzelkomponentenfunktionen f_k optimiert werden. Häufig dienen dazu multivariate Normalverteilungen (Gl. 4.34), die mit dem EM-Verfahren bestimmt werden.

5.4.4 Weitere neuronale Verfahren: CLM

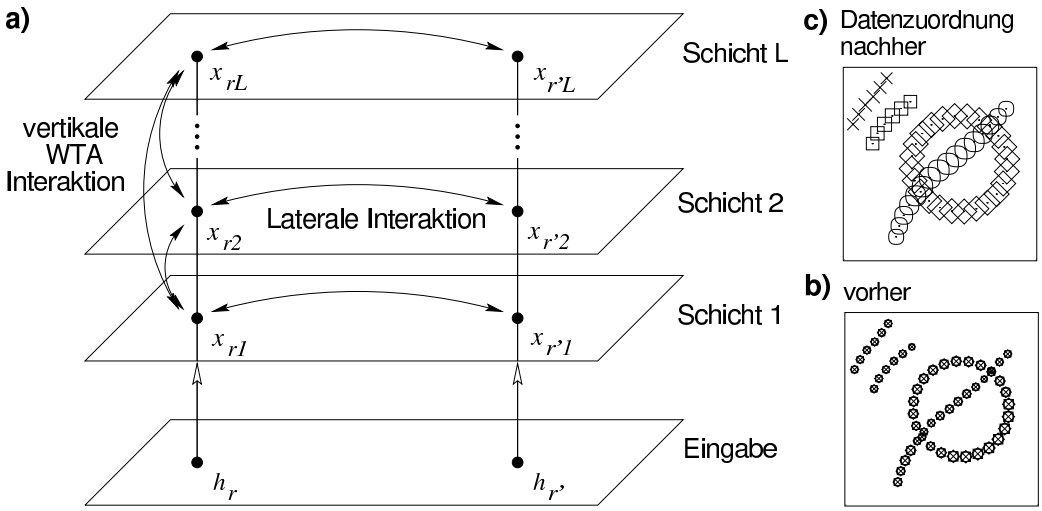


Abbildung 5.6: *(a, links)* Vertikale und laterale Wechselwirkungen bestimmen die Netzwerkdynamik im CLM-Netzwerk. *(Rechts)* Die Zuordnung einer Datenverteilung zu den Schichten vor *(unten, b)* und nach *(oben, c)* Ablauf der Dynamik wird durch die verschiedenen Marker illustriert. Jeder Markertyp ist Indikator für die Zuständigkeit zu einer Schicht. Die Gruppierung in Cluster erfolgt hier nicht nach Nähegesichtspunkten allein, sondern berücksichtigt auch eine „gute“ Fortsetzung von Strukturen.

Biologisch motivierte Ansätze, wie das *Competitive-Layer-Model* (**CLM**) Verfahren, machen sich ein in Schichten aufgebautes, stark vernetztes

neuronales Netzwerk zunutze, um die Daten auf verschiedenen Konzeptebenen abzubilden. Abb. 5.6a illustriert die Eingabeschicht der Daten und darüber L Schichten sowie die Wechselwirkungen der Neurone innerhalb und zwischen den Schichten. Ihre Aktivierung bezüglich der Daten ist so gestaltet, dass via vertikaler Wechselwirkung immer genau eines der L Neuronen aktiv ist (*Winner-Takes-All*, WTA). Die laterale Wechselwirkungsfunktion bestimmt, ob zwei Nachbarn sich positiv verstärken oder negativ inhibitieren. Im Gegensatz zu einfachen univariaten Paardistanzmaßen kann diese Funktion erheblich komplexere Zusammenhänge modellieren, zum Beispiel Wirkmechanismen, die den physiologischen Gestaltwahrnehmungsgesetzen entsprechen. Wie in Abb. 5.6bc gezeigt, erfolgt hier eine Clusterung nicht ausschließlich nach Nähe, sondern nach perzeptuellen Gruppiergesichtspunkten (Wersing, Steil und Ritter 2001).

5.5 Assoziationsregeln

Assoziationsregeln bilden die Essenz der so genannten „Warenkorbanalyse“. Ein Warenkorb kann bestimmte Kundentransaktionen repräsentieren, z.B. das Abrufen von Informationen oder den Kauf von Dienstleistungen oder Waren. Für die Analyse von Datensammlungen aus Scannerkassen einer Ladenkette kann man den Ausdruck Warenkorb wörtlich verstehen. Zusammenhänge, die hier im Vordergrund stehen, sind von der Art „die meisten Kunden, die Baguette und Rotwein kaufen, kaufen auch Käse“. Interessant ist diese Regel natürlich nur, wenn Baguettes und Rotwein genügend häufig gemeinsam gekauft werden. Wissen dieser Art kann zum Beispiel zur Kundensegmentierung oder zur Effizienzoptimierung von Cross-Marketing, Katalog-Designs oder Laden-Layouts verwendet werden.

Die Struktur einer Assoziationsregel hat die Form

$$\begin{aligned} &\text{WENN } A=1 \text{ UND } B=1 \quad \text{DANN} \quad C=1 \text{ mit Wahrscheinlichkeit } c \\ &\text{äquivalent zu:} \qquad \{A=1,\, B=1\} \quad \rightarrow \quad \{C=1\} \end{aligned} \tag{5.19}$$

wobei $A=1, B=1$ und $C=1$ so genannte *items* (Artikel) sind, die als binäre Variablen A, B, C das Vorhandensein eines Merkmals (z.B. Kauf eines Produktes) oder das Zutreffen einer Bedingung (z.B. Kaufpreis in einem bestimmten Intervall) notieren. c ist die bedingte Wahrscheinlichkeit $s_{ABC} = p(C=1|A=1, B=1)$ und wird auch als **Konfidenz** (*confidence, accurancy*) der Regel von Gl. 5.19 bezeichnet. Die Häufigkeit des gemein-

samen Auftreffens $s_{ABC} = p(A = 1, B = 1, C = 1)$, der ***support*** oder Träger der Regel, weist auf die Stärke der Regelbasis hin.

Das Assoziationsproblem ist nun, *alle* Assoziationsregeln zu finden, die eine Mindestkonfidenz c_{min} und einen Mindestsupport s_{min} erreichen oder überschreiten. Die Struktur der Regeln ist sehr einfach, überschaubar und gut interpretierbar. Allerdings ist der Begriff „Regel“ etwas irreführend, da keineswegs ein kausaler Zusammenhang (von links nach rechts in Gl. 5.19) beschrieben, sondern lediglich eine überschwellig starke Merkmalskorrelation festgestellt wird.

5.5.1 Der Apriori-Algorithmus

Das Assoziationsproblem ist ein Beispiel für Datamining-Algorithmen, für die die effektive Suche und die Datenhandhabung die kritischsten Faktoren sind. Als Basisalgorithmus ist der **Apriori**-Algorithmus von Agrawal und Srikant (1994) zu nennen. Er macht sich die Monotonieeigenschaft häufiger *itemsets* zur Abkürzung der systematischen Breitensuche nach Regel zunutze. Ein „häufiges“ k-*itemset* ist eine Kombination von k verschiedenen *items* mit *support* $\geq s_{min}$. Da jede *itemset*-Teilmenge weniger einschränkende Bedingungen (nach *items*) enthält, ist ihr *support* mindestens so groß und damit jede Teilmenge auch häufig. Dies wird bei der sukzessiven Generierung von Kandidaten für häufige k-*itemsets* aus kleineren und (geprüft) häufigen *itemsets* ausgenutzt und führt zu einer Einschränkung von zu testenden Regeln und damit zur großen Beschleunigung des Verfahrens durch Minimierung der Datenbasis-Durchläufe.

5.5.2 Verallgemeinerte und quantitative Assoziationsregeln

In der Praxis ergibt sich das Problem der Wahl der *items* und der Schwellenwerte c_{min}, s_{min}. Je feingranularer man die Artikel bestimmt (z.B. in der natürlichen Ordnungshierarchie Getränk, Wein, Rotwein, Marke, Lage/Jahr) desto geringer wird der *item support*. Senkt man den Minimalsupport c_{min}, findet der Apriori-Algorithmus ggf. eine unüberschaubare Menge von Regeln. Eine Lösung von Srikant und Agrawal (1995) berücksichtigt eine Zugehörigkeitshierarchie von *items* (Taxonomie) und kann damit die relative Interessantheit einer Regel formulieren. Z.B. entfallen in einem

Laden 10 % der Milchkäufe auf Biomilch; damit kann man die Konfidenz und den Support für {Müsli}→{Biomilch} aufgrund der allgemeineren Regel {Müsli}→{Milch} abschätzen. Werden die Erwartungen um einen bestimmten Prozentsatz überschritten, wird die speziellere Regel als „ungewöhnlich interessant“ markiert und ausgegeben. Auf diese Weise lässt sich die Menge an generierten Assoziationsregeln stark reduzieren und die anschließende manuelle Inspektion und Analyse entlasten.

Die Beschränkung auf kategoriale Merkmale wird durch das Verfahren für **quantitative Assoziationsregeln** aufgehoben (Srikant und Agrawal 1996). Dabei werden systematisch Intervallbereiche untersucht und daraus reichere Regelkombinationen generiert, zum Beispiel von der folgenden Art: {Alter zwischen 30 und 39} und {Familienstand verheiratet} → {Anzahl Autos 2}.

5.5.3 *Contrast Set Mining*

Möchte man den Unterschied zwischen zwei (oder mehr) Gruppen analysieren, liefert auch die Kodierung der Gruppe für Assoziationsregeln nicht notwendigerweise eine konsistente Darstellung. Zum einen ist die Schwellenwahl problematisch (s.o.), zum anderen wird nicht nach konstrastierenden Merkmalen gesucht, sondern separat nach Korrelationsmustern für jede Gruppe. Das *Search-and-Testing-for-Understandable-and-Consistent-Contrasts* (STUCCO) Verfahren von Bay und Pazzani (1999) generiert einen Baum von Merkmalskombinationen und sucht systematisch nach konstrastierenden Regeln, die eine unerwartet große Abweichung für die betrachteten Gruppen aufzeigen. Die statistische Signifikanz der Abweichung (χ^2-Test) wird mit der Suchstrategie verknüpft, um zum einen die Fehlerrate (Fehler 1ter Art, s. Abs. 4.6) zu begrenzen und zum anderen die Anzahl Regeln kompakt zu halten.

Zum Beispiel wurden in einem Datensatz von Herzoperationen (s. Kap. 9) vier interessante Kombinationen von Regeln gefunden, die in Zusammenhang stehen mit der Granolozytenkonzentration (einer Untergruppe der weißen Blutkörperchen) und der Häufigkeit für das perioperative Auftreten einen Schlaganfalls (d.h. während oder kurz nach einer Operation). Die entdeckte Kombination {Vorhofflimmern} UND {9.4< Granulozytenkonzentration $\leq$ 20.8} zeigt ein Quotenverhältnis von 23.45 (*odds ratio* mit p<0.0001 mit χ^2-Test) und kann in ein spezielles Risikomodell für die Schlaganfallvorhersage integriert werden (Albert et al. 2003).

5.6 Merkmals- und Modellselektion

5.6.1 Merkmalsselektion

Die Frage nach der Wahl geeigneter Merkmale tritt in vielen hochdimensionalen Problemen auf. Bei der Approximation von Y aus den Variablen $X_1, \ldots, X_m$ etwa kann eine Variable X gänzlich irrelevant sein (z.B. Rauschen) oder andere Variablen sind redundant, da sie im Wesentlichen dasselbe aussagen (z.B. sind die Variablen „Einkommen vor Steuer“ und „Einkommen nach Steuer“ hochkorreliert).

Eine theoretische Fundierung der Merkmalsauswahl liefert das Konzept der Unabhängigkeit zweier Variablen (Gl. 4.8). In einer endlichen Stichprobe kann sie abgeschätzt (Abs. 4.12.1) oder umgekehrt, die Assoziation zwischen kategorialen Variablen mittels der Entropie-basierten *mutual information* bewertet werden (Abs. 4.13.1). Kontinuierliche Variablen können z.B. mittels linearer Regression eingeschätzt werden, siehe Abs. 5.7.1.

Allerdings sind die k besten individuellen Variablen nicht unbedingt die beste Menge von k Variablen. Zur Verdeutlichung sei an das berühmte *parity* Problem zweier XOR-verknüpfter, unabhängiger Binärvariablen erinnert. Hier erscheint das Ergebnis durch jede Einzelvariable rein zufällig (also gar nicht) bestimmt, aber durch beide gemeinsam wird es vollständig beschrieben. Historisch gesehen, hatte dieses Problem großen Einfluss auf die Forschung über neuronale Netze, als die Unzulänglichkeit einfacher Ansätze von Minsky und Pappert (1969) angeprangert wurde. Unter der gleichen Unzulänglichkeit leiden praktisch alle Selektionsmethoden, die Einzelmerkmale beurteilen. Dennoch ist man in der Praxis auf Heuristiken zur Suche (*greedy selection*) angewiesen. Eine stark verbreitete Methode ist die stufenweisen Regression, die unten im Abs. 5.7.1 erläutert wird.

5.6.2 Modellselektion

Abb. 5.7 illustriert eine typische Fehlerkurve in Abhängigkeit der Modellkomplexität. Ist ein endlicher Trainingsdatensatz gegeben, kann man den Trainingsfehler evtl. auf Null bringen, indem man ein genügend flexibles Modell einsetzt. Der *Bias* des Modells, d.h. die Differenz des wahren Wertes zum Erwartungswert seiner Schätzung, ist dann klein, aber die *Vari-*

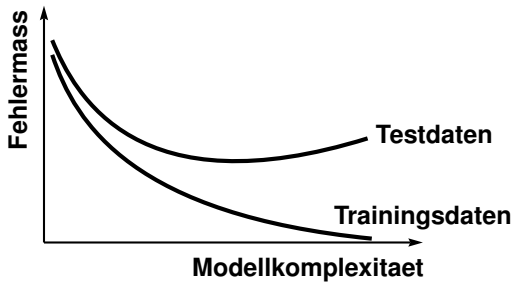


Abbildung 5.7: Ein typisches Diagramm des Missklassifikationsfehlers sowohl für den Test- als auch den Trainingsdatensatz in Abhängigkeit der Modellkomplexität (z.B. der Knotenzahl in einem Entscheidungsbaum).

anz bzgl. anderer Datensätze wird dafür größer. Dies nennt man auch das *Bias-Varianz-Dilemma*. Der entgegengesetzte Fall dieses Zielkonfliktes ist die Wahl eines zu inflexiblen Modells, das einen zu großen *Bias* hat.

Eine Strategie, diesen Kompromiss zu finden, ist der Einbau eines Strafterms für Komplexität in die allgemeine Bewertungsfunktion. Ein Teil misst die Modellgüte (*goodness of fit*) für die Daten und eine Extrakomponente sorgt für einen Bonus für Einfachheit (vgl. *principle of parsimony* oder *Ockham's razor*):

$$\text{Bewertungsfunktion(Modell)} = \text{Fehler(Modell)} + \text{Strafterm(Modell)}.$$

Für eine Reihe von Modellen $\{M_k\}$ mit jeweils m_k Parametern wählt man dann dasjenige, das die Gesamtbewertung optimiert.

Ein verbreiteter Ansatz ist das *Akaikes Information Criterion* (**AIC**), definiert als

$$S_{AIC}(M_k) = -2\,L(M_k) + 2m_k, \tag{5.20}$$

wobei L den Logarithmus der Likelihood notiert (s. Abs. 5.1 und Gl. 5.55) und der Strafterm proportional zur Parameterzahl m_k (Akaike 1973) ist. Eine Alternative ist das *Bayesian Information Criterion* (**BIC**)

$$S_{BIC}(M_k) = -2\,L(M_k) + m_k \log N\ , \tag{5.21}$$

dessen Strafterm auch die Datenanzahl N berücksichtigt. Da die negative Log-Likelihood i.d.R. linear mit N steigt, tritt der Komplexitätsstrafterm mit wachsender Datenzahl in den Hintergrund. Weitere Komplexitätsstrafterme basieren z.B. auf der *Minimum-Description-Length*-Methode (MDL) oder Vapnik's *Structural-Risk-Minimization*-Ansatz (SRM). Unter einer Reihe von Beweisannahmen lassen sich diese Ansätze theoretisch herleiten.

Jedoch sind die Voraussetzungen in der Praxis selten streng erfüllt und so bleibt nur die empirische Wahl (Weiteres s. Ripley 1996; Hand et al. 2001).

5.7 Wichtige Modelle zur Regression

Lineare Regression findet nicht nur zur Korrelationsanalyse und in der Merkmalsselektion Verwendung, sondern gehört auch mit logistischer Regression zu den wichtigsten Approximationsmodellen. Aus diesem Grunde seien sie hier ausführlicher erläutert.

5.7.1 Lineare Regression

In der einfachsten Form beschreibt die lineare Regression eine Verteilung mit (x_i, y_i)-Datenpaaren durch eine gerade Linie. In der *bivariaten Regression* wird ein lineares Modell

$$\hat{y} = \beta_0 + \beta_1 x \tag{5.22}$$

formuliert. Die Variable x heißt im Regressionskontext **„unabhängige Prädiktorvariable“** (auch *independent variable, predictor, regressor*) und y die **„abhängige Antwortvariable“** (*response, criterion*), was aber keinen realen Ursachen-Wirkungs-Zusammenhang impliziert und damit nicht verwechselt werden sollte. β_0 und β_1 heißen die Regressionskoeffizienten, die den Y-Schnittpunkt und die Geradensteigung beschreiben[2]. Sie können durch Minimierung der Standardkostenfunktion (Gl. 5.16), der Summe der quadratischen Fehler

$$QS_{residual} = LOF_{LSE} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \sum_i (\beta_0 + \beta_1 x_i - y_i)^2 \tag{5.23}$$

(*Least-Square-Error*, LSE) direkt und analytisch berechnet werden:

$$\begin{aligned} \beta_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \beta_0 &= \bar{y} - \beta_1 \bar{x} \; . \end{aligned} \tag{5.24}$$

[2] Üblicherweise werden im Kontext der Regression die Koeffizienten als α, β, oder allgemeiner als β_i notiert. Daher wird hier von der in Abs. 5.2 eingeführten Notation abgewichen.

Fehler und Signifikanz

Wie sicher beschreibt das lineare Modell die Daten? Eine (ungefähre) Normalverteilung der Daten vorausgesetzt, geben hierzu die Standardfehler der Parameterschätzungen Auskunft

$$s.e.(\beta_0) = s\sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}} \tag{5.25}$$

$$s.e.(\beta_1) = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} \quad \text{mit} \tag{5.26}$$

$$s = \sqrt{\frac{\sum_i (y_i - \bar{y})^2 - \beta_1^2 \sum_i (x_i - \bar{x})^2}{N-2}} \tag{5.27}$$

s wird gelegentlich auch als "Standardfehler der Regression" bezeichnet.

Erwarten wir eine bestimmte Geradensteigung β_1', kann die Abweichung einer empirische Steigung β_1 auf Signifikanz mit der t-Statistik

$$t = \frac{\beta_1 - \beta_1'}{s.e.(\beta_1)}\ , \quad \nu = N - 2 \text{ Freiheitsgrade} \tag{5.28}$$

und deren kritische Werte überprüft werden (s. Abschnitt 4.7.2). Insbesondere kann man damit erfahren, ob überhaupt ein signifikanter linearer Zusammenhang $y(x)$ besteht, indem man gegen die Nullhypothese $\beta_1' = 0$ testet.

Konfidenzbereich

Bei der Betrachtung des Konfidenzintervalls einer $\hat{y}$-Schätzung ist zu beachten, dass durch die Steigungsunsicherheit das Intervall zunimmt, je weiter der x-Wert von $\bar{x}$ entfernt ist. Der Standardfehler der $\hat{y}$-Schätzung in Gl. 5.22 ist

$$s.e.(\hat{y}(x')) = s\sqrt{\frac{1}{m} + \frac{1}{N} + \frac{(x' - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \tag{5.29}$$

und skaliert mit dem kritischen Wert der t-Statistik mit (N-2) Freiheitsgraden mit dem gewünschten Signifikanzniveau. Für einen konkreten Wert $\mathbf{x}$ ist $m = 1$ (bzw. noch allgemeiner ist m die Anzahl der Messungen, mit denen der neue x'-Wert bestimmt wurde, Zar 1996).

Wie zeichnet man den Bereich möglicher Regressionsgeraden? Hier ist die x'-Bestimmungsunsicherheit nicht wichtig und der $\frac{1}{m}$-Term wird 0 gesetzt. Abb. 5.8 zeigt zwei Regressionsgeraden mit je 95 % Konfidenzregionen, die schmale mit $m = 0$ und die weitere mit $m = 1$ (Albert, Arnrich, Rosendahl, Beller, Walter, Mortasawi, Dalladaku und Ennker 2002).

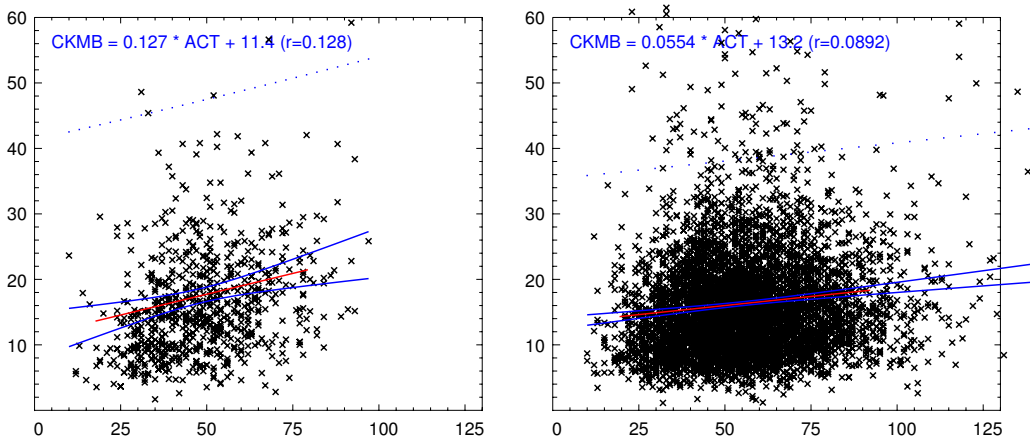


Abbildung 5.8: Beispiel für Konfidenzanalyse in der linearen Regression: zwei Verfahren zur Ruhigstellung des Herzmuskels bei einer Herzoperation mit Herz-Lungen-Maschine: (*a, links:*) für eine so genannte kristalloide Kardioplegie (injizierte Lösung) (*b, rechts:*) für eine Spülung mit Kalium-angereichertem Blut (s. auch Kap. 9). Aufgetragen ist ein spezifischer Schadensindikator für den Herzmuskel, dem CKMB-Blutwert versus der Aortenklemmzeit (*aortic clamp time* ACT mit Stillstand des Herzens). Die Verschiebung der Regressionsgeraden bedeutet einen Vorteil von Variante (*b*), der mit der Dauer des Eingriffs zunimmt. Die Absicherung der Signifikanz erfolgt zusätzlich über die stufenweise, multivariate Regression (s.u.). Hier werden zwei Terme hinzugenommen, *isTypeC, isTypeC*ACT*, die den Kardioplegietyp in dichotomer Kodierung enthalten. Beide Terme erwiesen sich als signifikant. Die inneren blau markierten Konfidenzregionen enthalten mit 95 % Wahrscheinlichkeit die (unter den Modellannahmen) „wahre“ Regressionsgerade – das äußere, durch blaue Punkte markierte Band beschreibt die Konfidenzintervalle für eine Schätzung *CKMB(ACT)*. Im Mittel liegen 5 % der Punkte außerhalb.

Varianzanalyse und Pearson's r

Die Gesamtvariabilität der abhängigen Variablen y wird als Quadratsumme der Mittelwertsabweichungen (*total sum of squares*) berechnet.

$$QS_{total} = \sum_i (y_i - \bar{y})^2 \tag{5.30}$$

Sie lässt sich durch die Regression in zwei Anteile zerlegen:

$$QS_{total} = QS_{regression} + QS_{residual} \tag{5.31}$$

wobei die Variabilität der linearen Regression sich darstellt als

$$QS_{regression} = \sum_i (\hat{y}_i - \bar{y})^2 = \frac{\left[\sum_i (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_i (x_i - \bar{x})^2} = \beta_1 \sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{5.32}$$

und $QS_{residual}$ der Kostenfunktion Gl. 5.23 entspricht.

Setzt man die beiden Größen ins Verhältnis, bekommt man

$$r^2 = \frac{QS_{regression}}{QS_{total}} = \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{[\sum_i (x_i - \bar{x})^2][\sum_i (y_i - \bar{y})^2]}\ , \tag{5.33}$$

den durch die Regression determinierten Anteil der Gesamtvarianz, der gleich dem Quadrat von Pearson's r ist (vgl. Gl. 4.84). r^2 wird daher auch *coefficient of determination* genannt. Zum Beispiel erklärt ein Regressionsmodell mit $r = 0.9$ genau 81 % der Varianz, während 19 % unerklärt bleiben.

Eine weitere geometrische Interpretation ergibt die N-dimensionale Einbettung der Mittelwertabweichung $\vec{X}_i = (x_i - \bar{x})$ und $\vec{Y}_i = (y_i - \bar{y})$. Dann beschreibt

$$r = \cos \angle(\vec{X}, \vec{Y}) = (||\vec{X}||\,||\vec{Y}||)^{-1} \vec{X} \cdot \vec{Y} \tag{5.34}$$

den Differenzwinkel der beiden Abweichungsvektoren.

Verallgemeinerte und multivariate lineare Regression

Das lineare Modell kann auf mehrere (m) unabhängige Regressionsvariablen X_j erweitert werden:

$$\hat{y} = f(X) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j \tag{5.35}$$

Die Variablen können verschiedenen Ursprungs sein:

- X_j sind quantitative Eingangsgrößen;
- Transformationen von quantitativen Eingangsgrößen ($\sqrt{x}, \log x, \ldots$);
- Nominalexpansionen von kategorialen Variablen;
- nichtlineare **Basisfunktionen**: Zum Beispiel kann ein Polynomfit m-ter Ordnung $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^m$ durch die Kodierung $X_j = X_1^j$ formuliert werden;
- Variableninteraktionen, zum Beispiel $X_3 = X_1 \cdot X_2$.

Diese Verallgemeinerung der linearen Regression ist auch unter dem Namen *Generalized Linear Least Square Model* bekannt, wobei sich der Terminus „linear“ hier auf die Linearität der Modellparameter β_j bezieht, auch wenn nichtlineare Basisfunktionen Verwendung finden.

Die N Regressionstrainingsdaten bestehen aus Tupeln $(\mathbf{x}_i, y_i)$, wobei $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$ ein Vektor aus den direkten oder durch Basisfunktionen abgeleiteten Eingangsgrößen ist. Bildet man eine $N \times (m+1)$-Matrix $\mathbf{X}$ mit den Zeilen der Eingabevariablen (mit einer vorangestellten 1) und einen Vektor $\mathbf{y}$ aus den N y_i-Werten, kann der quadratische Fehler der linearen Regression (vgl. Gl. 5.23) kompakt notiert werden:

$$QS_{residual}(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta). \tag{5.36}$$

Dies ist eine quadratische Funktion mit $m + 1$ Parametern. Differenzieren nach β liefert

$$\frac{\partial QS_{residual}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta). \tag{5.37}$$

Ist $\mathbf{X}$ nicht singulär und die Kovarianzmatrix $\mathbf{X}^T\mathbf{X}$ positiv definit, findet sich die eindeutige Lösung der Schätzung $\hat{\beta}$ durch Nullsetzen der ersten Ableitung

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{5.38}$$

Konfidenzbereiche und Signifikanztest

Die Varianz der Parameterschätzung ist

$$Var\{\hat{\beta}\} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \tag{5.39}$$

mit der erwartungstreuen Varianzschätzung

$$\hat{\sigma}^2 = \frac{1}{N-m-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{5.40}$$

Um die Nullhypothese zu testen, die besagt dass ein einzelner Parameter $\beta_j = 0$ ist, wird der Z-Score

$$z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{c_j}} \tag{5.41}$$

gebildet und gegen den kritischen $(1-\alpha)$-Wert der Standardnormalverteilung verglichen. Hierbei bezeichnet c_j das j-te Diagonalelement der Matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.

In ähnlicher Weise kann man den Konfidenzbereich C_β für den ganzen Parametersatz β abschätzen:

$$C_\beta = \left\{\beta \,|\, (\beta-\hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta-\hat{\beta}) \leq \hat{\sigma}^2\chi_{m+1}^{2^{(1-\alpha)}}\right\} \tag{5.42}$$

wobei $\chi_{m+1}^{2^{(1-\alpha)}}$ den kritischen $(1-\alpha)$-Wert der χ^2-Verteilung mit $(m+1)$ Freiheitsgraden ist. Eine Illustration dieses elliptischen Bereiches findet sich in Abb. 5.9.

Für einen neuen Datenpunkt $\mathbf{x}$ stellt sich die Frage, wie man das Konfidenzintervall der $\hat{y}(\mathbf{x})$-Schätzung ermittelt. Im Prinzip ergibt sich das Intervall aus dem Parameterbereich $\{\,(1:\mathbf{x}^T)\beta \,|\, \beta \in C_\beta\}$. Einfacher ist es, das Konfidenzintervall durch den Standardfehler

$$s.e.(\hat{y}(\mathbf{x})) = \sigma\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}} \tag{5.43}$$

und dem kritischen Wert der t-Statistik (mit $N-m-1$ Freiheitsgraden) abzuschätzen.

Sind Variablen X_i linear abhängig, leidet die Matrix unter Rangdefizit. Ein robustes Verfahren zur allgemeinen und robusten Lösung der Matrixgleichung Gl. 5.37 ist das SVD-Verfahren.

Das ***Singular-Value-Decomposition***-Verfahren (**SVD**) ist technisch eine Hauptkomponentenanalyse (**PCA**) und zerlegt eine $N \times M$ große $\mathbf{X}$-Matrix in drei Matrizen

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{5.44}$$

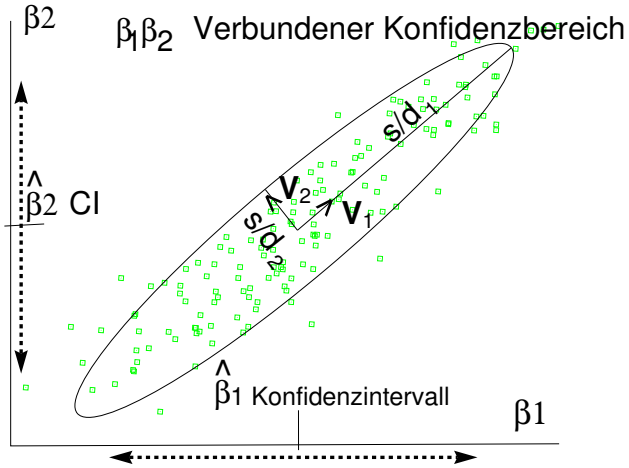


Abbildung 5.9: Die Punktwolke symbolisiert die Verteilung der Modellparameter in Projektion auf β_1 und β_2. Die isolierten Konfidenzintervalle für β_1, β_2 sind durch die gestrichelten Linien gekennzeichnet. Die Ellipse umschreibt den Konfidenzbereich C_β mit $s = \chi^2_{0.95}$ (Gl. 5.42). Die Halbachsenrichtung j ist die j-te Spalte der $\mathbf{V}$-Matrix, skaliert mit s/d_j. In diesem Fall erklären die isolierten Variablen X_1, X_2 unsicherer (große Konfidenzintervalle auf den $\beta_1, \beta 2$-Achsen) als beide gemeinsam bzw. deren Rekodierung (i.e. Projektion auf die V_1-, V_2-Achsen). Angenommen, das Konfidenzintervall von β_1 würde die 0 einschließen, dann würde die univariate Betrachtung X_1 als unsignifikant ablehnen. Für die Merkmalsauswahl vergleicht man daher besser komplette Regressionsmodelle, s. Abs. 5.7.1.

mit besonderen Eigenschaften: $\mathbf{U}$ ist eine $N \times M$ Matrix und spaltenweise orthonormal; $\mathbf{V}^T$ ist von der Größe $M \times M$ und orthonormal; $\mathbf{D}$ ist eine $M \times M$ große Diagonalmatrix mit nicht-negativen Elementen. Wegen der Orthonormalität $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{1}$ ist die inverse Matrix

$$\mathbf{X}^{-1} = \mathbf{V}[\text{diag}(1/d_j)]\mathbf{U}^T. \tag{5.45}$$

$\mathbf{U}$, $\mathbf{D}$, $\mathbf{V}$ sind effektiv (und bis auf Spalten- und Zeilenpermutationen eindeutig) mit der Householdertransformation bestimmbar. Häufig wird die Permutation ermittelt, die die d_j absteigend sortiert enthält. Damit ist Spalte $\mathbf{v}_1$ die so genannte „1. **Hauptachse**“ (*principal axis*). Sie ist die varianzstärkste Richtung in $\mathbf{X}$, $\mathbf{v}_2$ ist die dazu senkrechte Richtung mit der zweitstärksten Varianz usw. Die Transformation $\mathbf{V}^T x$ ist auch als *Karhunen-Loeve*-Verfahren bekannt.

Ein Rangdefizit der Matrix $\mathbf{X}$ wird durch ein (oder mehrere) verschwindende Diagonalelemente $d_j \approx 0$ angezeigt. Setzt man das entsprechende

inverse Diagonalelement von $\mathbf{D}^{-1}$ (statt ∞) zu 0, kann man zeigen, dass man den Vektor $\beta = \mathbf{X}^{-1}\mathbf{y}$ erhält, der optimal im Sinn des kleinsten Fehlerquadrates der Gl. 5.23 ist.

Ein Nebengewinn der Lösung mittels SVD ist das Zusatzwissen um die Korreliertheit der Parameterunsicherheiten. Der Konfidenzbereich ist gemäß Gl. 5.42 von elliptischer Form, die sich an $\mathbf{D}$ und $\mathbf{V}$ direkt ablesen lässt. Sind Variablen korreliert, dann zeigt sich dies in elongierten, gekippten Ellipsen in einem Parameterplot. Abb. 5.9 illustriert die Isolinien für zwei β-Komponenten.

Signifikanztest mit der F-Statistik

Einen äquivalenten Signifikanztest bietet die F-Statistik, die eng mit der ANOVA-Methode verknüpft ist (Abs. 4.8). Häufig möchte man eine Testaussage über den simultanen Einschluss von einer Gruppe von Variablen in die Regression. Dies ist zum Beispiel für die Expansion einer Nominalvariablen mit Kardinalität k oder bei verbundenen Basisfunktionen von Bedeutung. Mit der F-Statistik vergleicht man zwei Modelle, das kleinere mit $m_{small}+1$ Parametern und das größere mit $m_{large}+1$. Die Nullhypothese besagt, dass die $m_{large}-m_{small}$ Parameter im größeren Modell schadlos gleich Null gesetzt werden können. Die F-Statistik misst die Änderung der residualen Quadratsumme pro zusätzlichem Parameter, normiert auf die geschätzte Standardabweichung:

$$F = \frac{\frac{QS_{residual,small}-QS_{residual,large}}{m_{large}-m_{small}}}{\frac{QS_{residual,large}}{N-m_{large}-1}}. \tag{5.46}$$

Der kritische Vergleichswert ist die gewünschte Quantile der F-Verteilung $V^{\mathcal{F}_{m_{large}-m_{small}},N-m_{large}-1}$, die sich mit steigendem N der $V^{\chi^2}{}_{m_{large}-m_{small}}$ Statistik annähert (s.a. Abs. 4.7.5).

Stufenweise Merkmalsselektion

Alle verfügbaren X_j-Variablen in die Regression einzubeziehen, liefert oft eine bescheidene Schätzgenauigkeit für neue Daten. Eine verbreitete Methode ist die Beschränkung auf eine geeignete Teilmenge von Merkmalen. **Stufenweise multivariate Regression** bezeichnet Verfahren, die heuristisch nach den signifikantesten Teilmengen suchen.

Die **vorwärts stufenweise multivariate Regression** startet ohne erklärende Variable und fügt jeweils dasjenige Merkmal hinzu, das den größten Erklärungsgewinn im jeweiligen Iterationsschritt einbringt. Findet sich mittels F-Test keine signifikante Verbesserung mehr, wird gestoppt.

Umgekehrt startet das Verfahren der **rückwärts stufenweise multivariaten Regression** mit allen potentiellen Variablen und schließt sukzessive das am wenigsten signifikanteste Merkmal aus, i.e. die Variable, die mit dem kleinsten F-Wert den geringsten Erklärwert anzeigt.

Hybride stufenweise multivariate Regression: Aufgrund von komplexen Merkmalskorrelationen kann i.d.R. nicht garantiert werden, dass die optimale Merkmalskonfiguration gefunden wird. Eine vollständige Suche nach Kombinationen ist aber nur bei wenigen Merkmalen praktikabel. Eine Verfeinerung der Suchstrategien ist das Zulassen von intermittierenden Wachtums- und Schrumpfungsphasen. Es werden zwei verschiedene Signifikanzniveaus gewählt: zum Beispiel eine Irrtumswahrscheinlichkeit von $\alpha_+ = 5\%$ für einen Parametereinschluss und $\alpha_- = 10\%$ für das Auskoppeln des schwächsten β-Wertes. Dies erzeugt eine Hysterese und verhindert unerwünschtes Oszillieren. Die Suche wird beendet, wenn kein Parameter diese Kriterien mehr erfüllt, spätestens nach einer vorgegebenen Schrittzahl (z.B. Hastie et al. 2001).

Hauptachsenregression

Die Hauptachsenregression (*Principal Component Regression*, **PCR**) führt die Regression nicht in dem von $(1 : \mathbf{x})$ aufgespannten Raum durch, sondern von abgeleiteten, d.h. Karhunen-Loeve-transformierten Größen. Wie in Abb. 5.9 illustriert, kann es bei korrelierten Variablen sehr vorteilhaft sein, die Hauptkomponenten über SVD zu suchen, da sich dann varianzkleinere β_j von varianzgrößeren β_j trennen lassen. Die Kernidee ist also nicht alle, sondern nur die ersten q Hauptkomponenten $\mathbf{z} = \mathbf{V}^T(\mathbf{x}^T)^T$ einzubeziehen. Da die Hauptkomponentenanalyse nicht skaleninvariant ist, werden typischerweise in einem Vorverarbeitungsschritt alle $\{\mathbf{x}_i^T\}$- spaltenweise linear transformiert (zentriert auf Mittelwert 0 und skaliert auf Varianz 1). Die optimale Anzahl q der einbezogenen PC-Achsen kann mittels F-Tests bestimmt werden. Schließt man alle Komponenten ein ($q = p+1$), ist die PCR äquivalent zur normalen linearen Regression.

Parameterregularisierung

Die stufenweise Regression ist ein diskreter Suchprozess, der entscheidet, ob je ein Merkmal einbezogen wird oder nicht. Dabei sind suboptimale Entscheidungen nicht zu vermeiden. Das PCR-Verfahren nutzt nicht nur hinsichtlich Informationen über Korrelationen, sondern ist dabei auch ein diskreter Suchprozess. Eine andere Strategie der Varianzreduzierung ist mehr gradueller Natur. Zusätzlich zur Kostenfunktion in Gl. 5.23 werden Nebenbedingungen gestellt. Diese sind so formuliert, dass etliche varianzerzeugende β_j-Parameter schrumpfen oder gar verschwinden.

Die ***Ridge*-Regression** formuliert einen quadratischen Strafterm in der Minimierungsvorschrift für die Parametersuche

$$\hat{\beta}^{ridge} = \text{argmin}_\beta \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \qquad (5.47)$$

Hier ist $\lambda \geq 0$ ein Komplexitätsparameter, der große $||\beta||$ zunehmend bestraft (Hoerl und Kennard 1970). Die Idee ist analog zum *Weight-Decay*-Verfahren in Neuronalen Netzen. Im β-Parameterbild von Abb. 5.9 kann man die Wirkung erläutern. Richtungen mit hoher β-Varianz möchte man schrumpfen, zugunsten von Richtungen mit kleiner β-Varianz (und reziprok großem d). Man kann zeigen, dass Gl. 5.47 im Hauptkomponentenbild der Matrix $\mathbf{X}$ genau dies bewirkt: Es bringt den Korrekturfaktor $\frac{d_j^2}{d_j^2+\lambda} \leq 1$ in die j-Hauptkomponente ein. Dabei ist $\lambda = 0$ wirkungslos und $\lambda > 0$ schmälert die Projektionswirkung in der gewünschten Weise: und das umso stärker, je kleiner d_j (und damit die β_j-Varianz größer) ist. Den kontinuierlichen Komplexitätsparameter λ kann man durch Kreuzvalidierungsexperimente bestimmen.

In der Regel verschwinden die β-Parameter mit steigendem λ erst spät und nicht nacheinander. Dies ist für die Verbesserung der Interpretierbarkeit des Regressionsmodells also noch kein Gewinn.

Tibshirani (1996) schlug einen modifizierten Strafterm in der ***Lasso*-Regression** vor. Statt der quadratischen Metrik $\lambda \sum_{j=1}^{p} \beta_j^2$ in Gl. 5.47 führt die Absolutmetrik $\lambda \sum_{j=1}^{p} |\beta_j|$ dazu, dass mit wachsendem λ nach und nach β_j-Parameter zu Null werden. Damit verbindet die Lasso-Regression die kontinuierliche Parameterschrumpfung mit dem Interpretierbarkeitsgewinn durch weniger Eingangsterme X_i. Im Vergleichstest schneidet Lasso auch vorteilhaft beim Prädiktionsfehler ab (Hastie et al. 2001, S. 57).

5.7.2 Logistische Regression

Die logistische Regression gehört zur Klasse der verallgemeinerten linearen Modelle und verwendet die **Fermi-Funktion**

$$\text{fermi}(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} \quad \in [0,1] \tag{5.48}$$

um eine Zielfunktion aus dem Wertebereich $[0,1]$ zu modellieren. Damit eignet sie sich besonders um die Eintrittswahrscheinlichkeiten π

$$\pi = \pi(\eta(\mathbf{x})) = \text{fermi}(\eta(\mathbf{x})) \tag{5.49}$$

eines dichotomen Ereignisses aufgrund einer Linearkombination η

$$\eta = \eta(\mathbf{x}) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j = \text{logit}(\pi). \tag{5.50}$$

von m erklärenden („unabhängig“ genannten) Variablen x_i zu beschreiben. Die **logit-Funktion**

$$\text{logit}(\pi) = \ln \frac{\pi}{1-\pi} \tag{5.51}$$

ist genau die Umkehrfunktion der Fermi-Funktion in Gl. 5.48, d.h.

$$\eta = \text{logit}(\pi) \quad \text{und} \quad \pi = \text{fermi}(\eta).$$

Logit ist ferner der Logarithmus des erwarteten Eintrittsverhältnisses „1“ zu „0“ (Quote, *odds*). Der Erwartungswert der dichotomen (in {0,1}-Kodierung) Zielvariable Y ist

$$E\{Y\} = p(Y=1)\cdot 1 + p(Y=0)\cdot 0 = \pi \cdot 1 + (1-\pi)\cdot 0 = \pi. \tag{5.52}$$

Das Datenmodell erlaubt flexibel sowohl metrische als auch kategoriale Merkmale zu kombinieren. Ist die unabhängige (Eingangs-) Variable x_i binär {0,1} kodiert, so ist das Quotenverhältnis (s. Gl. 4.69) für diese Variable

$$\textit{odds ratio} \quad \textit{OR}_i = \frac{\frac{\pi_{x_i=1}}{1-\pi_{x_i=1}}}{\frac{\pi_{x_i=0}}{1-\pi_{x_i=0}}} = \exp\beta_i \ . \tag{5.53}$$

Die logistische Regressionsaufgabe besteht darin, aus den vorliegenden Daten $\{x_{ij}\}$ den Parametervektor $\beta \in \mathbb{R}^{m+1}$ zu ermitteln. Abb. 5.10 zeigt ein $m=1$-dimensionales Beispiel mit zwei zu schätzenden Variablen β_0, β_1.

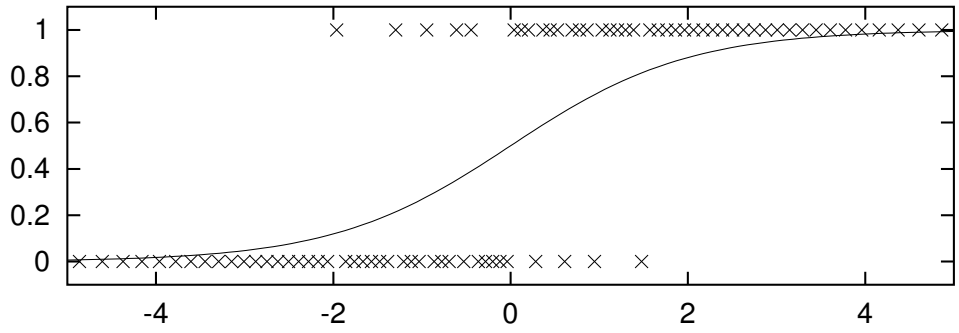


Abbildung 5.10: Logistische Regression: Die Kreuze zeigen 85 Punkte (x_i, y_i) mit x_i als erklärender Variablen und $y_i \in \{0, 1\}$ als binärem Ergebnis. Die Daten deuten an, dass $y = 0$ für $x < 0$ und $y = 1$ für $x > 0$ wahrscheinlicher ist. Die durchgezogene *Fermi*-Kurve zeigt eine damit konsistente Schätzung der Eintrittswahrscheinlichkeit $p(y = 1|x) = \pi(x) = 1/[1 + \exp(-(\beta_0 + \beta_1 x))]$.

5.7.3 Lokale Logistische Regression mittels der Maximum-Likelihood-Methode

Die Notation der folgenden Herleitung ist an zwei Stellen verallgemeinert. *(i)* Insbesondere bei kategorialen Daten können mehrfache Datenbeobachtungen vorliegen (zum Teil mit uneinheitlichem Ergebnis); *(ii)* ferner wird hier eine Gewichtung g_i jedes Datenpunktes eingeführt, die später erlaubt, von einem globalen ($g_i = 1, \ \forall i$) zu einem lokal gewichteten Regressionsmodell überzugehen.

Gegeben seien N Datenpunkte aus $\mathbb{R}^m$, davon seien I verschieden. Mit einer vorangestellten Konstante 1 werden alle Datenpunkte i so zu einer $I \times (m + 1)$ dimensionalen Matrix $\mathbf{X}$ vereinigt, dass die i-te Zeile gleich $\mathbf{X}_i = (1, x_{i1}, x_{i2}, \ldots x_{im})$ ist. Der i-te Datenpunkt sei n_i mal beobachtet und zähle y_i positive Ausgänge (und damit $n_i - y_i$ nicht positive). Die Gesamtzahl der unabhängigen Beobachtungen ist $\sum_{i=1}^{I} n_i = N$. I.d.R. ist für kontinuierliche Variablen $I = N$ mit $n_i = 1$ ($\forall i$).

Die logistische Regressionsaufgabe besteht darin, den Parametervektor $\beta \in \mathbb{R}^{m+1}$ zu ermitteln, der die Likelihood $p(X|\beta)$ gemäß Gl. 5.2 maximiert. Alle Datenpunktschätzungen

$$\pi_i = \pi(\eta_i) = \frac{1}{1 + e^{-\eta_i}} \quad \text{mit} \quad \eta_i = \mathbf{X}_i \beta = \sum_{j=0}^{m} \beta_j x_{ij} \tag{5.54}$$

bilden gemeinsam die Verbundwahrscheinlichkeit der **Likelihood**

$$lik(\beta) = \prod_{i=1}^{I} \left[\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}\right]^{g_i} \tag{5.55}$$

als Produkt der I unabhängigen Binomialdichtefunktionen, potenziert mit den Gewichten g_i. Z.B. hätte damit ein Gewicht $g_i = 2$ dieselbe Wirkung wie das doppelte Auftreten des Datenpunktes i.

Die Likelihood $lik(\beta)$ wird leichter handhabbar durch Logarithmierung, was zur **Log-Likelihood**

$$L(\beta) = \log(lik(\beta)) = \sum_{i=1}^{I} g_i \left[y_i \ln \pi_i + (n_i - y_i)\ln(1-\pi_i)\right] \tag{5.56}$$

und mit Gl. 5.51 zu

$$L(\beta) = \sum_{i=1}^{I} g_i \left[y_i \eta_i + n_i \ln(1-\pi_i)\right] \tag{5.57}$$

führt. Die Maximierung von $L(\beta)$ erfolgt durch Nullstellensuche in den partiellen Ableitungen

$$\frac{\partial L}{\partial \beta_a} = \sum_{i=1}^{I} g_i(y_i - n_i\pi_i)x_{ia} \quad \text{mit} \quad a \in \{0, 1, \ldots p\}. \tag{5.58}$$

Da sich die zweite Ableitung, die **Hessematrix**

$$\frac{\partial^2 L}{\partial \beta_a \partial \beta_b} = -\sum_{i=1}^{I} g_i n_i \pi_i (1-\pi_i) x_{ia} x_{ib} \tag{5.59}$$

recht einfach darstellen lässt, können Optimierungverfahren zweiter Ordnung effizient eingesetzt werden.

Das *Expectation-Maximization*-Verfahren

Bei genauer Betrachtung zeigt sich, dass das Problem, die unbekannten β-Parameter zu schätzen, verwoben ist mit der Maximierung der Likelihood. Eine vielseitig Lösungsstrategie ist das ***Expectation-Maximization***-Verfahren (**EM**-Verfahren) (Dempster et al. 1977). Hierbei nähert man sich iterativ der Lösung, indem zwei Schritte alternieren:

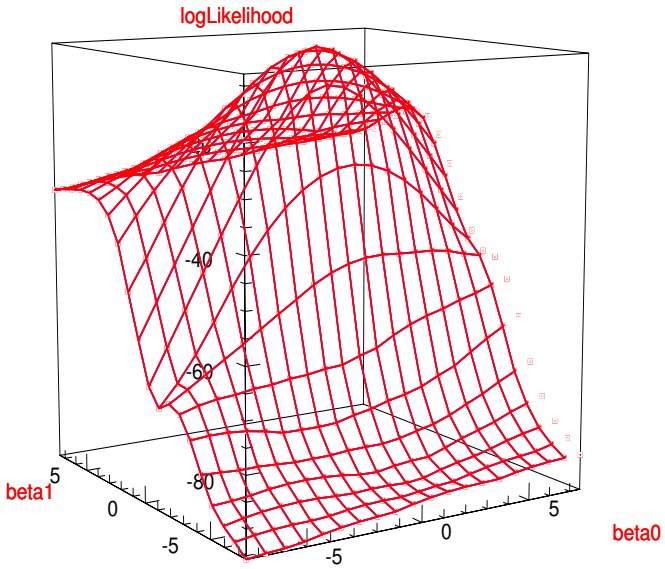


Abbildung 5.11: Die Log-Likelihoodfunktion $L(\beta)$ über der β_0, β_1 Landschaft eines $m = 1$ dimensionalen Problems, s. Abb. 5.10. Das Maximieren von L bezüglich des β-Vektors liefert die Lösung der logistischen Regression. Die Form des Maximums trägt Informationen über die Güte der Parameterschätzung (s. Varianzschätzung Gl. 5.67).

E-step: Im E-Schritt wird die Erwartungschätzung erneuert, d.h. anhand des besten β-Vektors werden Gl. 5.54 und 5.56 ausgewertet.

M-step Die Maximierung der Likelihood wird M-Schritt genannt.

Der M-Schritt nach dem Newton-Raphson-Verfahren verbessert $\beta^{(t)}$, ausgehend von einem Startvektor $\beta^{(t=0)}$, wie folgt:

$$\beta^{(t+1)} = \beta^{(t)} - (\nabla_\beta \nabla_\beta L^{(t)})^{-1} \nabla_\beta L^{(t)}. \tag{5.60}$$

Dies lässt sich in Matrixschreibweise formulieren:

$$\beta^{(t+1)} = \beta^{(t)} + (\mathbf{X}\mathbf{G}\mathbf{A}^{(t)}\mathbf{X}^T)^{-1}\,\mathbf{X}^T\mathbf{G}(\mathbf{y} - \hat{\mathbf{y}}^{(t)}) \quad \text{mit} \tag{5.61}$$

$$\hat{y}_i^{(t)} = n_i \pi_i^{(t)}\ ,\quad i \in \{1, \ldots, I\}, \tag{5.62}$$

$$\mathbf{G} = \text{diag}[g_i] \quad \text{und} \tag{5.63}$$

$$\mathbf{A}^{(t)} = \text{diag}[n_i \pi_i^{(t)}(1 - \pi_i^{(t)})] \tag{5.64}$$

Die oberen Indizes $^{(t)}$ markieren, welche Größen nach jedem Iterationsschritt erneut zu berechnen sind, denn mit $\beta^{(t)}$ verändern sich ja die erwarteten Wahrscheinlichkeiten $\pi_i^{(t)}$ und die Likelihood $L^{(t)}$.

Levenberg-Marquardt für logistische Regression

Die Abb. 5.11 zeigt die Log-Likelihoodlandschaft über die Parameter β_0, β_1. Hierbei wird deutlich, dass eine lokale, elliptische Formanpassung nicht überall das Maximum gut annähert. Um möglichen Konvergenzproblemen zu begegnen, gibt es mehrere Möglichkeiten. Eine besonders robuste, die gleichzeitig mögliche Probleme mit der Invertierbarkeit der Hessematrix elegant löst, wurde von Levenberg und Marquardt (1963) vorgeschlagen. Es ist eine graduelle Mischung aus Verfahren erster und zweiter Ordnung. Die Kernänderung dabei ist, die Gl. 5.60 durch eine nicht-negative Diagonalbeimischung zur Hessematrix $\nabla_\beta \nabla_\beta L^{(t)} + \lambda \mathbf{I}$ zu ergänzen. Der Algorithmus startet mit einem kleinen Levenberg-Marquardt-Parameter λ

1. Start mit $t = 0,\ \lambda = 0.001$ und $\beta^{(0)}$. Dann Iteration:
2. E-Schritt: Berechne $\eta_i, \pi_i, \nabla_\beta L^{(t)}, \nabla_\beta \nabla_\beta L^{(t)}$ mit Gl. 5.58 und 5.59
3. M-Schritt:
$$\beta^{(t+1)} = \beta^{(t)} - (\nabla_\beta \nabla_\beta L^{(t)} + \lambda \mathbf{I})^{-1} \nabla_\beta L^{(t)} \tag{5.65}$$
4. IF $L(\beta^{(t+1)}) < L(\beta^{(t)})$ THEN $\lambda := 10\lambda$ und GOTO 3
5. IF(Abbruchkriterium) THEN STOP
6. $\lambda := \lambda/10$ und $t := t + 1$ und GOTO 2

Die Iteration wird abgebrochen, wenn die Konvergenz festgestellt wird, i.e. entweder

- die Parameteränderungen klein werden: $||\beta^{(t+1)} - \beta^{(t)}|| < \epsilon_1$,
- die Likelihood-Änderung klein wird: $(L^{(t+1)} - L^{(t)})/L^{(t)} < \epsilon_2$ oder
- die maximale Iterationszahl überschritten wird: $t > t_{max}$;
- alle erwarteten Werte entweder 0 oder 1 sind,
d.h. $\forall i: \quad \pi_i(1 - \pi_i) < \epsilon_3 \approx 10^{-8}$.

Mit kleinem Levenberg-Marquardt-Parameter λ verhält sich der Algorithmus wie das Newton-Raphson-Verfahren. Bei erfolgreicher Extremalisierung wird der Schritt akzeptiert ($t := t+1$) und λ noch weiter verkleinert. Wenn es zu Verschlechterungen kommt, wird der Schritt wiederholt und λ zügig vergrößert. Mit wachsendem λ wird die Matrix diagonal dominiert und das Verfahren wird zum *steepest-descent*-Verfahren, das sich in gleicher Weise robust bei einer schlecht konditionierten Hessematrix verhält (s. a. Press et al. 1988).

Varianzschätzung und Konfidenzintervalle

Im Limes großer Datenpunktanzahlen N konvergiert die negative Hessematrix (auch Informationsmatrix genannt) gegen die Kovarianzmatrix der Datendichte. Nachdem die ML-Schätzung $\hat{\beta}$ vorliegt, ergibt sich daraus die asymptotische Kovarianzmatrix der $\hat{\beta}$-Verteilung

$$\hat{\mathbf{C}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{A}} \mathbf{G} \mathbf{X})^{-1} \; . \tag{5.66}$$

Der Standardfehler der einzelnen Parameterschätzung $\hat{\beta}_j$ ist gleich der Quadratwurzel des j-ten Diagonalelementes der Matrix $\hat{\mathbf{C}}(\hat{\beta})$. Anhand der Standardnormalverteilungsperzentilen $Q^{\mathcal{N}}_{1-\alpha/2}$ werden daraus Konfidenzintervalle für die $\hat{\beta}_j$ bestimmbar (Agresti 1990).

Für einen konkreten Datenpunkt $\mathbf{x}_q$ kann das Konfidenzintervall CI der Eintrittswahrscheinlichkeit $\pi(\mathbf{x}_q)$ ermittelt werden. Hierzu werden die Intervallendpunkte mit $\hat{L} = \mathbf{x}_q \hat{\beta}$ und $\sigma^2(\hat{L}) = \mathbf{x}_q^T \hat{\mathbf{C}}(\hat{\beta}) \mathbf{x}_q$ auf der Logit-Skala zu Wahrscheinlichkeiten transformiert:

$$CI_{(1-\alpha)}(\pi(\mathbf{x}_q)) = \left[1 + \exp\left(-\mathbf{x}_q \hat{\beta} \mp Q^{\mathcal{N}}_{1-\alpha/2} \sqrt{\mathbf{x}_q^T \hat{\mathbf{C}}(\hat{\beta}) \mathbf{x}_q} \right) \right]^{-1} . \tag{5.67}$$

Modellvergleich mit dem Likelihood-Ratio- und dem Wald-Test

Man kann zeigen, dass sich die Differenz der Log-Likelihood-Statistiken zweier Modelle M_1, M_2 (s. Gl. 5.56), von dem das erste ein Spezialfall des zweiten ist, sich (für große Fallzahlen) näherungsweise wie die χ^2 Statistik verhält (Kleinbaum 1994). Daraus ergibt sich die *Likelihood-Ratio* oder **LR-Statistik**

$$LR = -2 \ln \frac{lik(M_1)}{lik(M_2)} = 2 \left(L\left(M_2\right) - L\left(M_1\right) \right) \; . \tag{5.68}$$

Die Nullhypothese ist die Zufälligkeit des Modellunterschiedes, d.h. $LR \approx \chi^2_{d.f.}$. Die Anzahl der Freiheitsgrade $\nu = d.f.$ für die χ^2-Vergleichsstatistik wird gleich der Differenz der Parameterzahl der beiden Modelle gewählt.

Ein anderer Weg des Hypothesentests ist unter dem Namen **Wald-Test** bekannt. Er eignet sich insbesondere um Modellunterschiede bzgl. eines Parameters zu testen. Die Nullhypothese H_0 ist, dass ein Parameter i einer logistischen Regression nicht zuträglich ist, d.h. $\beta_i = 0$. Für große Stichproben verhält sich

$$Z_{Wald} \approx \mathcal{N}(0,1) \quad \text{oder} \quad Z^2_{Wald} \approx \chi^2_{d.f.=1} \quad \text{mit} \quad Z_{Wald} = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)} \tag{5.69}$$

unter der Nullhypothese H_0. Zusätzlich nähern sich mit wachsender Stichprobengröße Z^2_{Wald} und LR an. Der Wald-Test ist einfach auswertbar. Die Tab. 5.1 listet ein logistisches Regressionssergebnis inklusive des Standardfehlers (s.e.) von β, Z_{Wald}, den ermittelten p-Wert und die *odds-ratio*.

Variable	**Ergebnis** β	**s.e.** (β_i)	Z^2_{Wald}	**p Wert**	**OR** e^{β_i}
Konstante	β_0= -4,542	0,280	263,62	0,000	
Kritischer präop. Zustand	β_{CRIT}= 1,322	0,253	27,35	0,000	3,752
Chronische Lungenkrank.	β_{PULM}= 1,256	0,237	28,10	0,000	3,511
Creatinine Clearance <55	$\beta_{CC<55}$= 0,908	0,193	22,23	0,000	2,481
Herzauswurfvolumen niedrig	β_{LVEF}= 0,596	0,170	12,27	0,000	1,816
Pulmonaler Bluthochdruck	β_{SPre}= 0,667	0,225	8,77	0,003	1,948
Notfall-Op	β_{EMER}= 1,076	0,382	7,95	0,005	2,933
Alterskodierung	β_{AGE}= 0,156	0,059	6,98	0,008	1,169
...					

Tabelle 5.1: Ergebnisse der logistischen Regression für das Versterbensrisiko bei herzchirurgischen Operationen unter Einschluss der Merkmale des EuroSCORE-Risikobewertungssystems, das in Abs. 9.1.2 beschrieben wird.

5.7.4 Integrales Gütemaß für dichotome Merkmalsschätzer: ROC-Analyse

Modelle, die das Auftreten eines dichotomen Merkmals schätzen, sind in vielen Anwendungen bedeutsam. Zum Beispiel kann ein Risikomodell

Schlaganfall oder Versterben im Umfeld eines chirurgischen Eingriffes prognostizieren (s. Kap. 9) oder eine Alarmanlage entscheidet anhand von Sensorsignalen, ob eine Person in einen Überwachungsbereich eindringt oder nicht. I.d.R. wird in solchen Systemen intern eine ordinale oder kontinuierliche Größe entwickelt, die monoton mit der in Frage stehenden Eintrittswahrscheinlichkeit zusammenhängt. Sie kann letztendlich mit einem Schwellenwert λ verglichen werden, um z.B. Alarm auszulösen oder von einer Operation Abstand zu nehmen.

Einschätzung: \ **Goldstandard:**	definitiv normal **x=1**	wahrs. normal **x=2**	fraglich **x=3**	wahrsch. abnorm **x=4**	definitiv abnorm **x=5**
Normal $(-)$	33	6	6	11	2
Abnormal $(+)$	3	2	2	11	33
Total	36	8	8	22	35

Tabelle 5.2: Zahlenbeispiel für ärtzliche Befundung von 109 CT-Bildern von 58 gesunden (Normal) und 51 kranken Patienten mit Klassifikation in 5 Kategorien.

Dichotome Klassifikatoren (oder allgemein Merkmalsschätzer) mit einem solchen internen Parameter λ können mit der ***Receiver Operating Curve* (ROC)** Analyse in einer integrierten Form evaluiert werden. Im Falle der logistischen Regression wird der Erwartungswert nach Gl. 5.52 zum Vergleich herangezogen. Es können aber auch ordinale Schätzungen sein: Tab. 5.2 zeigt ein Zahlenbeispiel für die ärztliche Beurteilung von Gewebe nach Röntgen-CT-Bildern. Der „interne“ Parameter ist hier die ordinale Klassifizierung in fünf Stufen $x \in \{1, 2, 3, 4, 5\}$ von „definitiv normal“ über „fraglich“ bis „definitiv krank“. Legt man die Entscheidungsschwelle auf λ=1.5, 2.5 oder 3.5?

Grundlage des Vergleichs ist das Wissen um die wahre Klasse bzw. eine Referenzklassifikation, des „Goldstandards“. Ändert sich die Schwelle λ effektiv, ergibt sich eine Verlagerung innerhalb der 2×2-Kontingenztabelle oder „Konfusionsmatrix“ (Tab. 5.3), die sich durch die Zusammenfassung der Fälle in der Tabelle 5.2 kleiner und größer λ darstellt. Abb. 5.12 illustriert den prinzipiellen Zusammenhang graphisch.

Das ROC-Diagramm zeichnet genau diese schätzerspezifische Abhängigkeit mittels der so genannten **Sensitivität** (s.u.) versus der so genannten **Spezifität** für alle λ-Wert auf (d.h. per Konvention 1-Spezifität, s. Abb. 5.7.4). Die Fläche Φ unter dem Linienzug wird vermessen und gibt einen einzigen

aussagekräftigen Wert, der ein integrales Maß für die Prädiktionsgüte des Schätzers darstellt.

Sensitivität und Spezifität sind anhand der vier möglichen Fälle definiert, die in der Entscheidungstabelle 5.3 präzisiert werden:

$$\begin{array}{rcll}
\text{Sensitivität} & := & \frac{\text{TP}}{\text{Goldstandard positiv}} = \frac{\text{TP}}{\text{TP + FN}} & = \frac{a}{a+c} \\
\text{Spezifität} & := & \frac{\text{TN}}{\text{Goldstandard negativ}} = \frac{\text{TN}}{\text{FP + TN}} & = \frac{d}{b+d} \\
\textit{Accuracy} & := & \frac{\text{Einschätzung richtig}}{\text{alle}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} & = \frac{a+d}{a+b+c+d} \\
\textit{Precision} & := & \frac{\text{TP}}{\text{Einschätzung positiv}} = \frac{\text{TP}}{\text{TP + FP}} & = \frac{a}{a+b} \\
\textit{Recall} & := & \text{Sensitivität} & = \frac{a}{a+c}
\end{array}$$

Im Bereich des Text-Mining finden die verwandten Begriffe *precision,*

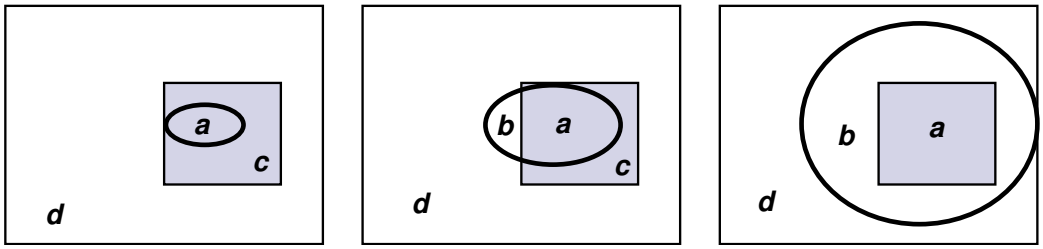


Abbildung 5.12: Klassifikation in Abhängigkeit des gewählten Schwellenwertes λ: das äußere Rechteck steht für alle Fälle (a+b+c+d), das innere für die positiven Fälle (a+c). Durch die Wahl des binarisierenden Schwellenwertes λ, ändert sich die Entscheidungsgrenze des Klassifikators, hier durch die wachsende Ellipse illustriert. (*Links*) Spezifität=1: Nur ein Teil der positiven Fälle wird richtig klassifiziert (a=TP), etliche bleiben unerkannt (c=FN) – dafür gibt es keine falsch positiven (b=FP=0) und die Spezifität ist maximal (=1). (*Mitte*) λ mittel: Mehr Fälle werden positiv entschieden, manche aber auch falsch (b=FP>0). (*Rechts*) Sensitivität=1: Der Klassifikator ist jetzt maximal sensibel und erkennt alle positiven Fälle, allerdings auch zahlreiche irrtümlich (b=FP, hier c=FN=0). Welcher λ-Wert macht den Klassifikator am empfindlichsten? Vergleicht man die Fälle links und rechts, wird die Unzulänglichkeit des umgangssprachlichen Begriffs „Empfindlichkeit“ deutlich, denn beide Extremfälle sind in ihrem jeweiligen Sinne hochempfindlich. Die Fläche unter der ROC-Kurve ist das richtige Maß, um die Diskriminationsgüte zu bewerten. Hier würde sich eine Verbesserung durch eine Adaptation der Ellipsenform an das Rechteck ausdrücken.

	positive Einschätzung	**negative Einschätzung**
Goldstandard positiv (hier abnorm)	korrekte positive Entscheidung (a) TP (*true positive*)	inkorrekte negative Entscheidung (c) FN (*false negative*)
Goldstandard negativ (hier normal)	inkorrekte positive Entscheidung (b) FP (*false positve*)	korrekte negative Entscheidung (d) TN (*true negative*)

Tabelle 5.3: Entscheidungstabelle eines binären Tests mit Definitionen von üblichen Bezeichnungen (TP,TN,FN,FP) und (a,b,c,d).

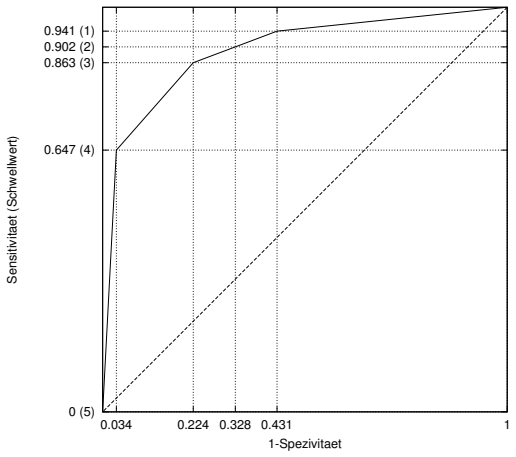


Abbildung 5.13: *Receiver Operating Characteristic* (ROC) Kurve für das in Tabelle 5.2 dargestelle einfache Zahlenbeispiel. Jeder Vertex des Linienzuges entspricht einem effektiven Schwellenwert. Die Diagonale zeigt die Vergleichskurve für einen Zufallsklassifikator mit der Fläche $\frac{1}{2}$.

recall und *accuracy* Verwendung, s. Abs. 8.5. Hierbei ist zu beachten, dass die Begriffe auf eine konkrete „positiv"-Definition bezogen und nicht vertauschbar sind. Die Fläche unter der ROC-Kurve Φ wird i.d.R. als Summe von Trapezflächen unter dem Polygonzug (Abb. 5.7.4) berechnet und ist als Integralmaß invariant gegen Richtungstausch (positiv–negativ).

Die Fläche $\Phi \leq 1$ spiegelt die insgesamte Leistungsfähigkeit des Klassifikators in allen möglichen Arbeitspunkten λ wieder. Mit einem unwissenden Zufallsgenerator, der die Klassenzugehörigkeit auswürfelt, wird eine Vergleichs-ROC-Fläche von $\Phi = 0.5$ erzielt (s. Abb. 5.7.4; ROC-Werte $\Phi < 0.5$ sollten daher nicht vorkommen, sie deuten auf eine falsche Klassendefinition hin). Konstruktionsbedingt steigt die ROC-Kurve stets monoton von (0,0) auf (1,1) (siehe korrespondierende Extremfälle in Abb. 5.12). Je weiter sich die Kurve dem Eckpunkt (0,1) und damit die Fläche Φ der 1 annähert, desto mehr Vertrauen darf man dem Klassifikator schenken.

Die praktische Wahl des Schwellenwertes λ hängt von der gewünsch-

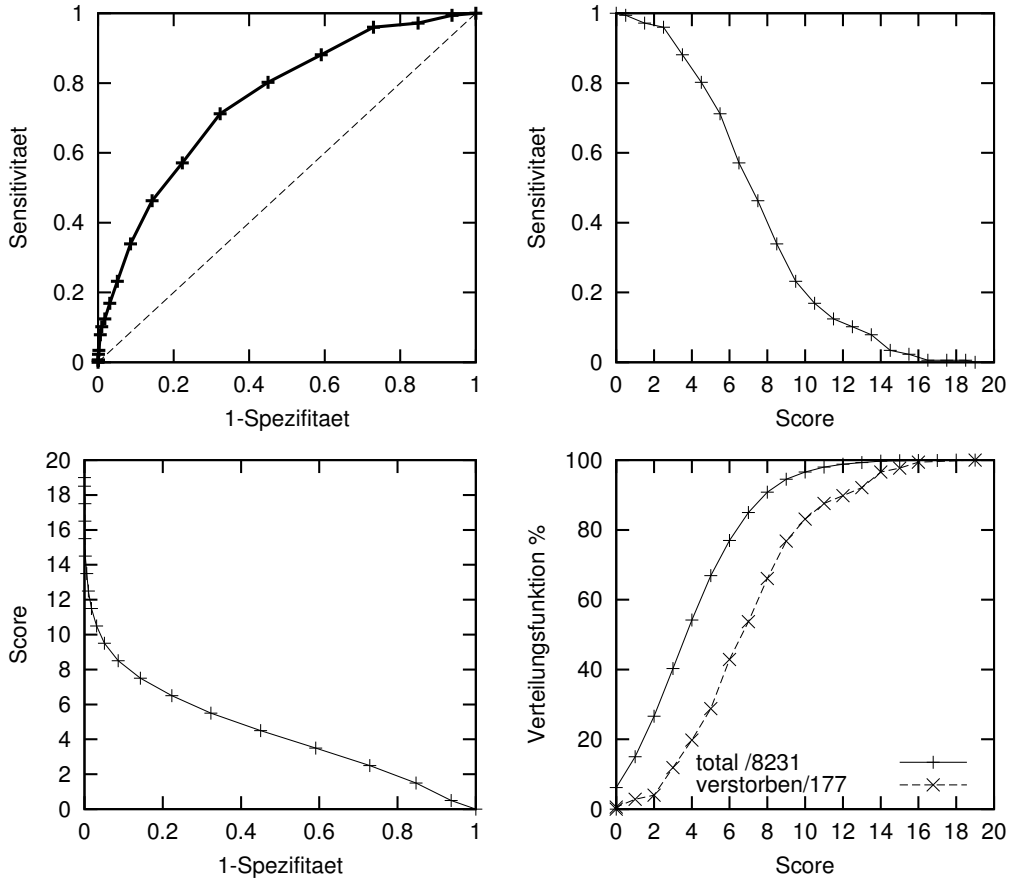


Abbildung 5.14: *(a, links oben:) Receiver-Operating-Characteristics-* (ROC)-Kurve für das EuroSCORE-Risikoschätzmodell (s. Abs. 9.1.2). Die Standardform schätzt das Operationsrisiko eines herzchirurgischen Patienten auf einer ganzzahligen Skala (*Score*-Wert). Zur ROC-Kurve *(links oben)* sind die korrespondierenden effektiven Schwellenwerte λ (i.e. hier halbzahlig) in den benachbarten Graphiken *(b) rechts* und *(c) unten* ablesbar. *(d) rechts unten* sind die akkumulierten Verteilungsfunktionen des *Score*-Wertes für die betrachtete Gesamtpopulation $V^{alle}(Score)$ und für die Untergruppe der Verstorbenen $V^{verst.}(Score)$ aufgetragen. In diesen vier Graphiken sind alle wichtigen Entscheidungsdetails des betrachteten Risikoklassifizierungssystems integriert präsentiert. Die ultimative Verdichtung ist die Angabe der Fläche unter der ROC-Kurve ($\Phi = 0.753$), die die Prädiktionsqualität des Klassifikators auf eine einzige Zahl abbildet.

ten Anwendung ab, genauer gesagt von der relativen Bewertung einer möglichen Fehlklassifikation vom Typ FP und vom Typ FN. Sind sie beispielsweise äquivalent, findet sich der optimale Schwellenwert durch Anlegen einer 45°-Geraden an die ROC-Kurve. Der extremste Kontaktpunkt ist mit der optimalen Schwelle λ assoziiert.

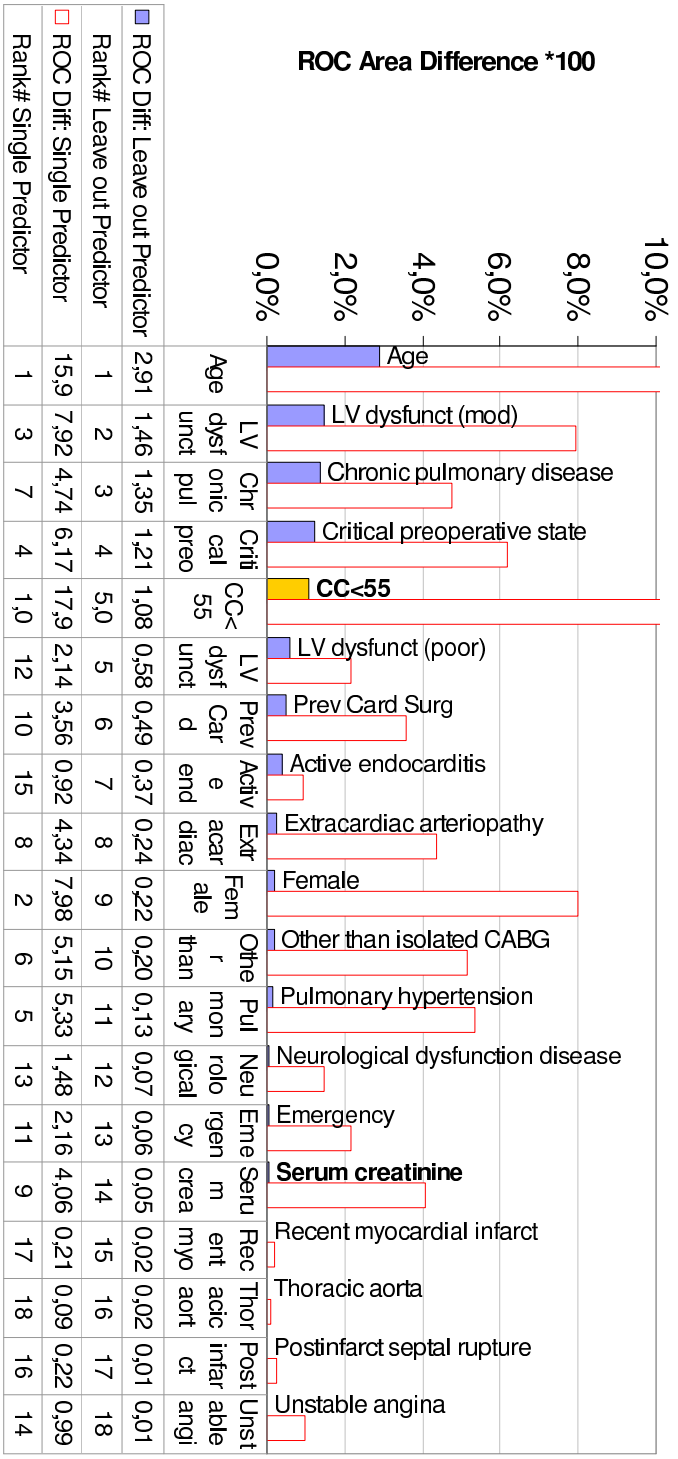
Abb. 5.14 zeigt als weiteres Beispiel die ROC-Kurve für ein Risikomodell. Durch die horizontal und vertikal assoziierten Graphiken wird die jeweilige Korrespondenz zur Wahl einer Binärgrenze λ dargestellt. Die bedingten Verteilungsfunktionen Abb. 5.14d komplettieren das Bild über das Schätzervermögen.

Man kann zeigen, dass die Fläche unter der ROC-Kurve Φ sich als Erwartungswert der Konkordanz interpretieren lässt: D.h. Φ ist die Wahrscheinlichkeit, dass der „interne“ Parameter eines zufällig gewählten positiven Falles höher ist als der eines zufällig gewählten negativen Falles. Würde man z.B. annehmen, von zwei Patienten stirbt eher der Ältere, liegt man bei zufälliger Wahl einer zweielementigen Stichprobe, die genau einen Verstorben enthält, zu 65,9 % richtig (50 % wäre natürlich trivial, s. Abb. 5.7.4).

Konfidenzintervalle für die ROC-Fläche Φ können über die Wilcoxon-*Matched-Pairs-Signed-Ranks*-Teststatistik (s. Abs. 4.12.3) bestimmt werden (Hanley und McNeil 1982). Allerdings sind bei Vergleich zweier Modell anhand ihrer Φ Differenzen und dem s.e.(Φ) Vorsicht geboten (s.u.).

Um die Bedeutung von einzelnen Merkmalen in einer multivariaten Regression zu analysieren, können auch differenzielle ROC-Analysen durchgeführt werden. Mit dieser Technik kann man zum Beispiel untersuchen, wie sich das so genannte EuroSCORE-Risikobewertungssystem verbessert lässt (s. Abs. 9.1.2), indem durch Austausch eines Prädiktormerkmals die Nierenfunktion besser erfasst wird. Da dieser Risikoscore im Bereich der Herzchirurgie weit verbreitet ist (s. Abs. 9.1.2), bedarf ein solcher Verbesserungsvorschlag genauer Analyse.

Abb. 5.7.4 tabelliert und visualisiert die Ergebnisse der differenziellen ROC-Analyse. Die logistische Regression mit den 18 Standardwerten liefert Φ=0.776 (s.e. 0.018). Tauscht man hier den *Serum-Creatinin*-Blutwert gegen einen abgeleiteten Wert, i.e. den binarisierten *Creatinin Clearance* (CC) (Nierenfiltrationsrate <55 ml/min), so verbessert sich die ROC-Fläche um 0.0108 auf 0.787. Wie wichtig der Beitrag zur Prädiktionsgüte ist, kann zum einen *(i)* anhand der ROC-Fläche für jedes isolierte Merkmal bestimmt; zum anderen *(ii)* kann jedes Merkmal einzeln ausge-



	Age	LV dysfunct	Chronic pulmonary	Critical preoperative	CC<55	LV dysfunct	Prev Card Surg	Active endocarditis	Extracardiac arteriopathy	Female	Other than isolated CABG	Pulmonary hypertension	Neurological dysfunction	Emergency	Serum creatinine	Recent myocardial infarct	Thoracic aorta	Postinfarct septal rupture	Unstable angina
ROC Diff: Leave out Predictor	2,91	1,46	1,35	1,21	1,08	0,58	0,49	0,37	0,24	0,22	0,20	0,13	0,07	0,06	0,05	0,02	0,02	0,01	0,01
Rank# Leave out Predictor	1	2	3	4	5,0	5	6	7	8	9	10	11	12	13	14	15	16	17	18
ROC Diff: Single Predictor	15,9	7,92	4,74	6,17	17,9	2,14	3,56	0,92	4,34	7,98	5,15	5,33	1,48	2,16	4,06	0,21	0,09	0,22	0,99
Rank# Single Predictor	1	3	7	4	1,0	12	10	15	8	2	6	5	13	11	9	17	18	16	14

Abbildung 5.15: Beitrag der Einzelprädiktoren des EuroSCORE-Risikosystems als Änderung der Fläche unter der ROC-Kurve. Die „ROC-Diff-Leave-out-Predictor“-Werte zeigen den marginalen Beitrag des Merkmals versus den kompletten Standardmerkmalssätzen (Φ=0.776); die „ROC Diff Single Predictor“-Werte sind der ROC-Beitrag der isolierten Merkmale (Gewinn über 0.5 mal 100). Zusätzlich ist das CC<55-Merkmals eingereiht, dessen eingeschobene Rangfolgezahl mit einem „,0“-Suffix markiert ist (d.h. 5,0 vor 5 und 1,0 vor 1).

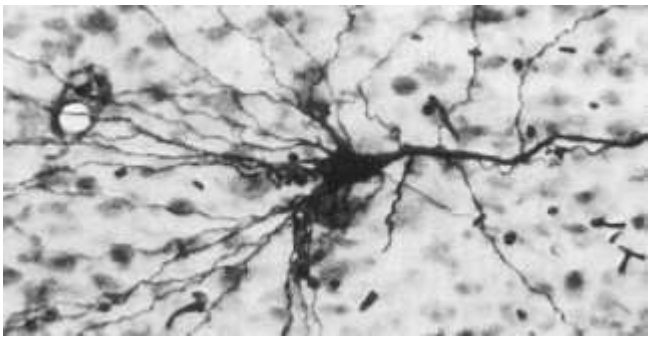


Abbildung 5.16: Ein Neuron im Mikroskopbild nach Färbung (visueller Kortex der Katze). Erkennbar sind (von links nach rechts) der Dentritenbaum, der Zellkörper (Soma) und ein langes Axon

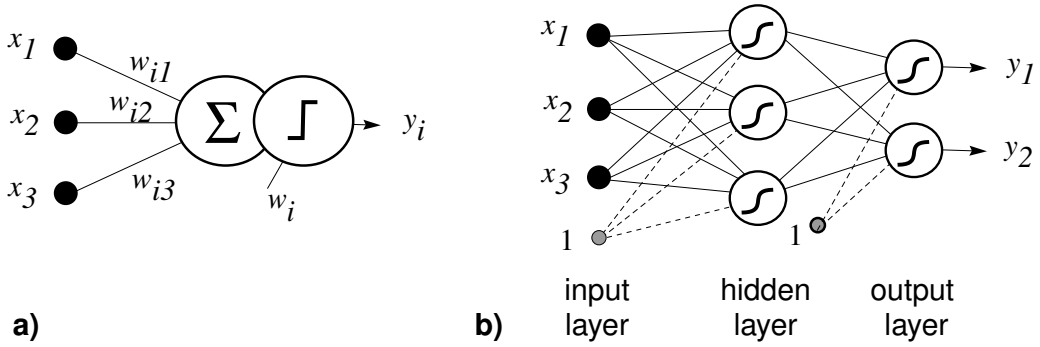
schlossen werden (1 aus 18) und der entstehende Φ-Verlust bestimmt werden. Abb. 5.7.4 zeigt sie nach Wichtigkeit *(ii)* geordnet und reiht den (CC<55)-Wert in der Rangfolge ein. Würde man die Modelle allein nach Φ-Differenzen bewerten, würde man nur das Alter außerhalb eines Standardfehlerintervalls finden. Tatsächlich aber ist CC in der ROC-Analyse das 5.-wichtigste Merkmal in der multivariaten Rangfolge und das allerwichtigste univariate Merkmal und ist damit vorteilhafter als *Serum Creatinin* mit jeweils Platz 14 und 9 (Walter et al. 2003).

Die ROC-Analyse ist sehr aussagekräftig und nicht auf die im Kontext vorgestellten Regressionmodelle beschränkt.
Im nächsten Abschnitt liegt der Schwerpunkt auf der Familie der so genannten neuronale Netze. Wie in Abs. 5.2 bereits erwähnt, sind die lineare und die logistische Regression strukturell sehr eng mit dem Perzeptron, einem einfachen neuronalen Netzwerkverfahren, verknüpft.

5.8 Neuronale-Netz-Modelle: MLP

„If the brain were so simple that we could understand it then we'd be so simple that we couldn't." (Lyall Watson)

Neuronale Netze bezeichnen eine Gruppe von Algorithmen, die durch das Vorbild des biologischen Gehirns motiviert sind. Das menschliche Ge-



a) **b)**

Abbildung 5.17: *(a, links)* Das **McCulloch-Pitts-Neuron** (1946) als frühes Modell eines biologischen Neurons „feuert“ (Ausgabe y_i=1 sonst 0), wenn die gewichtete Summe der Eingänge $\sum_j w_{ij}x_j$ den Schwellenwert w_i übersteigt. Die binäre Transferfunktion, auch „Aktivierungsfunktion“ genannt, wird häufig zur nichtlinearen Sigmoide Gl. 5.11 verallgemeinert und ist Bestandteil des *(b, rechts)* Standard-***Multi-Layer-Perceptron*** **(MLP)**. Es ist schichtweise aufgebaut: links die Eingabeschicht, dann ein oder mehrere „verborgene“ Schichten (*hidden layers*) und eine Ausgabeschicht, die, je nach Anwendungszweck, auch eine lineare Transferfunktion besitzen kann.

hirn verarbeitet Informationen gänzlich anders als unsere üblichen Computer (Von-Neumann-Architektur). Es besteht aus sehr vielen ($\approx 10^{10}$) Neuronen, die zudem hochgradig miteinander vernetzt sind (Vernetzungsgrad $\approx 10^4 - 10^5$). Ein Neuron strukturiert sich (s. Abb. 5.16) in seinen *(i)* Dentritenbaum, den *(ii)* Zellkörper (Soma) und *(iii)* ein langes Axon, welche sich meist den Aufgaben *(i)* Informationseingabe, *(ii)* -verarbeitung und *(iii)* Ausgabe zuordnen lassen. Die Information wird in Form von transienten Potentialänderungen (*Spikes*) entlang der Zellausläufer transportiert. Die Verbindungsstellen zu anderen Neuronen bilden *Synapsen*: Kontakte, die die Information mittels *Neurotransmitter* (spezifische chemische Botenstoffe) in eine Richtung übertragen und zu einer Erregung des Zielneurons beitragen (*exzitatorisch*) oder dieser abträglich sind (*inhibitorisch*).

Neuronen arbeiten im Vergleich zu Computern sehr langsam, hochparallel und fehlertolerant. Das Studium der Funktionsweise des Gehirns hat bislang zwar noch nicht zu einer umfassenden Aufklärung geführt, aber dennoch wurden ein Reihe von Mechanismen kognitiver Leistungen erhellt und Lernalgorithmen und -strukturen entdeckt. Sie lösen sich unterschiedlich weitgehend vom biologischen Vorbild und bilden technische Abstraktionen, „künstliche neuronale Netze“ genannt.

Abb. 5.17 illustriert und erläutert das historische McCulloch-Pitts-Neuron (1943) und den Prototyp des neuronalen Netzwerkes, das klassische Multi-Lagen-Perzeptron (MLP). Es ist ein *feed-forward network*, d.h. die Aktivierungen werden schichtweise von der Eingabe- zur Ausgabeschicht propagiert. In der umgekehrten Richtung wird zum Nachtrainieren der Netzwerkgewichte $\mathbf{w}_{ij}$ das Fehlersignal (Soll-Ist-Differenz) nach der so genannten Deltaregel propagiert. Die (mehrfache und unabhängige) Entdeckung dieses Verfahrens Mitte der 90er Jahre löste einen großen Aufschwung im Forschungsfeld der neuronalen Netze aus und gab dem MLP auch den Zweitnamen ***Back-Propagation Net***. Zur Beschleunigung des Lernens finden einige raffinierte Lernverfahren (z.B. konjugiertes Gradientenabstiegsverfahren) Verwendung (s. z.B. Orr und Müller 1998).

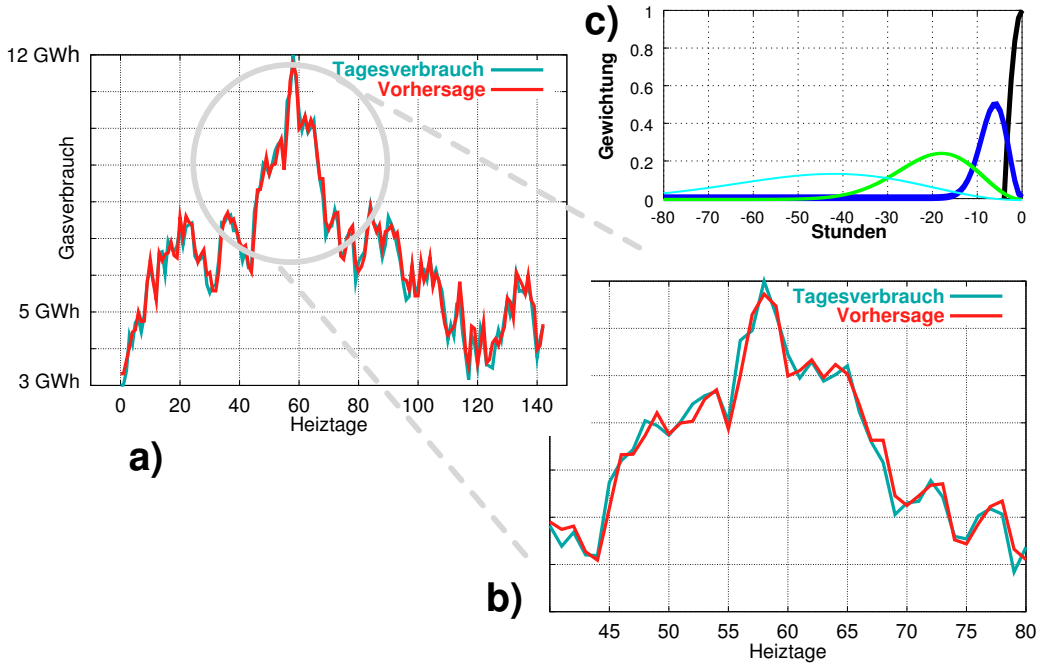


Abbildung 5.18: *(a,b)* Zeitserie des wahren und prognostizierten städtischen Gasverbrauchs (seit Anfang der Heizperiode 1. November). Die Schätzung erfolgt mittels eines 12-3-1 MLP (mit einem linearen Ausgangsneuron und drei verborgenen Neuronen mit sigmoider Aktivierungsfunktion). Die Merkmale entstehen aus Faltungen der stündlichen Messungen von Temperatur, Wind und Gasmenge mit vier Kernelfunktionen verschiedener Breite *(c)* .

Abb. 5.18 zeigt ein Anwendungsbeispiel aus dem Bereich Energieverbrauchsprognose aus Zeitserien. In einem Kooperationsprojekt mit den Stadtwerken Bielefeld, einem kommunalen Energieversorger, wurde an-

hand von stündlichen Temperatur- und Windmessungen der Tagesbedarf an Stadtgas prognostiziert. Ziel war es, überschwelligen Leistungsspitzen, die hohe Vertragszusatzkosten zu Folge hatten, entgegenwirken zu können (Arnrich und Walter 2000).

Gibt es zyklische Signalverbindungen, z.B. von der Ausgabeschicht zur Eingabeschicht, spricht man von rekurrenten Netzen, die sich durch ein reiches Spektrum von möglichen Dynamiken auszeichnen. Sie werden u.a. als Assoziationsspeicher zu Steuerungsaufgaben oder zur Gruppierung eingesetzt, s. Abs. 5.4.4.

Das oben beschriebene Modell radialer Basisfunktionen (RBF, s. Abb. 5.2) kann als Netzwerk aus parallel arbeitenden Neuronen dargestellt werden und wird daher auch zu den neuronalen Verfahren gezählt.

5.9 Selbstorganisierende Karten

Weitere wichtige Vertreter neuronaler Modelle sind die selbstorganisierenden Karten (*Self-Organizing Maps* **SOM**), die ursprünglich von Teuvo Kohonen (1982) als mathematisches Modell für die Morphogenese bestimmter Hirnareale formulierte wurden. Ein Beispiel ist der somatosensorische Kortex, ein Rindenfeld des Großhirns, das sich als eine zweidimensionale Abbildung, eine topographische Karte der ganzen Hautoberfläche im Gehirn charakterisieren lässt. Die neuronalen Verbindungen zu den verschiedenen Hautsensoren sind viel zu zahlreich, um sie genetisch im Detail vorzuprogrammieren. Stattdessen werden sie in einer frühen Lebensphase datengetrieben und selbstorganisiert ausgebildet. „Topographisch“ heißt, dass benachbarte Hautregionen (allgemeine Merkmale) auf benachbarte Neuronen im Kortex abgebildet werden. Dies ist für die Weiterverarbeitung sehr vorteilhaft, z.B. wenn zusammenhängende oder sich verschiebende Hautkontakte registriert werden. Ein weiteres Beispiel sind die retinotopischen Karten im primären visuellen Kortex (z.B. Obermayer et al. 1990, weiterführende SOM-Darstellungen finden sich z.B. in Ritter et al. 1991; Kohonen 2001).

Abb. 5.19 zeigt den Aufbau der SOM oder Merkmalskarte als Gitter A von Knoten oder „Neuronen“. Jedes Neuron a ist mit einem Referenzvektor $\mathbf{w_a}$ verknüpft, der in den Eingaberaum X projiziert.

Die Antwort der SOM auf einen neuen Eingabevektor $\mathbf{x}$ ist bestimmt

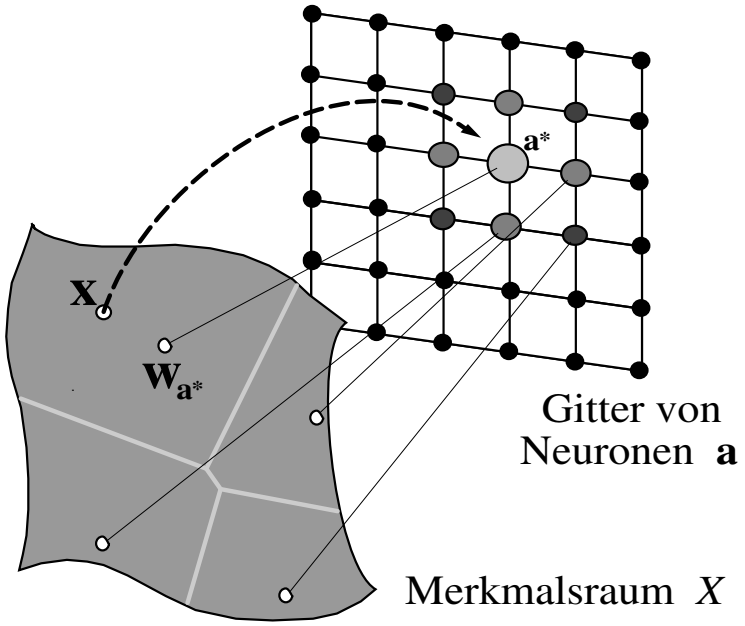


Abbildung 5.19: Die „Self-Organizing Map“ („SOM“) wird durch ein Gitter von verarbeitenden Knoten oder „Neuronen“ gebildet. Jedes Neuron hat einen Prototypen- oder Referenzvektor $\mathbf{w}_a$ zugeordnet, der im Eingaberaum X eingebettet ist. Ein neuer Vektor $\mathbf{x}$ wird auf dasjenige Neuron abgebildet, dessen $\mathbf{w}_a$ am nächsten liegt. Dieser kompetitive Mechanismus teilt den Eingaberaum in diskrete Abschnitte, den so genannten *Voronoi-Zellen*.

durch den am besten passenden Knoten (*best matching unit*, BMU, oder „Gewinner“):

$$\mathbf{a}^* = \underset{\forall \mathbf{a}' \in \mathbf{A}}{\operatorname{argmin}} \; \|\mathbf{w}_{\mathbf{a}'} - \mathbf{x}\| \,. \tag{5.70}$$

Dieser Wettbewerb zwischen den Knoten kann biologisch als Ergebnis einer lateralen Inhibition in der Neuronenschicht interpretiert werden. Mathematisch wird dadurch eine dimensionsreduzierende Abbildung des Eingaberaumes X auf das Gitter A konstituiert.

Die Referenzvektoren oder Gewichte $\mathbf{w}_\mathbf{a}$ werden normalerweise iterativ durch eine Sequenz von Trainingsdaten adaptiert. Nachdem der *best-matching* Knoten $\mathbf{a}^*$ bestimmt wurde, wird nicht er allein, sondern es werden *alle* Knotenvektoren $\mathbf{w}_\mathbf{a}^{(new)} := \mathbf{w}_\mathbf{a}^{(old)} + \Delta\mathbf{w}_\mathbf{a}$ adaptiert, d.h. gemäß

$$\Delta\mathbf{w}_\mathbf{a} = \epsilon \, h(\mathbf{a}, \mathbf{a}^*) \, (\mathbf{x} - \mathbf{w}_\mathbf{a}) \tag{5.71}$$

mehr oder weniger dem Eingabevektor $\mathbf{x}$ angeglichen. Wie stark, hängt neben dem (globalen) Lernschrittparameter ϵ von der so genannten Nachbarschaftsfunktion

$$h(\mathbf{a}, \mathbf{a}^*) = h(\,\|\mathbf{a} - \mathbf{a}^*\|\,) = \exp -\frac{\|\mathbf{a} - \mathbf{a}^*\|^2}{\sigma^2}\,, \tag{5.72}$$

einer glockenförmigen Funktion die an der BMU $\mathbf{a}^*$ zentriert ist und mit wachsendem Abstand $|\mathbf{a}-\mathbf{a}^*|$ im Gitter abnimmt. Nach anfänglicher Zufallsinitialisierung der Knoten werden während des Trainings der Lernparameter ϵ und die Glockenbreite σ sukzessiv verkleinert, um von einem groben, aber schnellen Lernen graduell zu einem „Feinlernen“ zu gelangen.

Durch dieses kooperative Lernverfahren erreicht der SOM-Algorithmus mehrere Vorteile:

- Er kann eine topographische Ordnung zwischen den Referenzvektoren $\mathbf{w_a}$ ausbilden;
- Die Konvergenz ist verbessert, da statt nur einem Knoten eine ganze Gruppe an jedem Lernschritt partizipiert;
- Das Lernen eines Ausgabewertes wird robuster.

Durch Verknüpfung jedes Knotens mit einem Ausgabewert oder einer lokal gültigen multivariaten linearen Regression (*Local Linear Map*, **LLM**) erhält man ein Approximationsmodell. Die LLM haben sich in vielen Anwendungen bewährt, z.B. beim Lernen der visuo-motorischen Koordination eines Industrieroboters (Ritter et al. 1989; Walter und Schulten 1993) oder zur Zeitserienvorhersage (Walter et al. 1990), u.a. auch in Verbindung mit dem Neuronen-Gas-Netzwerk (Walter 1991; Martinetz et al. 1993). Der *Neural Gas* Algorithmus modifiziert die Nachbarschaftsfunktion Gl. 5.72 auf eine dynamische, Abstandsrang-basierte Form, die von der Gitterstruktur losgelöst (und damit gasartig) ist. Eine Weiterentwicklung der SOM, speziell zum schnellen Lernen von kontinuierlichen und glatten Abbildungen, ist die „Parametrisierte SOM“ (**PSOM**). Sie wird in Abs. 7.2.1 bzgl. ihres Interpolationsvermögens diskutiert.

Neben dem Einsatzziel der Approximation wird die SOM zu Cluster- und Klassifikationssaufgaben und sehr häufig zur Visualisierung eingesetzt.

Abb. 5.20 zeigt ein Beispiel einer SOM-Visualisierung zur Identifizierung erfolgversprechender Behandlungsstrategien im Bereich Herzchirurgie (s.a. Kap. 9). Das zweidimensionale rechteckige Gitter von Knoten ist

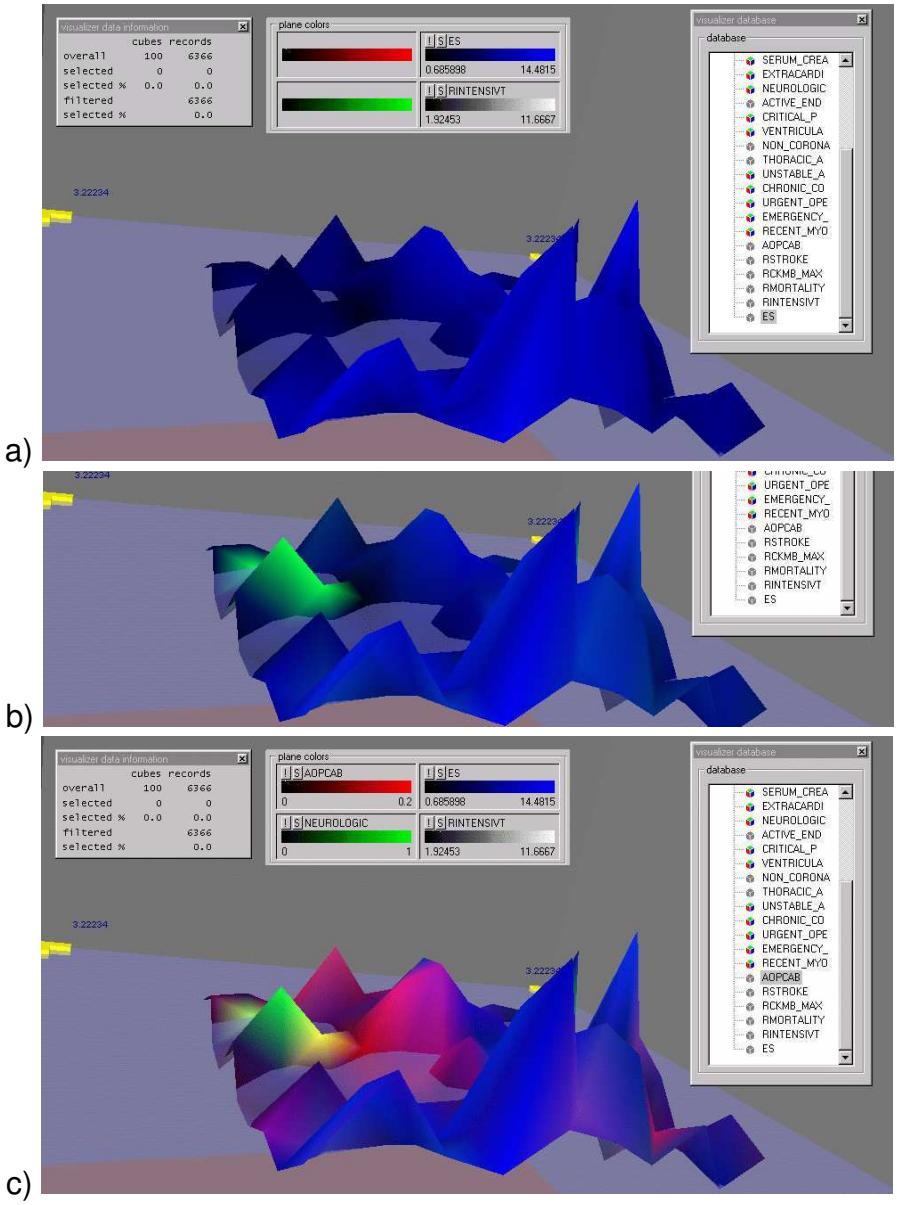


Abbildung 5.20: SOM-Visualisierung und Hypothesenbildung. 18 Merkmale (EuroSCORE, s. Abs. 9.1.2) wurden im SOM Training verwandt. Hier eine integrierte Visualisierung von vier Merkmalen der SOM-Knoten: (*i*) Die Vertikalachse zeigt die Intensivaufenthaltsdauer (die verstellbare, graue Wasserfläche hilft den Höhenverlauf zu verdeutlichen), (*ii*) die Blauintensität der Flächenfärbung zeigt den mittleren EuroSCORE-Wert, (*iii*) die Grünintensität zeigt die Häufigkeit präoperativer neurologischer Dysfunktionen und (*iv*) die Rotintensität zeigt den OPCAB-Anteil. Die Mischfarbe hellgelb bedeutet hier, dass in diesem Bereich wenig Blau, aber viel Rot und Grün auf kleiner Höhe zusammentreffen.

hier dreimal als Gebirge mit verschiedenen Einfärbungen der Zwischenflächen dargestellt. Die Höhe markiert den (BMU-bedingten) Erwartungswert des Merkmals Intensivaufenthaltsdauer, die Farben Rot, Grün, Blau jeweils das Merkmal Therapieform OPCAP, Vorliegen einer neurologischen Dysfunktion und die Risikobewertung nach EuroSCORE (s. Abs. 9.1.2).

Die auftretende Mischfarbe Hellgelb bedeutet in Abb. 5.20c, dass in diesem Bereich wenig Blau, aber viel Rot und Grün auf geringer Höhe zusammentreffen. Sie ist benachbart zu einer Spitze gleichen Grüns (*b*), aber ohne Rotanteil (andernfalls Gelb). Das bedeutet, dass bei Vorliegen einer neurologischen Dysfunktion die OPCAB-Technolgie im Mittel zu kürzeren Aufenthalten in der Intensivstation führt (s.a. Albert, Walter, Rosendahl, Schröder und Ennker 1999).

In Abs. 7.2 wird eine Erweiterung des SOM-Algorithmus für nicht-euklidische Räume vorgestellt und es werden weitere Visualisierungs- und Explorationsbeispiele im Bereich semantischer Karten von Textdokumenten präsentiert. Im folgenden Abschnitt wird eine weitere Methode zur Visualisierung vorgestellt, die besonders flexibel bezüglich der Datenrepräsentation ist.

5.10 Multidimensionale Skalierung (MDS)

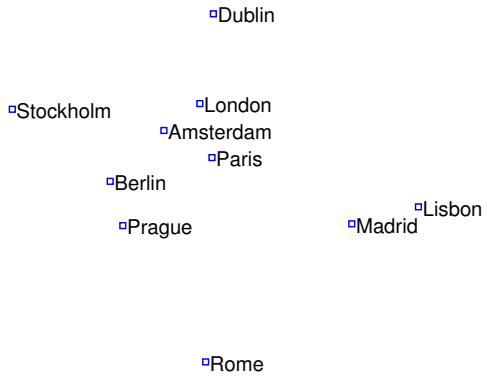
Die multidimensionale Skalierung (*Multi-Dimensional Scaling*, **MDS**) bezeichnet eine Klasse von Techniken für die Analyse von Objekten, deren paarweise Distanz- bzw. Unähnlichkeitsmaße vorliegen (s. Abs. 2.2). Ziel ist es, die in den Daten verborgene Struktur sichtbar zu machen, indem man eine geeignete Objektplatzierung in einem niedrigdimensionalen, typischerweise euklidischen Raum erzeugt. Im Folgenden bezeichnen $\delta_{ij} \in \mathbb{R}_0^+$ die Unähnlichkeit (*dissimilarity*) zwischen Objekt i und Objekt j. Sie ist üblicherweise symmetrisch $\delta_{ij} = \delta_{ji}$.

Das Ziel von MDS ist, eine räumliche, möglichst **abstandstreue** Repräsentation $\mathbf{x}_i$ für jedes Objekt i im L-dimensionalen Raum $\mathbb{R}^L$ zu finden. Abstandstreue bedeutet, dass die Paarabstände im Zielraum $d_{ij} \equiv d(\mathbf{x}_i, \mathbf{x}_j)$ so gut als möglich die Disparitätsstruktur der Objekte widerspiegeln, also $\forall_{i \neq j} \mathbf{D}_{ij} \approx d_{ij}$.

Häufig wird ein geeigneter δ_{ij}-Vorverarbeitungsschritt mit einer monotonen Funktion $T_{disp}(\cdot)$ zwischengeschaltet, der die rohen Unähnlichkeiten

#	Stadtabst.	1	2	3	4	5	6	7	8	9	10
1	London	0	569	667	530	141	140	357	396	570	190
2	Stockholm	569	0	1212	1043	617	446	325	423	787	648
3	Lisbon	667	1212	0	201	596	768	923	882	714	714
4	Madrid	530	1043	201	0	431	608	740	690	516	622
5	Paris	141	617	596	431	0	177	340	337	436	320
6	Amsterdam	140	446	768	608	177	0	218	272	519	302
7	Berlin	357	325	923	740	340	218	0	114	472	514
8	Prague	396	423	882	690	337	272	114	0	364	573
9	Rome	569	787	714	516	436	519	472	364	0	755
10	Dublin	190	648	714	622	320	302	514	573	755	0

Abbildung 5.21: *(Oben:)* Entfernungen zwischen zehn europäischen Hauptstädten. *(Unten:)* Multidimensionale Skalierung auf $\mathbb{R}^2$.



δ_{ij} in **Disparitäten** $\mathbf{D}_{ij}$ überführt:

$$\mathbf{D}_{ij} = T_{disp}(\delta_{ij}). \tag{5.73}$$

Im Standardfall ist dies die Identitätsabbildung $\mathbf{D}_{ij} = \delta_{ij}$.

5.10.1 Klassische multidimensionale Skalierung

Klassisches multidimensionales Skalieren geht zurück auf Young und Householder (1938) und ist eng verknüpft mit der Rekonstruktion von Koordinaten. Ein Beispiel ist die Kartenerstellung aus Entfernungen zwischen Ortspaaren. Abb. 5.21 zeigt die Ortsrekonstruktion von europäischen Hauptstädten aus Distanztabellen (s. Borg und Groenen 1997, S. 16, und z.B. Cox und Cox 1994 als weitere Standardreferenz).

Seien $\mathbf{x}_r = (x_{r1}, \ldots, x_{r_L})^T$ die gesuchten Koordinaten von N Punkten in einem L-dimensionalen euklidischen Raum ($r = 1, \ldots, N$), dann sind die euklidischen Paardistanzen offensichtlich durch

$$\delta_{ij}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s) = \mathbf{x}_r^T\mathbf{x}_r + \mathbf{x}_s^T\mathbf{x}_s - 2\mathbf{x}_r^T\mathbf{x}_s \tag{5.74}$$

beschreibbar. Der Rekonstruktionsweg der $\mathbf{x}_r$ baut auf einer spektralen Zerlegung der Matrix $[\mathbf{B}]_{rs} = b_{rs} = \mathbf{x}_r^T \mathbf{x}_s$ auf. Wie bekommt man die innere Produktmatrix $\mathbf{B}$? Man kann leicht zeigen (Cox und Cox 1994), dass die Spalten- und Zeilenzentrierung der Hilfsmatrix $\mathbf{A}$ mit $[\mathbf{A}]_{rs} = a_{rs} = -\frac{1}{2}\delta_{ij}^2$ dies bereits leistet.

$$\mathbf{B} = \mathbf{A} - N^{-1}(\mathbf{A}\mathbf{1})^T\mathbf{1} - N^{-1}\mathbf{1}(\mathbf{A}\mathbf{1})^T + N^{-2}\mathbf{1}(\mathbf{A})^T\mathbf{1}. \tag{5.75}$$

Die Subtraktion der Spalten- und Zeilenmittelwerte und Addition des Gesamtmittels ist äquivalent dem Abzug des Spaltenmittels und dann des neuen Zeilenmittels.

Fügt man die $\mathbf{x}_r$-Vektoren zur Matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times p}$ zusammen, kann man die symmetrische Matrix $\mathbf{B}$ als Matrixprodukt darstellen und spektral zerlegen

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T. \tag{5.76}$$

Die Eigenwertzerlegung liefert die Diagonalmatrix $\mathbf{\Lambda}$ =diag$(\lambda_1, \lambda_2, \ldots, \lambda_N)$ der (absteigend sortierten) Eigenwerte $\{\lambda_i\}$ von $\mathbf{B}$ und deren korrespondierenden, orthonormalen Eigenvektoren $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$. Stammen die $\{b_{rs}\}$ von diesem idealen Matrixprodukt, ist $\mathbf{B}$ positiv semi-definit mit Rang $rg(\mathbf{B}) = L$. Damit gibt es L nicht-negative Eigenwerte und $N - L$ Eigenwerte, die 0 sind. Dieser Nullraum wird verworfen und $\mathbf{B}$ kann kompakter als

$$\mathbf{B} = \mathbf{V}_L\mathbf{\Lambda}_L\mathbf{V}_L^T \quad \text{mit } \mathbf{\Lambda}_L = \text{diag}(\lambda_1, \ldots, \lambda_L), \quad \mathbf{V}_L = [\mathbf{v}_1, \ldots, \mathbf{v}_L] \tag{5.77}$$

beschrieben werden. Die gesuchte Koordinatenrekonstruktion ist (bis auf isometrische Transformationen) gegeben durch

$$\mathbf{X} = \mathbf{V}_L\mathbf{\Lambda}_L^{\frac{1}{2}} \qquad \text{mit } \mathbf{\Lambda}_L^{\frac{1}{2}} = \text{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_L}\right) . \tag{5.78}$$

In der Praxis sind die Distanzen häufig empirischen Ursprungs und die Frage nach der optimalen Einbettungsdimension L ist offen. Aufschluß geben die Eigenwerte $\{\lambda_i\}$. Gower (1966) führte hierfür den Begriff der **Analyse der prinzipalen Koordinaten** (*Principal COordinate analysis*, PCO) ein. Analog zur PCA wird der Beitrag einer Komponente i zur Varianz der Summe der quadrierten Distanzen durch λ_i erklärt. Damit kann die Wahl der Anzahl einflußstärkster Koordinaten L mit dem Maß der relativen Varianzabdeckung gesteuert werden:

$$\text{relative Varianzerklärung}(\sum_{\forall ij} \delta_{ij}^2) = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{N-1} \lambda_i} . \tag{5.79}$$

Ist $\mathbf{B}$ nicht positiv semi-definiert, treten negative Eigenwerte auf und die modifizierte Summe im Nenner (Gl. 5.79) läuft dann über $|\lambda_i|$ oder nur über die positiven λ_i. Treten nur betragsmäßig kleine, negative $|\lambda_i|$ auf, können diese Koordinaten ignoriert werden.

Dieses klassische, metrische Skalierverfahren zielt auf die optimale metrische Einbettung. Erfüllen die δij nicht die Metrikeigenschaften oder wird L (auch absichtlich) zu klein gewählt, ist die folgende, modernere Problemformulierung vorteilhafter.

5.10.2 Least-Square- oder Kruskal-Scaling

Kruskal führte (1964) als erster eine Kostenfunktion ein, die MDS als nichtlineares Optimierungsproblem formuliert. Es gilt, die Stressfunktion zu minimieren

$$STRESS(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{\sqrt{\sum_{i=1}^{N} \sum_{j>i} (d_{ij} - \mathbf{D}_{ij})^2}}{\sqrt{\sum_{i=1}^{N} \sum_{j>i} \mathbf{D}_{ij}^2}}. \qquad (5.80)$$

Dies ist im Wesentlichen die Summe über alle paarweisen Abstandsverzerrungen mit einer Normierungskonstante im Nenner. Dabei ist die Wurzel Konvention.

Der Zielraum (*target space*) ist typischerweise niedrigdimensional, oft mit L =2 oder 3 und alle Paarabstände werden darin euklidisch bestimmt:

$$d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j|| \quad \text{mit} \quad \mathbf{x}_i \in \mathbb{R}^L,\ i, j \in \{1, 2, ..N\}. \qquad (5.81)$$

Es gibt einige Varianten von Gl. 5.80. Eine mit dem Namen **S**STRESS bekannte Kostenfunktion ersetzt die Terme d_{ij} und $\mathbf{D}_{ij}$ durch deren Quadrate. STRESS**2** benennt eine Normierungsvariante, die im Nenner statt $\mathbf{D}_{ij}$ den Term $\mathbf{D}_{ij} - \mathbf{D}_{\cdot\cdot}$ einsetzt ($\mathbf{D}_{\cdot\cdot}$ = Mittelwert der $\mathbf{D}_{ij}$).

5.10.3 MDS nach Sammon

Eins der bekanntesten Verfahren wurde von John Sammon (1969) vorgeschlagen. Er formulierte *Multi-Dimensional Scaling* als Optimierungspro-

blem bezüglich folgender Kosten- oder Stressfunktion

$$E(\{\mathbf{x}_i\}) = \sum_{i=1}^{N} \sum_{j>i} w_{ij}(d_{ij} - \mathbf{D}_{ij})^2. \tag{5.82}$$

Hierbei gewichten und normieren die Faktoren w_{ij} die Disparitäts-Distanzverzerrungs-Paare. Damit Gl. 5.82 als Gütekriterium tauglich ist, muss es invariant gegen die Skalierung des Ausgangsproblems sein. Abhängig von der Aufgabenstellung kann man alle Disparitäts-Distanzverzerrungen gleich bewerten, die „globale“ Variante ($w_{ij}^{(g)} = const$) wählen, oder die Wiedergabe lokaler Feinstrukturen betonen, indem man den Einfluss von großen Disparitäten reduziert ($w_{ij}^{(l)}$)

$$w_{ij}^{(g)} = \frac{1}{\sum_{k=1}^{N} \sum_{l>k} \mathbf{D}_{kl}^2}\,, \qquad w_{ij}^{(l)} = \frac{2}{N(N-1)} \frac{1}{\mathbf{D}_{ij}^2}\,. \tag{5.83}$$

Zu beachten ist, dass der letzte Term undefiniert wird, wenn Paarabstände von 0 auftreten. In seiner Originalarbeit schlug Sammon einen Mittelweg vor

$$w_{ij}^{(m)} = \frac{1}{\sum_{k=1}^{N} \sum_{l>k} \mathbf{D}_{kl}} \frac{1}{\mathbf{D}_{ij}}\,, \tag{5.84}$$

der auch nachfolgend verwendet wird. Die Lösung des Minimierungsproblems Gl. 5.82 wird iterativ mit einem Gradientenabstieg erreicht: ausgehend von einer Startkonfiguration $\{\mathbf{x}_i\}$, wird in jedem Schritt ein Objekt i^* zufällig ausgewählt und Gl. 5.82 bezüglich $\mathbf{x}_{i^*}$ minimiert

$$\mathbf{x}_{i^*}^{(new)} = \mathbf{x}_{i^*}^{(old)} + \eta\, \Delta_{i^*}. \tag{5.85}$$

Mit Hilfe des (diagonalen) Newtonverfahrens wird Δ_{i^*} bestimmt, in Komponentenschreibweise ($q \in \{1, .., L\}$)

$$\Delta_{i^*,q} = -\frac{\partial E}{\partial x_{i^*,q}} \Big/ \left| \frac{\partial^2 E}{\partial x_{i^*,q}^2} \right|. \tag{5.86}$$

Die Nebendiagonalelemente der Hessematrix werden hier ignoriert. Der Algorithmus konvergiert in der Regel nach wenigen Epochen auf ein Stressminimum. Üblicherweise begegnet man dem Problem von lokalen Minima, indem von mehreren Initialkonfigurationen ausgehend gestartet und das Ergebnis mit dem kleinsten Stress E gewählt wird. Gut ausgewählte Startkonfigurationen $\{\mathbf{x}_i\}$ sparen dabei Rechenzeit. Liegen Vektordaten vor, können die ersten L Komponenten einer PCA verwendet werden. Für Distanzdaten kann direkt das FastMap-Verfahren eingesetzen werden, das in Abs. 5.10.4 beschrieben ist.

Ein vergleichendes Anwendungsbeispiel aus dem Bereich Farbkognition wird im nächsten Abschnitt in Abb. 5.24 gezeigt.

5.10.4 Dimensionsreduktion mit FastMap

Faloutsos und Lin (1995) entwickelten ein dem MDS zielverwandtes und besonders gut skalierbares Einbettungsverfahren für Distanzdaten. Zentral ist dabei die mutige Annahme, dass die gegebenen Paardistanzwerte die euklidischen Distanzen von N Punkten in einem k-dimensionalen Vektorraum sind, wobei k nicht im Voraus bekannt sein muss.

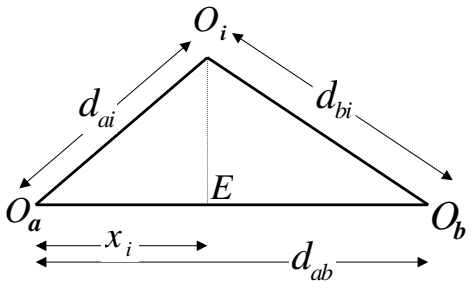


Abbildung 5.22: Illustration des Kosinussatzes: Projektion des Punktes O_i auf die von O_a, O_b aufgespannte Gerade.

Der Algorithmus arbeitet rekursiv und wählt zuerst zwei Datenpunkte, die so genannten Pivotelemente O_a, O_b, aus der Menge von möglichen Datenpunkten aus (die Details werden unten erläutert). Jeder weitere Punkt O_i ($i \in \{1, \ldots, N\}$) wird auf die davon aufgespannte Gerade projiziert, wie in Abb. 5.22 illustriert. Mittels des Kosinussatzes

$$d_{b,i}^2 = d_{a,i}^2 + d_{a,b}^2 - 2x_i d_{a,b} \tag{5.87}$$

kann die Projektionskoordinate x_i einfach aus den Abständen ermittelt werden

$$x_i = \frac{d_{a,i}^2 + d_{a,b}^2 - d_{b,i}^2}{2d_{a,b}} \,. \tag{5.88}$$

Die Schreibweise $d_{b,i}$ ist dabei die Abkürzung des Abstandes $D(O_b, O_i)$ in der gültigen Metrik. Für $k = 1$ ist das Quasi-MDS-Problem bereits durch Gl. 5.88 gelöst.

Stimmt die Annahme, kann man das Vorgehen auf einer Hyperebene im $k-1$-dimensionalen Raum wiederholen. Wie in Abb. 5.23 zu sehen ist,

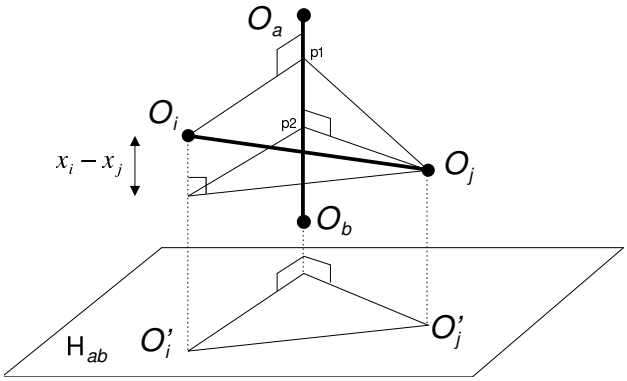


Abbildung 5.23: Projektion auf eine Hyperebene H_{ab}, die senkrecht auf der von O_a, O_b aufgespannten Gerade aus der vorigen Abb. 5.22 steht.

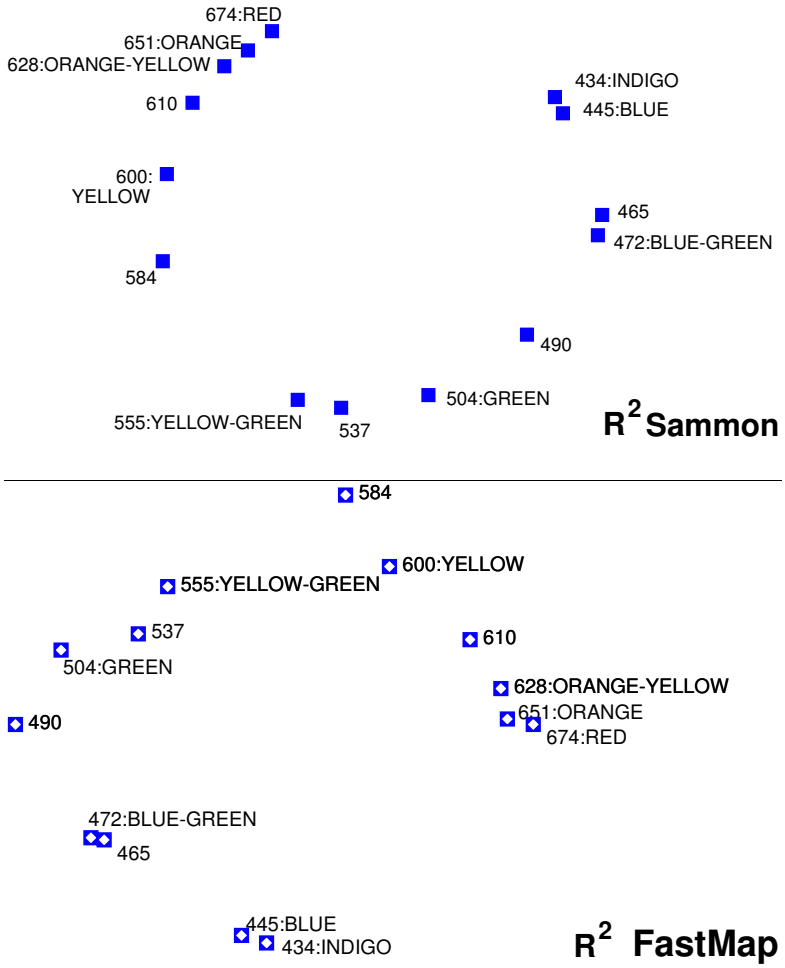
steht sie senkrecht auf der von O_a, O_b aufgespannten Geraden. Wie bekommt man den Abstand der Projektionsorte O'_i, O'_j zweier anderer Punkte? Mittels des Satzes von Pythagoras, angewendet auf das Dreieck im Bildvordergrund, findet man

$$[D'(O'_i, O'_j)]^2 = [D(O_i, O_j)]^2 - (x_i - x_j)^2 \qquad i \in \{1, \ldots, N\}. \tag{5.89}$$

Damit können alle Paarabstände im Unterraum berechnet werden. Dieser steht per Konstruktion senkrecht auf der Projektionsgeraden. Nun kann das Verfahren im Prinzip $k - 1$ Schritte rekursiv weiterlaufen. Für jeden Punkt werden die rekursiv ermittelten k Projektionsorte $\{x_i^{(\prime)}\}$ als Koordinaten in der Einbettung in den (um eine Dimension reduzierten) Unterraum verwendet.

#	**Farbähnlichkeiten**	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	434:INDIGO	100	86	42	42	18	6	7	4	2	7	9	12	13	16
2	445:BLUE	86	100	50	44	22	9	7	7	2	4	7	11	13	14
3	465	42	50	100	81	47	17	10	8	2	1	2	1	5	3
4	472:BLUE-GREEN	42	44	81	100	54	25	10	9	2	1	0	1	2	4
5	490	18	22	47	54	100	61	31	26	7	2	2	1	2	0
6	504:GREEN	6	9	17	25	61	100	62	45	14	8	2	2	2	1
7	537	7	7	10	10	31	62	100	73	22	14	5	2	2	0
8	555:YELLOW-GREEN	4	7	8	9	26	45	73	100	33	19	4	3	2	2
9	584	2	2	2	2	7	14	22	33	100	58	37	27	20	23
10	600:YELLOW	7	4	1	1	2	8	14	19	58	100	74	50	41	28
11	610	9	7	2	0	2	2	5	4	37	74	100	76	62	55
12	628:ORANGE-YELLOW	12	11	1	1	1	2	2	3	27	50	76	100	85	68
13	651:ORANGE	13	13	5	2	2	2	2	2	20	41	62	85	100	76
14	674:RED	16	14	3	4	0	1	0	2	23	28	55	68	76	100

Tabelle 5.4: Ähnlichkeitsbewertung s_{ij} für 14 monochrome Farben, gemittelt über 31 Probanden (nach Ekman, 1954). Eine geeignete Unähnlichkeits-Transformation der Skala von 0 bis 100 ist $\delta_{ij} = 1 - s_{ij}/100$. MDS-Ergebnis in Abb. 5.24.



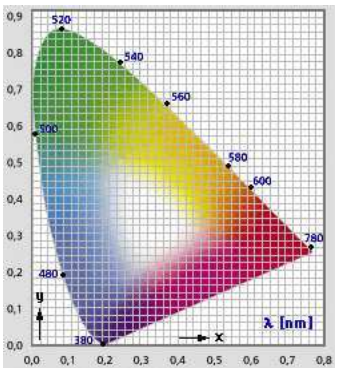


Abbildung 5.24: Einbettungsergebnis der Ähnlichkeitseinschätzungen für die Farbwahrnehmungsdaten aus Tab. 5.4. Die Zahlen sind Wellenlängenangaben in *nm*. *(a, oben)* mit dem Sammon-MDS-Algorithmus und (*b, Mitte*) mittels der FastMap. Die U-förmige Platzierung der Daten korrespondiert mit dem äußerem Rand der „reinen Farben“ im heutigen Standardfarbmodell, wie der Vergleich mit der Illustration der CIE-xy-Farbnormtafel *(c, links unten)* zeigt.

Bleibt zu klären, wie man in jedem Rekursionsschritt die Pivotobjekte a, b auswählt. FastMap lehnt sich hier konzeptionell am Karhunen-Loeve- oder PCA-Verfahren an: Die 1. Hauptachse ist die varianzstärkste Achse, die zweite Hauptachse ist die zweitstärkste, dazu senkrechte u.s.w. Im Unterraum, der durch Projektion entlang der 1. Hauptachse entsteht, ist die 2. Hauptachse wiederrum die varianzstärkste. In der Sequenz von Unterraumprojektionen verhält sich FastMap in der Wahl der Achsen ähnlich, mit einer weiteren mutigen Vereinfachung: Statt auf die Varianz der Datenverteilung einzugehen, wird einfach das Objektpaar mit dem größten Abstand gesucht und als Pivotelement ausgewählt. Da alle Abstände ja gegeben oder zwischenberechnet sind, umfasst dies die Suche nach dem größten Eintrag in der Distanztabelle. Um auch hier eine Beschleunigung herbeizuführen, wurde eine iterative Liniensuche vorgeschlagen. Man startet bei einem Objekt i', läuft die entsprechende i'-te Zeile entlang, sucht den größten Eintrag j'' und wiederholt den Vorgang in dessen Spalte, dann in dessen Zeile i'' u.s.w. Damit landet man sehr schnell zumindest in einem lokalen Maximum. Die Gegenstrategie ist der Mehrfachstart an mehreren (s) zufälligen Startpunkten i' und die Verwendung des besten gefundenen Pivotpaares.

Berechnungsaufwand: Mit diesem Trick bleibt das gesamte Verfahren linear in der Anzahl der Datenobjekte $O(N)$, denn es müssen tatsächlich nicht alle $N(N-1)/2$ Distanzen ausgewertet werden, sondern nur die zu den Pivotpaaren gehörenden, also $O(kNs)$ viele.

Stabilität: Die beschriebene Auswahlheuristik macht das FastMap-Ergebnis instabil, denn zwei Durchläufe ergeben nicht notwendigerweise gleiche Ergebnisse.

Vergleich mit Karhunen-Loeve: Das FastMap-Ergebnis kann zum PCA-Verfahren sehr unterschiedlich sein, wie in Abb. 5.25 abstrakt gezeigt ist. Gegenüber der FastMap ist die PCA auf Fälle mit vektoriell vorliegenden Objektdaten beschränkt.

Abbildung neuer Datenobjekte: Durch Speichern der Pivotdaten lässt sich eine sehr performante Abbildung eines neuen Objektes in den k-dimensionalen Zielraum darstellen.

Reduktion der Abbildungskosten: Der Umstand, dass für die Abbildung eines neuen Objekts nur die Abstände zu den Pivotelementen zur Verfügung stehen müssen, um eine k-dimensionale Einbettung zu bekom-

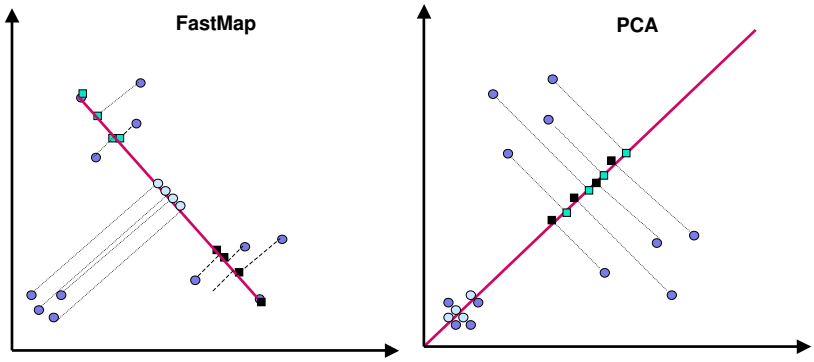


Abbildung 5.25: Vergleich der Projektionsachsen Fastmap (Pivotpaar) und PCA (varianzstärkste Richtung). Im Extremfall kann der Datensatz wegen der unterschiedlichen Auswahl der Projektionsachse zu sehr unterschiedlichen Ergebnissen führen.

men, ist insbesondere vorteilhaft, wenn die Beschaffung von Abstandsdaten d_{ij} aufwändig ist. Sind die Beschaffungskosten objektabhängig (z.B. wegen Experimentkosten oder Laufzeiten), kann dies zudem in der Pivotwahl berücksichtigt werden.

Robustheit: Offensichtlich ist das Verfahren nicht resistent gegen große d_{ij}-Ausreißer. Wenn sie vom Pivotsuchverfahren visitiert werden, dominieren sie unweigerlich das Projektionsergebnis.

Verletzung der Grundannahme: Stammen die Paarabstände nicht von einem euklidischen Vektorraum, sind also allgemeine Unähnlichkeitsdaten, werden die Voraussetzungen einer Metrik nicht mehr garantiert. Da die Cauchy-Schwarz'sche-Dreiecksungleichung (Gl. 2.5) dann nicht allgemein gültig ist, kann der Ausdruck in Gl. 5.89 negativ werden. Entweder ignoriert man den Imaginärteil und setzt $D'(O'_i, O'_j) = 0$, oder man ignoriert das Vorzeichen bzw. die Phasenlage und iteriert weiter. Auf alle Fälle leidet die Interpretierbarkeit des FastMap-Ergebnisses. Wang et al. (2000) schlugen ein Kompromissverfahren vor, das Fastmap in einigen Bereichen überlegen ist.

Fazit: Fastmap lässt an Zulässigkeit und Zuverlässigkeit zu wünschen übrig, ist aber ein bewährtes Verfahren, wenn es darum geht sehr große Datenvolumina schnell zu bewältigen.

Ein Anwendungsbeispiel aus dem Bereich der Wahrnehmungspsychologie zeigt Abb. 5.24. Probanden wurden in diesem klassischen Experi-

ment gebeten, 14 monochrome Farben auf ihre paarweise Ähnlichkeit auf einer 0–100-Skala einzuschätzen. Die Mittelwerte sind in Tab. 5.4 gegeben (Ekman 1954). Multi-dimensionale Skalierung ist hier das Verfahren der Wahl, um eine räumliche Anordnung der Farben zu finden. Die hufeisenförmige Gestalt in Abb. 5.24a entspricht der Linie der reinen Farben, die gleichzeitig den Rand der heutigen CIE-xy-Farbnormtafel bilden (s. Abb. 5.24 unten, sowie Abs. 2.4.3, CIE 1931).

Grundsätzlich ist die MDS-Einbettung nicht eindeutig bzgl. Rotation und Translation, während die FastMap sukzessiv die gefundenen Pivotelemente den Achsen zuordnet (s. horizontale Mittelachse 490:674 in Abb. 5.24b). Die FastMap liefert ein auffallend „unebeneres“ Bild als MDS, da sie reine Projektionen mittels der (hier) vier Pivotelemente ausführt und alle anderen Informationen unberücksichtigt lässt.

Um eine Startkonfiguration für iterative Verfahren wie z.B. SOM und MDS zu finden, ist Fastmap aber sehr gut geeignet (für eine Anpassung an den hyperbolischen Raum für HSOM und HSOM siehe Abs. 8.6). Hier spielen die Nachteile keine Rolle und es genügt, schnell und effektiv eine niedrigdimensionale Vorstrukturierung zu erlangen. Die Feinstrukturierung ist hier dem Zielverfahren überlassen, aber die Grobstrukturierungsphasen lassen sich durch Integration dieser „Starthilfe“ erheblich verkürzen.

Kapitel 6

Grundlagen hyperbolischer Geometrie

In Kap. 3 wurden verschieden Verfahren zur Visualisierung und Exploration erläutert. Alle sind auf die eine oder andere Weise durch die zur Verfügung stehende Anzeigefläche limitiert. Im euklidischen, zweidimensionalen Raum wächst bekanntermaßen die Fläche nur quadratisch mit dem Radius, was für das Layout der Darstellung ein großes Hemmnis werden kann. Ein Schlupfloch bietet der hyperbolische Raum.

Wie folgender Exkurs darlegt, wächst die Fläche im zweidimensionalen hyperbolischen Raum exponentiell mit dem Radius, er erscheint „intensiver unendlich“ und bietet sehr ansprechende Möglichkeiten zur Kontexterhaltenden Visualisierung und interaktiven Navigation.

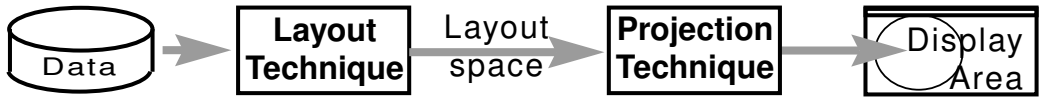


Abbildung 6.1: Visualisierung mit begrenztem Platz ist (mindestens) ein zweistufiger Prozess. Nach Datenvorauswahl und Transformation wird ein Layout generiert und interaktiv steuerbar zur Anzeige gebracht. Meist sind Ausschnitt und Auflösung steuerbar (s.a. Abs. 3).

6.1 Geschichte

Vor etwa 2300 Jahren brauchte der griechische Mathematiker Euklid fünf Axiome, um seine komplette Geometrie zu gründen.

A1: Durch zwei Punkte P1, P2 geht genau eine Verbindungs**gerade** G. Sie ist die eineindeutig kürzeste Verbindungslinie;

A2: G kann beliebig weit fortgesetzt werden (= allgemeine Gerade);

A3: Um jeden Punkt kann man einen Kreis mit beliebigem Radius legen;

A4: Alle rechten Winkel sind kongruent zueinander;

A5: Wenn eine Gerade zwei andere Geraden schneidet und die beiden Schnittwinkel auf einer Seite zusammen weniger als zwei rechte Winkel ergeben, dann scheiden sich die Fortsetzungen der beiden Geraden auf dieser Seite.

Das letzte Axiom ist auffallend kompliziert und kann in das äquivalente **Parallelenaxiom** überführt werden:

A5': Durch einen Punkt P3 außerhalb einer Geraden G (durch P1, P2) gibt es genau eine **Parallele**, i.e. eine Gerade, die G nicht schneidet.

Alle wesentlichen Elemente der Geometrie, wie wir sie in der Schule lernen, hat Euklid in seinem Buch „Die Elemente“ beschrieben: Linien, Kreise, Quadrate etc. Das Axiom A5 bemühte er selten, und etliche Mathematiker glaubten an einen Irrtum Euklids. Sie waren intensiv bemüht, die Ableitbarkeit – und damit die Entbehrlichkeit – des **Parallelenpostulates** nachzuweisen. Es war vergeblich. Fast zweittausend Jahre später entdeckten die Mathematiker Nicolai Lobachevski (1793–1856), János Bolyai (1802–1860) und Karl Friedrich Gauß (1777–1855) unabhängig voneinander neue, nicht-euklidische Geometrien. Sie beschreiben gekrümmte Räume, in denen alle Axiome gelten – nur nicht das fünfte, A5. Betrachtet man die Räume mit konstanter Krümmung (ungleich Null), kann man genau zwei nicht-euklidische Geometrien unterscheiden: die **sphärische Geometrie** mit positiver Krümmung und ihr Gegenstück, die **hyperbolische Geometrie** mit negativer Krümmung.

2D	**Sphärische** Geometrie S^2	**Euklidische** Geometrie $\mathbb{R}^2$	**Hyperbolische** Geometrie $\mathbb{H}^2$
Krümmung	positiv	null	negativ
Winkelsumme im Δ	$> 180^o$	180^o	$< 180^o$
... rel. zur Ebene	größer als ...	=	kleiner als ...
A5 # Parallelen (*)	0	1	∞
Anwendung	Kartographie	Standard	Relativitätstheorie
Metrik (**)	$ds^2 = dx^2 + dy^2 + dz^2$	$ds^2 = dx^2 + dy^2$	$ds^2 = dx^2 + dy^2 - dw^2$
Isometrische 3D	$x = sin(\theta)cos(\phi)$		$x = sinh(\theta)cosh(\phi)$
Einbettung	$y = sin(\theta)sin(\phi)$		$y = sinh(\theta)sinh(\phi)$
Polarkoord. (**)	$z = cos(\theta)$		$w = cosh(\theta)$
Entwürfe ... zur ... Abbildungen ... auf die Ebene (u.a.)	Stereographische P., Mercator P., P. nach Wollweide, Eckert, Lambert, ...	1:1	Minkowski (*), Obere Halbebene, Klein-Beltrami, Poincaré

Tabelle 6.1: Eigenschaftsvergleich zwischen zwei-dimensionalen Geometrien mit uniformer Krümmung (* s. a. Abb. 6.8; ** bezieht sich auf die Minkowski-metrik).

Sphärische Geometrie in zwei Dimensionen ist unserer Wahrnehmung und unserer Vorstellungskraft wohlvertraut. Sie beschreibt die Oberflächeneigenschaften einer 3D-Kugel, z.B. der Erdoberfläche oder einer Orangenschale. Die „Verbindungsgeraden“ zwischen zwei Weltpunkten P1, P2 sind die **Geodäten** und liegen auf Großkreisen G (wie der Äquator). Sie entsprechen den kürzesten Punkt-zu-Punkt Flugrouten um den Globus. Ein Dreiecksflug Richtung Südpol mit zwei 90° Kurven nach je einem Viertelkreisbogen würde sich mit einer Winkelsumme von 270° schließen. Dies widerspricht der Erwartung eines Euklidischen Raumes mit einer Winkelsumme von genau 180°. Auch kann man sich leicht vergewissern, dass das Parallelenpostulat nicht gilt: denn alle Großkreise durch P3 ($\notin$ G) schneiden G zweimal. Damit gibt es keine Parallele durch P3.

Heute kennen wir Satellitenaufnahmen der Erde oder phantastische Bilder anderer Himmelskörper und es fällt uns nicht leicht, die historischen Widerstände nachzuvollziehen, die mit der Aufgabe der Weltvorstellung einer flachen Scheibe verbunden war. Dieser Schritt war leicht, sobald wir den „globalen“ Blick bekamen. Der nächste ist schwerer. Die spezielle Relativitätstheorie hat uns gelehrt, dass die Universalität der Lichtgeschwindigkeit ein vierdimensionales Raum-Zeit-Kontinuum nahelegt. Im Minkowski-Raum wird mittels einer imaginären Zeitachse die Beschrei-

bung in allen Inertialsystemen geschlossen formulierbar. Eine Hauptidee ist die Betrachtung der räumlich radialen Lichtausbreitung als raum-zeitliche Kegelflächen mit der Norm 0 ($dr^2 - c^2dt^2 = 0$). In der allgemeinen Relativitätstheorie ging Einstein noch einen Schritt weiter und legte die Krümmung des Raumes durch Gravitationswirkung dar.

Beides entzieht sich unserer unmittelbaren Wahrnehmung, da wir uns viel zu langsam bewegen und keinem nennenswerten Gravitationsgradienten ausgesetzt sind. Unser Raum erscheint euklidisch und ist doch gekrümmt. Nicht zuletzt deshalb ist die Betrachtung nicht-euklidischer Geometrien bedeutungsvoll. Einige Eigenschaften sind kurios und ungewohnt, aber eröffnen auch eine neue Welt.

Die Krümmung von Flächen und des Raumes lässt sich mit Methoden der Differentialgeometrie allgemein lokal bestimmen. Tabelle 6.1 betrachtet die einfachsten Fälle – zweidimensionale Räume mit konstanter Krümmung – und listet einige wichtige Unterschiede und Struktursymmetrien zwischen den Geometrietypen auf.

6.2 Wie kann man sich ein Bild vom hyperbolischen Raum machen?

Es liegt in der Natur der gekrümmten Räume, dass sie sich einer optimalen Abbildung in einem flachen Raum widersetzen. Gewünscht sind folgende Abbildungseigenschaften:

- Geradentreue: Geraden bleiben Geraden;
- Konformität, Winkeltreue: Schnittwinkel zwischen Geraden bleiben erhalten;
- Längentreue: Längenverhältnisse bleiben erhalten;
- Flächentreue: Flächenverhältnisse bleiben erhalten.

Diese Anforderungen können nicht simultan erfüllt werden. Daher gibt es zahlreiche „Abbildungsentwürfe“ mit verschiedenen Eigenschaften. Zur Anfertigung von Karten der Erdoberfläche, zum Beispiel, sind u.a. die orthographische Projektion, die Projektionen nach Mercator, vielleicht auch

nach Lambert oder Mollweide bekannt. Mindestens eine der obigen Anforderungen ist unerfüllbar.

Im zwei-dimensionalen hyperbolischen Raum $\mathbb{H}^2$ gibt es ganz analog wie im sphärischen Raum S^2 mehrere Abbildungsentwürfe, auch Karten oder Modelle genannt. Die fünf wichtigsten analytischen Modelle und ihre mnemonischen Abkürzungen sind:

M, das **M**inkowski-Modell;

K, das **K**lein-Beltrami-Modell;

S, das Halb**s**phären-Model (auch *Jemi**s**phere model*);

P, das **P**oincaré-Modell (auch *disk model*);

H, das Modell der oberen **H**albebene (*upper-half-plane model*).

Jedes ist auf einer Teilmenge des $\mathbb{R}^{n+1}$, seiner Domäne, definiert, die auch die Namensgebung von H und S erklärt:

$$\begin{aligned} M = &\ \{(x_1,\ldots,x_n,x_{n+1}) &&: x_1^2+\ldots+x_n^2-x_{n+1}^2=-1 \wedge x_{n+1}>0\};\\ K = &\ \{(x_1,\ldots,x_n,\ 1\ ) &&: x_1^2+\ldots,x_n^2<1\};\\ S = &\ \{(x_1,\ldots,x_n,x_{n+1}) &&: x_1^2+\ldots+x_{n+1}^2=1 \wedge x_{n+1}>0\};\\ P = &\ \{(x_1,\ldots,x_n,\ 0\ ) &&: x_1^2+\ldots+x_n^2<1\};\\ H = &\ \{(\ 1,x_2,\ldots,x_{n+1}) &&: x_{n+1}>0\}; \end{aligned} \tag{6.1}$$

Abb. 6.2 illustriert den eindimensionalen Fall $n=1$, bzw. im mehrdimensionalen Fall eine Projektion von x_1, x_{n+1}. Die Hilfslinien deuten die Isometrien zwischen den Modellen an. Ausgehend von S als zentralem Hilfsmodell sind alle anderen Modelle durch stereographische oder orthographische Projektionen von und nach S zu erreichen. Die wichtigsten konkreten Abbildungen seien hier vorgestellt:

- die Karte $\phi_{MS}: M \to S$ von der Hyperbel M (bzw. Rotationshyperboloid) zur Halbkugel ist eine Zentralprojektion vom „Südpol“ $(0,\ldots,0,-1)$ aus

$$\phi_{MS}: M \to S, \quad (x_1,\ldots,x_{n+1}) \mapsto \left(\frac{x_1}{x_{n+1}},\ldots,\frac{x_n}{x_{n+1}},\frac{1}{x_{n+1}}\right); \tag{6.2}$$

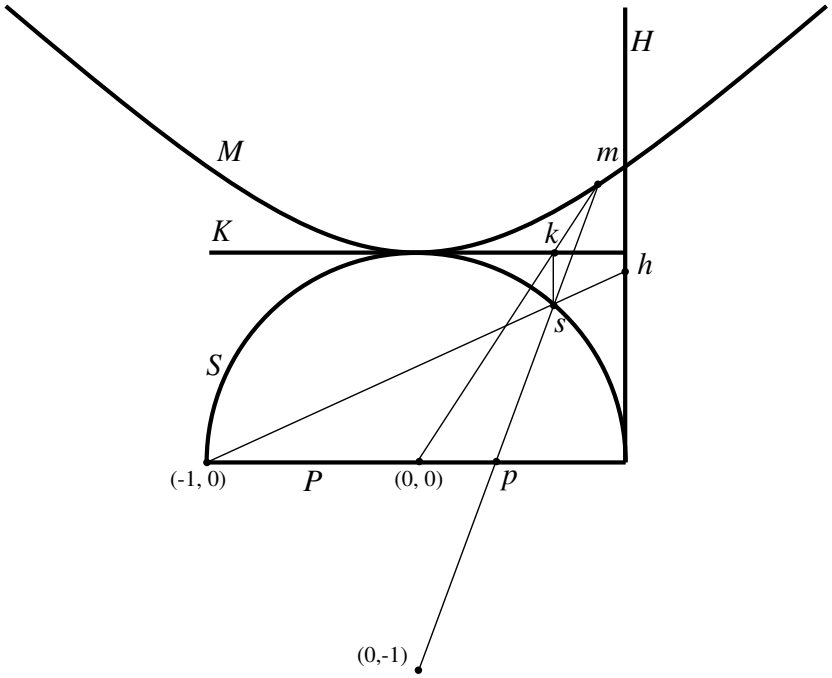


Abbildung 6.2: Fünf analytische Modelle des ein-dimensionalen hyperbolischen Raumes und ihre verbindenden Isometrien. Die Punkte $m \in M$, $p \in P$, $s \in S$, $k \in K$ und $h \in H$ stellen den selben Punkt im synthetischen $\mathbb{H}^1$ dar. (**M**inkowski-Modell, **K**lein-Beltrami-Modell, Halb-**S**phären-Model, **P**oincaré-Modell und das Modell der oberen **H**albebene)

- die Karte $\phi_{SP} : S \to P$ zur Poincaré-Scheibe ist auch eine Zentralprojektion vom „Südpol“ $(0, \ldots, 0, -1)$

$$\phi_{SP} : S \to P, \quad (x_1, \ldots, x_{n+1}) \mapsto \left(\frac{x_1}{x_{n+1}+1}, \ldots, \frac{x_n}{x_{n+1}+1}, 0 \right); \quad (6.3)$$

- die Karte $\phi_{SK} : S \to K$ zur Klein-Beltrami-Scheibe ist eine orthographische, vertikale Projektion auf die Höhe $x'_{n+1} = 1$

$$\phi_{SK} : S \to K, \quad (x_1, \ldots, x_{n+1}) \mapsto \left(x_1, \ldots, x_n, \sqrt{1 - x_1^2 - \cdots - x_n^2} \right); \quad (6.4)$$

- die Karte $\phi_{SH} : S \to H$ zur *upper-half-plane* ist eine Zentralprojektion vom „Westpunkt“ $(-1, 0, \ldots, 0)$ auf die Halbebene H

$$\phi_{SH} : S \to H, \quad (x_1, \ldots, x_{n+1}) \mapsto \left(1, \frac{2x_2}{x_1+1}, \ldots, \frac{2x_{n+1}}{x_1+1} \right). \quad (6.5)$$

6.3 Metriken für die fünf hyperbolischen Modelle

Alle fünf analytischen Modelle sind differenzierbare Mannigfaltigkeiten mit Riemann'scher Metrik und damit messbaren Längen, Winkeln und Flächeninhalten. Allgemein muss man beim Übergang zwischen krummlinigen Koordinatensystemen bei Messungen die lokalen Verzerrung durch zweifache Tensorfelder berücksichtigen:

$$ds^2 = d\mathbf{x}^T \mathbf{G}\, d\mathbf{x} = \sum_{ij} g_{ij} dx_i dx_j. \quad (6.6)$$

Längenelemente werden als $ds = \sqrt{ds^2}$ und Flächenintegrale mittels $\int_a^b dA = \int_a^b \sqrt{\det \mathbf{G}} d\mathbf{x}$ berechnet. Wie sehen die assoziierten Riemann'schen Metriken für die fünf Modellen aus?

$$ds_M^2 = dx_1^2 + \cdots + dx_n^2 - dx_{n+1}^2 \quad (6.7)$$

$$ds_S^2 = \frac{dx_1^2 + \cdots + dx_{n+1}^2}{x_{n+1}^2} = \left(\frac{||d\mathbf{x}||}{x_{n+1}} \right)^2 \quad (6.8)$$

$$ds_K^2 = \frac{dx_1^2 + \cdots + dx_n^2}{1 - dx_1^2 - \cdots - dx_n^2} + \frac{x_1 dx_1 + \cdots + x_n dx_n}{(1 - dx_1^2 - \cdots - dx_n^2)^2} \quad (6.9)$$

$$ds_P^2 = 4\frac{dx_1^2 + \cdots + dx_n^2}{(1 - x_1^2 - \cdots - x_n^2)^2} = \left(\frac{2\,||d\mathbf{x}||}{1 - ||\mathbf{x}||^2}\right)^2 \quad (6.10)$$

$$ds_H^2 = \frac{dx_2^2 + \cdots + dx_{n+1}^2}{x_{n+1}^2} = \left(\frac{||d\mathbf{x}||}{x_{n+1}}\right)^2 \quad (6.11)$$

Alle fünf Metriken beschreiben Varianten des hyperbolischen Raumes, wobei man an ds_M^2 (Gl. 6.7) den Bezug zur Hyperbel (in der Grundform $x^2 - y^2 = 1$) am deutlichsten erkennnen kann.

6.3.1 Ein weitere $\mathbb{H}^2$ Einbettungen in den $\mathbb{R}^6$

Eine isometrische Einbettung der hyperbolischen Ebene $\mathbb{H}^2$ in $\mathbb{R}^2$ oder $\mathbb{R}^3$ gibt es nicht – dies wusste schon Riemann. Aber gibt es vielleicht eine in noch höheren Dimensionen? Eine isometrische Konstruktion wurde von Blanusa (1955) vorgeschlagen. Sie bildet den Koordinatenpunkt (u, v) im den sechsdimensionalen Raum ab:

$$\Phi : \ (u, v) \mapsto \mathbf{x} = (x_1, \ldots, x_6) \in \mathbb{R}^6 \ .$$

Die Einzelkomponenten

$$\begin{aligned} x_1 &= x_1(u) \\ x_2 &= f_1(u)\sin(v\,\psi_1(u)) \\ x_3 &= f_1(u)\cos(v\,\psi_1(u)) \\ x_4 &= f_2(u)\sin(v\,\psi_2(u)) \\ x_5 &= f_2(u)\cos(v\,\psi_2(u)) \\ x_6 &= v \end{aligned} \quad (6.12)$$

sind regelhaft aus den Hilfsfunktionen $x_1(u), f_1(u), f_2(u), \psi_1(u), \psi_2(u)$ und ferner $\phi_1(u), \phi_2(u)$ zusammengesetzt. Die zweite Koordinate v fungiert als Drehwinkel in vier senkrechten Richtungen x_2, x_3, x_4, x_5. Um Selbstdurchdringung zu verhindern, erzeugt sie in x_6 zusätzlich eine helikale Verschiebung. u hat komplexeren Einfluss auf die Hilfsfunktionen. Mit wachsendem $|u|$ erhöht sie (u.a.) über $\psi_i(u)$ schrittweise die Drehfrequenz:

$$\begin{aligned} \psi_1(u) &= \exp\left(2\left\lfloor\frac{|u|+1}{2}\right\rfloor + 5\right) \\ \psi_2(u) &= \exp\left(2\left\lfloor\frac{|u|}{2}\right\rfloor + 6\right) \end{aligned} \quad (6.13)$$

Hierbei bezeichnet $\lfloor x \rfloor$ den abgerundet ganzzahligen Anteil von x und Gl. 6.13 erzeugt Treppenfunktionen. Die Diskontinuitäten an den Stufen ($\forall u \,|\, u = \lfloor u \rfloor$) werden an anderer Stelle geglättet.

Zwei weitere Funktionen ϕ_i werden als normalisierte Stammfunktionen definiert:

$$\phi_1(u) = \frac{1}{A} \int_{x=0}^{u+1} F(x)\, dx \tag{6.14}$$

$$\phi_2(u) = \frac{1}{A} \int_{x=0}^{u} F(x)\, dx \quad \text{mit} \tag{6.15}$$

$$F(x) = \sin(\pi x) \exp(-sin^{-2}(\pi x)) \quad \text{und} \tag{6.16}$$

$$A = \int_{x=0}^{1} F(x) dx = 0.141327... \tag{6.17}$$

Die Funktionen ϕ_i sind nicht-negativ, periodisch und erfüllen die Eigenschaften $\phi_1 + \phi_2 = 1$ und $\phi_1(u) = \phi_1(u+2) = \phi_2(u+1)$. Sie verhalten sich ähnlich wie perfekt geglättete Versionen von $|\sin(\pi u)|$ und $|\cos(\pi u)|$, i.e. an den Nulldurchgängen sind *sämtliche* Ableitungen ebenso Null.

Die Funktionen f_1, f_2 sind als

$$f_i(u) = \frac{\sinh(u)}{\psi_i(u)} \sqrt{\phi_i(u)} \tag{6.18}$$

assembliert. Es bleibt x_1, das als Integral definiert ist:

$$x_1(u) = \int_0^u \left(1 - f_1'(u)^2 - f_2'(u)^2\right)\, du. \tag{6.19}$$

Damit wird insgesamt die gewünschte konstante, negative Krümmung mit der Metrik $ds^2 = du^2 + \cosh^2(u) dv^2$ erzeugt, was den zweidimensionalen Raum von (u, v) zur hyperbolischen Ebene $\mathbb{H}^2$ macht – stetig und glatt eingebettet in $\mathbb{R}^6$.

Abb. 6.7 zeigt eine 3D-Projektion eines rechteckigen u, v-Gitterstückes. Das radial stufenweise Auftreten von weiteren Auffaltungen ist erkennbar ähnlich der Illustration in Abb. 6.7. Zu Visualisierungszwecken ist Blanusas funktionale Einbettungskonstruktion nicht direkt geeignet – ganz im Gegensatz zu den in Abs. 6.2 eingeführten geometrisch motivierten Modellen. Im Folgenden werden die wichtigsten darstellungsorentierten Eigenschaften insbesondere des Poincaré-Modells, vorgestellt.

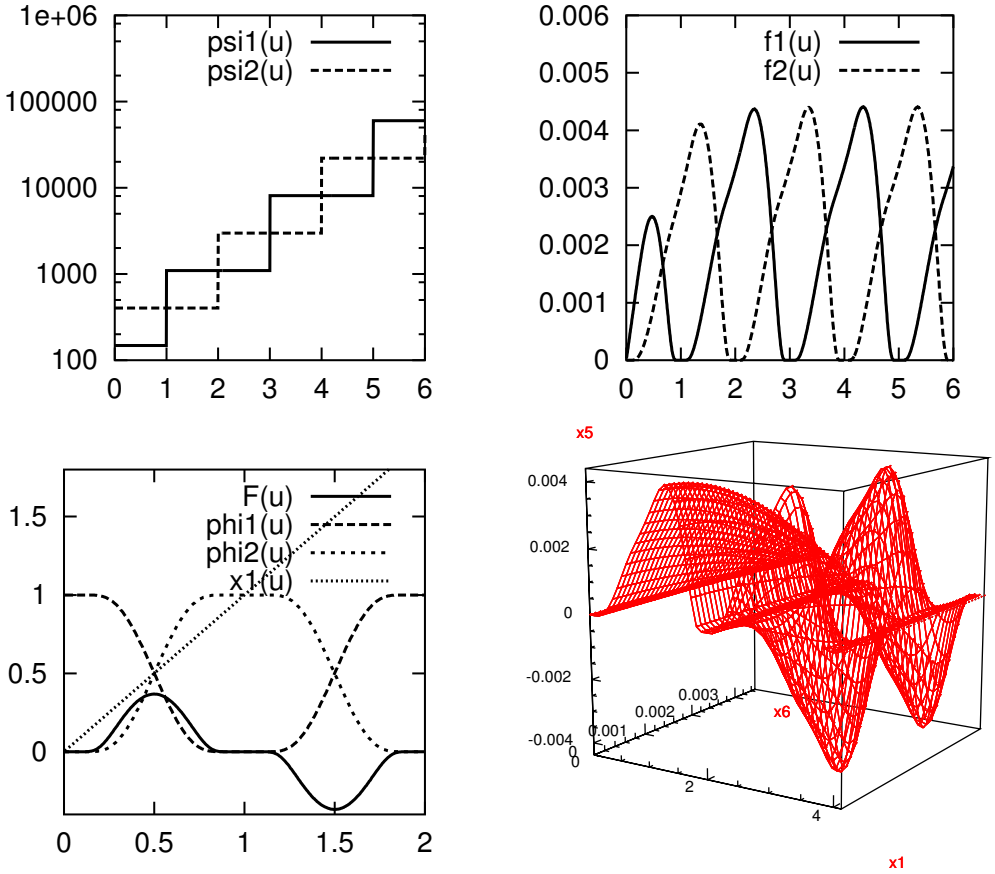


Abbildung 6.3: Teilaspekte Blanusa's Einbettung des $\mathbb{H}^2$ in den $\mathbb{R}^6$ mit einem 50×40 Gitter ((x_1, x_5, x_6) mit $u \in [0, 4.2], v \in [0, 0.1]$).

6.4 Eigenschaften des $\mathbb{H}^2$: Geodäten, Flächen etc.

Die Riemann'sche Distanz $\overline{P_1 P_2}$ ist das kürzestmögliche Pfadintegral zwischen den Punkten P_1 und P_2. Dieser Pfad ist immer Teil einer geodätischen Linie. Wie sehen diese **Geodäten** aus? In der Ebene sind es die geraden Linien. Jede Pfadabweichung von der $P_1 P_2$-Geraden verlängert den Weg. Die Geodäten der Kugel sind die Großkreise (z.B. der Äquator und die Längenkreise, s. a. Abb. 6.8).

Abb. 6.4 und 6.5 illustrieren eine hyperbolische Geodäte und verdeutlichen den projektiven Zusammenhang der fünf Modelle M, K, S, P und

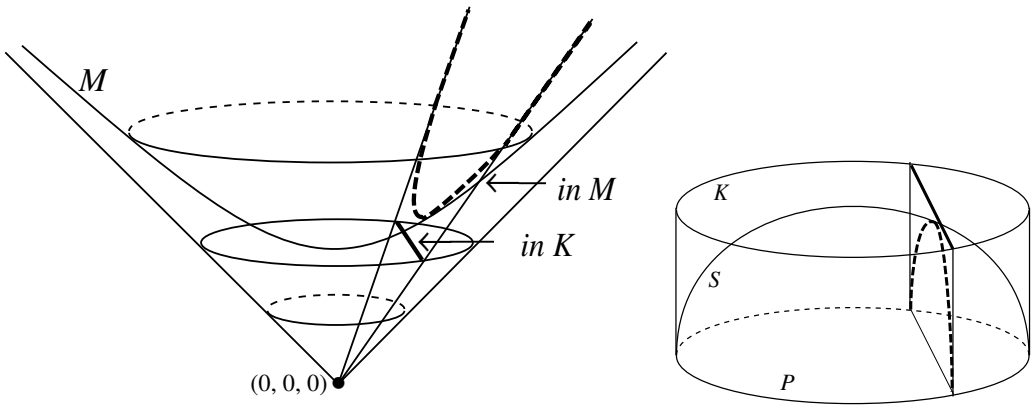


Abbildung 6.4: (*a, links:*) Eine Geodäte oder auch „$\mathbb{H}^2$-Gerade“ ist als breite Linie in M, K, S gezeigt. Im Minkowski-Modell M und in der Klein-Beltrami-Scheibe K entstehen sie durch Ebenenschnitte, die den Ursprung einschließen. Die kürzesten Verbindungen und damit die $\mathbb{H}^2$-Geraden stellen sich als Hyperbeln in M dar. Das Klein-Beltrami-Modell K ist offensichtlich geradentreu. (*b, rechts:*) Im Halbkugelmodell S sind die Geodäten vertikale Halbkreisbögen.

H. Im Gegensatz zum eindimensionalen Fall in Abb. 6.2 wird nun der $\mathbb{H}^2$ dargestellt. Anhand dieser $\mathbb{H}^2$-“Geraden“ lassen sich eine Reihe von Eigenschaften der hyperbolischen Geometrie aufzeigen.

Für die Visualisierung ist das **Poincaré-Kreisbild** P am interessantesten. Der Grund dafür wird in den nächsten Abschnitten erläutert.

Displaygerecht: Die unendlich große Fläche des $\mathbb{H}^2$ wird auf eine feste Kreisscheibe in K und P abgebildet (natürlich nicht längen- oder flächentreu). Dieser Umstand faszinierte den vielseitigen Künstler Maurits Escher und veranlasste ihn zu einigen Darstellungen eines kompakten Unendlichen, u.a. Abb. 6.6;

Rand $= \infty$***:*** Alle fernen Punkte sind nahe an den Kreisrand in P gedrängt, ohne ihn zu berühren;

Fovea: Die Punkte nahe der Vertikalachse (nicht in H) werden, wie mit einem „Fischaugen“-Objektiv, auf P abgebildet (mit Vergrößerung mittig 0.5);

Fokus $+$ ***Kontext:*** Der „Fokus“ kann (wie unten gezeigt) wie eine „Fovea“ an jeden Ort verschoben werden. Der Kontext wird mit dargestellt,

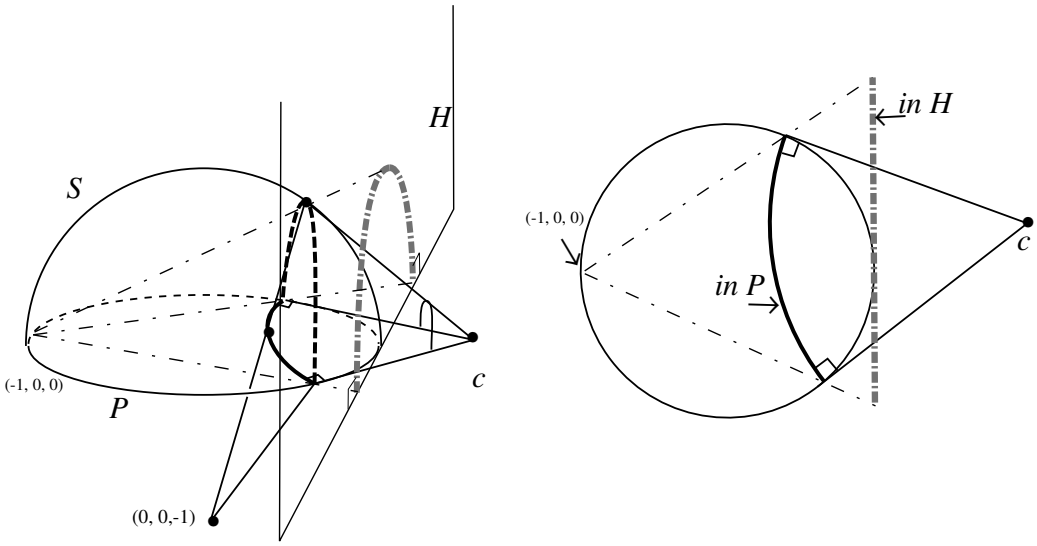


Abbildung 6.5: (*a, links:*) Eine $\mathbb{H}^2$-Geodäte ist als breite Linie im S-, P- und H-Modell dargestellt. Die Zentralprojektion ist winkeltreu und überführt Kreise in Kreise. Die S-*Gerade*, der vertikale Halbkreisbogen, wird in ein Kreisbogensegment in der Poincaré-Scheibe P projiziert (Zentralpunkt Südpol). Da die 90°-Winkel erhalten bleiben, wird die Kreisscheibe rechtwinklig erreicht. c ist der Kreissegmentmittelpunkt in P und gleichzeitig die Spitze des die Halbkugel S tangierenden Kegels. Die graue, unterbrochene Linie ist das Pendant der Geodäte in der obereren Halbebene H (vgl. Abb. 6.2). Durch Zentralprojektion entsteht wiederum ein vertikaler Halbkreis, nun über der Grundlinie $x_1 = 1$. (*b, rechts:*) orthogonale Projektion von x_1x_2. Sonderfälle: *(i)* Außer den Kreissegmenten in P, die den Rand senkrecht schneiden, sind auch Geraden durch den Mittelpunkt Geodäten. Sie werden auch „Verallgemeinerte Kreise“ genannt, mit unendlichem Radius und Mittelpunktabstand; *(ii)* auch senkrechte Halbgeraden in H sind Geodäten (neben den Halbkreisbögen).

mit wachsendem „Fovea“-Abstand zunehmend vergröbert;

Geraden werden P-Kreisbögen: Alle $\mathbb{H}^2$-Geraden werden auf P-Kreise oder P-Geraden abgebildet, die alle den Einheitskreis senkrecht schneiden. P-Geraden senkrecht zum Rand gehen durch die P-Kreismitte (0,0). Oft erweitert man den Kreisbegriff zum *verallgemeinerten Kreis*, der Geraden als Kreise mit unendlich großem Radius und unendlich weitem Mittelpunkt einschließt. Wie auch in Abb. 6.5 dargestellt, sind die Projektionen der Geraden in die obere Halbebene H ebenfalls H-Halbkreisbögen oder H-Vertikale;

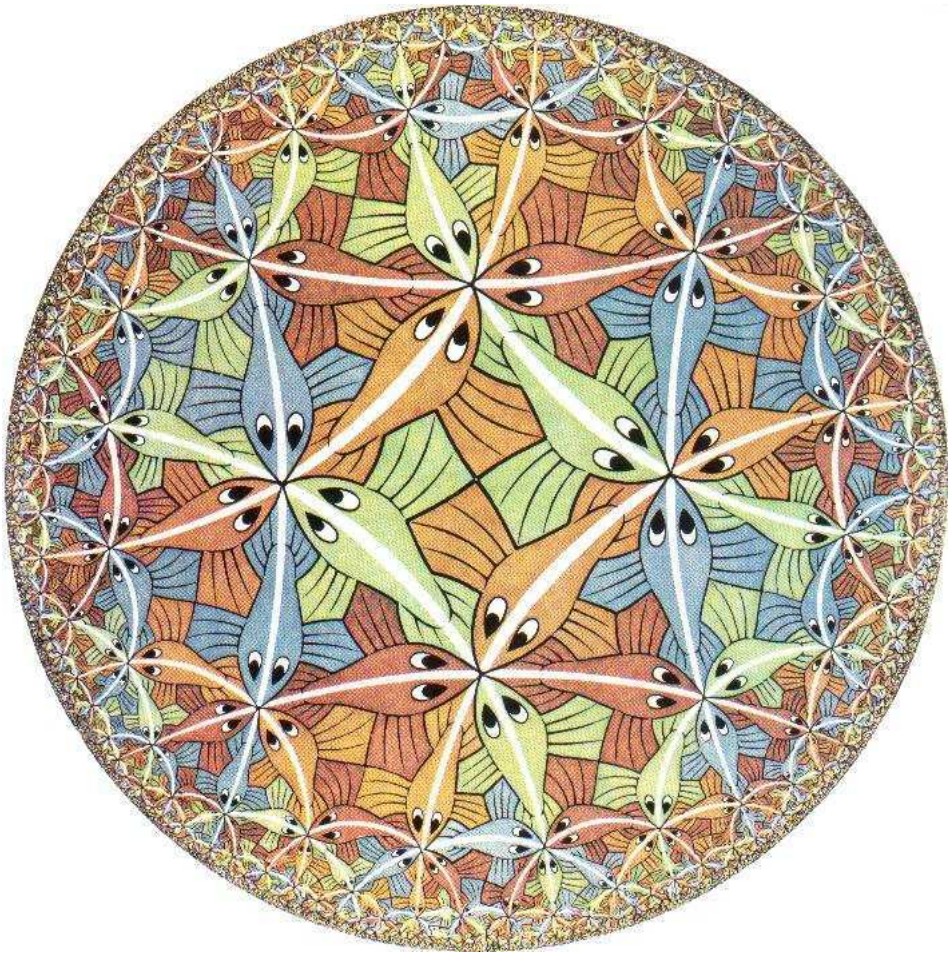


Abbildung 6.6: Holzschnitt von Maurits C. Escher, „Kreislimit III“ (1958). Die Möglichkeit, kompakt und vollständig das Unendliche darzustellen, faszinierte ihn, als er Bilder des Poincaré-Modells des $\mathbb{H}^2$ in Coxeters Buch sah (1957). Alle weißen Linien schneiden senkrecht den Rand, der unendlich weit weg ist. Man beachte den “Fischaugen“-Effekt: Alle Fische sind im $\mathbb{H}^2$ tatsächlich gleich groß, aber je weiter außen sie liegen, um so weniger Blickwinkel nehmen sie ein. (Mit freundlicher Genehmigung von Cordon Art - Baarn - Holland)

Konforme Abbildung in P, H, S: Bei Projektion bleiben Kurvenschnittwinkel erhalten, d.h. die Abbildungsentwürfe P, H und S sind *konform* (im Gegensatz zum Klein-Beltrami-Modell K, das stattdessen geradentreu ist). Der ∞-Rand wird von allen Geodäten „winkeltreu“, also senkrecht erreicht. Kreise, i.e. Punkte gleichen Abstands zu P sind auch Kreise, allerdings ist i.A. P nicht der Kreismittelpunkt, sondern (vom Ursprung radial) nach außen verschoben;

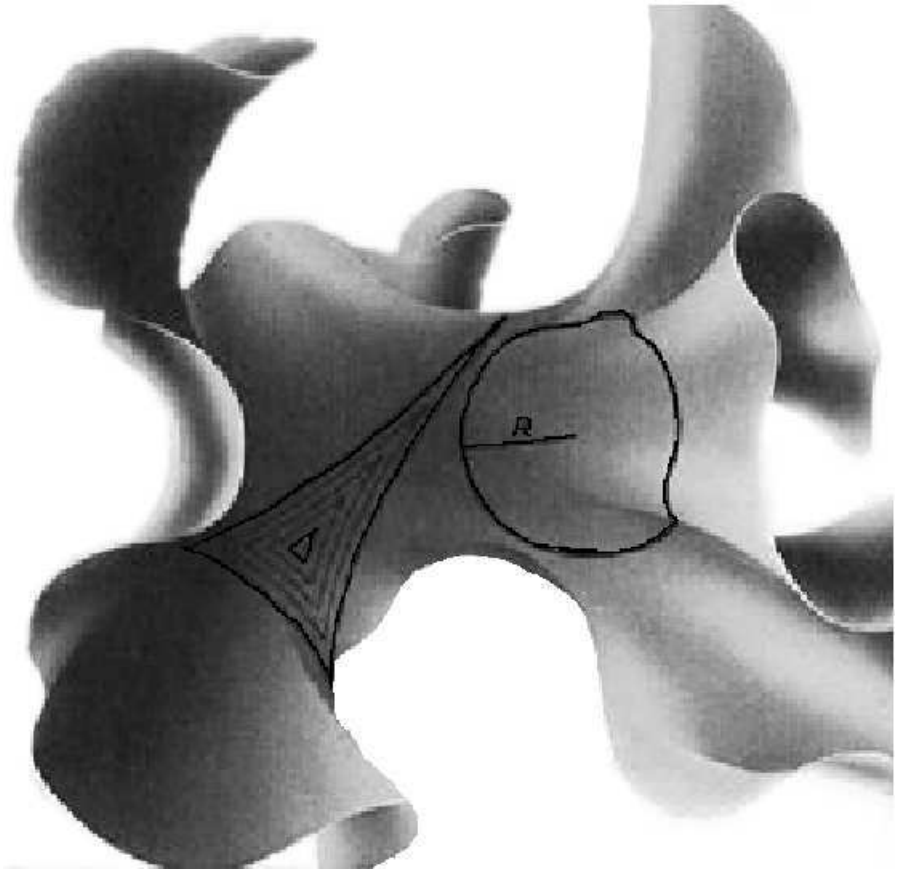


Abbildung 6.7: Die lokale Einbettung des $I\!H^2$ als $\mathbb{R}^3$-Ansicht: Der hyperbolische Raum bietet mehr Platz als der euklidische. (*Links:*) Je größer die konzentrischen Dreiecke, umso spitzer werden sie und die Winkelsumme kleiner 180° nimmt ab. (*Rechts:*) In einer kleinen Nachbarschaft um einen Punkt ist die Struktur flach. Wächst der Radius R eines Kreises, werden mehr und mehr „Falten“ des hyperbolischen Raumes eingeschlossen, was zum exponentiellen Wachsen der Fläche (Gl. 6.26) und des Umfanges führt.

Winkeldefizit und „Zipfelkissen“: Ein Dreieck (in den konformen $\mathbb{H}^2$-Modellen) weist eine Winkelsumme von stets kleiner 180° auf. Der $\mathbb{H}^2$-Flächeninhalt ist sogar durch dieses Winkelsummendefizit definierbar (Gauß-Bonnet-Theorem). In Abb. 6.6 ist jedes Dreieck, das den Ursprung enthält, durch eine konkave dritte Seite abgeschlossen, siehe auch Tab. 6.1 und Abb. 6.7. Ferner ist sowohl die Fläche eines Dreiecks als auch die Dicke (kürzester Abstand eines Randpunktes zu den gegenüberliegenden Seiten) beschränkt;

Parallelenaxiom A5’: Anhand von Abb. 6.8 lässt sich das Axiom A5 gut aufzeigen: Durch einen dritten Punkt P3, der nicht auf dem durch zwei Punkte P1 und P2 definierten Kreisbogen K liegt, lassen sich leicht beliebig viele andere Kreise zeichnen, die K *nicht* schneiden. Damit sind sie alle, per definitionem, Parallelen in H2.

Zwei Arten von Parallelen: Es gibt Kreise *(i)*, die sich in P nicht schneiden (Parallelen) und *(ii)* Kreise, die sich genau am P-Kreisrand berühren. Letztere sind auch Parallelen, sie sind *„asymptotisch-parallel“* – „mit gleicher ∞-Richtung“. Erstere sind *„divergent-parallel“* oder auch *„ultra-parallel“*. In Abb. 6.8 werden sie illustriert und mit der Situation in S^2 und $\mathbb{R}^2$ gegenübergestellt. Ein weiteres Indiz dafür, dass es viel „mehr“ Platz im Unendlichen gibt.

Tesselation mit Dreiecken: Es lassen sich gleichschenklige Dreiecke mit sehr kleinen Winkeln (alle identisch und <60°) konstruieren. Eine vollständige Tesselation der Fläche wird ermöglicht, indem man jeweils äquilaterale Dreiecke so fügt, dass an jedem Eckpunkt stets n Dreiecke zusammenkommen. Da der Innenwinkel $\angle = 360°/n < 180°/3 = 60°$ hier immer kleiner 60° ist, muss $n \geq 7$ sein (s. a. Abb. 7.5). Damit gibt es sehr vielfältige Möglichkeiten, den $\mathbb{H}^2$ durch reguläre Tesselation zu teilen (im Vergleich zur euklidischen Ebene, die nur $n = 3, 4, 6$ erlaubt, oder der Kugel mit $n = 3, 4, 5$).

Einbettungsmodell: Ein grobes Modell des $\mathbb{H}^2$ lässt sich z.B. aus Papier anfertigen: Nach obigem Strickmuster nehme man viele gleichseitige Dreiecke und hefte sie paarweise an den Kanten, so dass genau n an jeder Ecke zusammenstoßen. Es entsteht ein Gebilde, das an jedem Ort lokal eine Sattelstruktur aufweist. Eine glatte Einbettung eines Stückes $\mathbb{H}^2$ in $\mathbb{R}^3$ wird in Abb. 6.7 illustriert.

Der $\mathbb{H}^2$ ist „intensiver unendlich“ und wächst exponentiell:

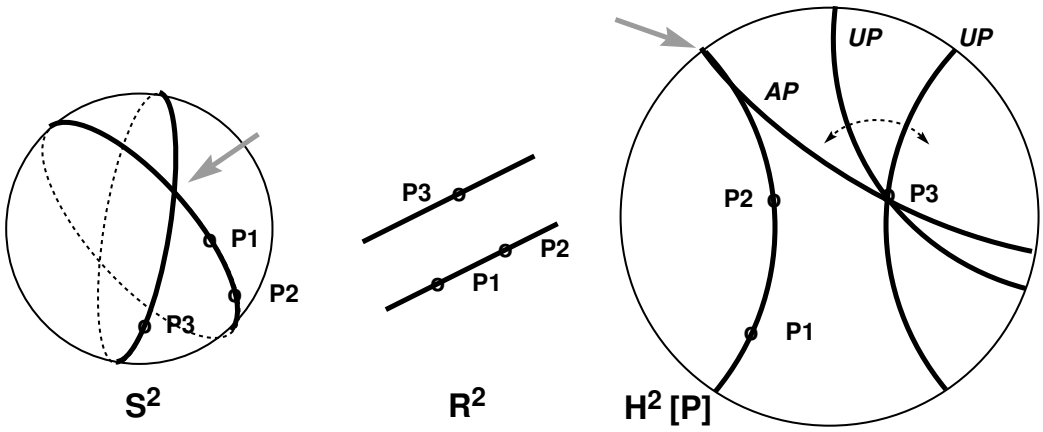


Abbildung 6.8: Das Parallelenaxiom in den drei uniform gekrümmten Räumen: *(a, links)* Auf der Kugeloberfläche S^2 findet sich kein Großkreis (=Gerade) durch P3, der den durch P1,P2 definierten Großkreis nicht schneidet. *(b, Mitte:)* Der flache, euklidische Raum $\mathbb{R}^2$ ermöglicht genau eine Parallele. *(c, Rechts:)* Im hyperbolischen $I\!H^2$ gibt es beliebig viele „Ultra-Parallelen“ („UP“). Die Geraden (im Poincaré-Bild Kreisbögen), die sich im selben Randpunkt treffen, heißen auch „asymptotische Parallelen“ („AP“).

Im Minkowski- oder Hyperboloidbild M wird deutlich, dass ein nach oben wachsender Hyperboloidstreifen auf immer weniger Platz nahe des Kreisrandes in P oder S projiziert wird (Abb. 6.4). In der ds_M^2 Metrik Gl. 6.7 kompensiert das Nach-oben-Wegelement $-d_{x+1}^2$ weitgehend das Nach-außen-Wegelement. Trotz Umfangwachstums gibt es hier radial kaum Zuwachs, s. a. Abb. 6.9.

Es stellt sich heraus, dass sowohl der Umfang als auch die Kreisfläche exponentiell mit dem Radius wachsen. Betrachten wir nun den zweidimensionalen Fall in P und transformieren das Abstandselement in Gl. 6.10 und das Flächenelement in Polarkoordinaten

$$ds = \frac{2}{1-r^2}\,dr\ , \qquad dA = \frac{4r}{(1-r^2)^2}\,dr\,d\theta. \tag{6.20}$$

Ausgehend vom Ursprung berechnen sich der hyperbolische Radius ρ, der Umfang C und die Fläche A als Integrale des euklidischen Radius R zu

$$\rho = \int_0^R \frac{2}{1-r^2}\,dr = \ln\frac{1+R}{1-R}; \tag{6.21}$$

$$C = \int_{\theta=0}^{2\pi} \frac{2R}{1-R^2}\,d\theta = \frac{4\pi R}{1-R^2} \tag{6.22}$$

$$A = \int_{\theta=0}^{2\pi}\int_{r=0}^{R} \frac{4r}{(1-r^2)^2}\, dr\, d\theta = \frac{4\pi R^2}{1-R^2}\,. \tag{6.23}$$

Mit Gl. 6.21 entsteht

$$R = \frac{e^\rho - 1}{e^\rho + 1} = \frac{\cosh\rho - 1}{\sinh\rho} \tag{6.24}$$

$$C = 2\pi \sinh\rho \tag{6.25}$$

$$A = 4\pi \sinh^2\frac{\rho}{2}\,. \tag{6.26}$$

Dieses Ergebnis hat zwei bemerkenswerte asymptotische Verhalten, die auch in Abb. 6.9 detailliert beschrieben sind:

- Für kleine hyperbolische Radii ρ wachsen der Umfang $C(\rho) \approx 2\pi\rho$ und die Kreisfläche $A(\rho) \approx \pi\rho^2$ wie in der euklidischen Ebene. Der Darstellungszoomfaktor $\partial R/\partial\rho$ ist 0.5;
- Für großes ρ wachsen $C(\rho)$ und $A(\rho)$ exponentiell, während R gegen 1 und der Zoomfaktor $\partial R/\partial\rho$ gegen 0 geht.

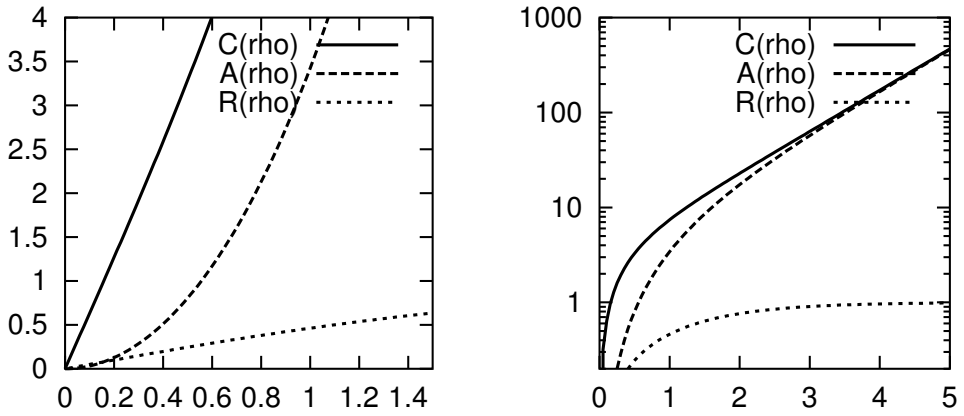


Abbildung 6.9: Entwicklung von $C(\rho)$, $A(\rho)$ und $R(\rho)$ in (*links*) linearer und (*rechts*) halblogarithmischer Darstellung (s. Gl. 6.24 ff). Für kleine hyperbolische Radii ρ wachsen C und A euklidisch und für großes ρ exponentiell, während R sich asymptotisch der 1 nähert (wie C/A).

Für Visualisierungszwecke ist dieses exponentielle Wachstum des $\mathbb{H}^2$ ein Kernvorteil, denn es gibt ideale Platzverhältnisse für, zum Beispiel,

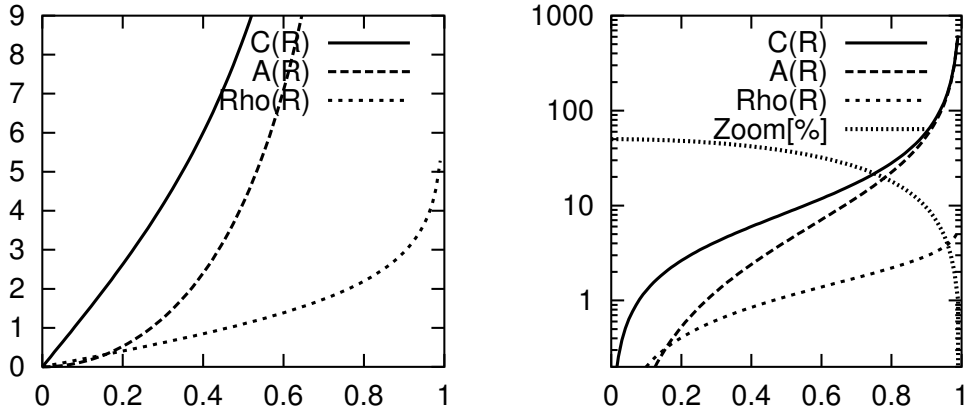


Abbildung 6.10: Entwicklung von $C(R)$, $A(R)$ und $\rho(R)$ in (*links*) linearer und (*rechts*) halblogarithmischer Darstellung (s. Gl. 6.21 ff). Im Gegensatz zu Abb. 6.9 in Abhängigkeit zum Darstellradius R in der Poincaré-Scheibe P (statt ρ). Es lassen sich grob drei Bereiche unterscheiden: *(i)* In einem Kreisbereich bis $R \leq 0.3$ wachsen C und A etwa euklidisch, *(ii)* in einem mittleren Kreisring ca. $0.3 \leq R < 0.85$ exponentiell. *(iii)* Im äußeren Kreisbereich, ca. $0.85 \leq R < 1$ wächst die Fläche A superexponentiell gegen ∞. Der Grund liegt im asymptotischen Grenzverhalten $R(\rho) \to 1$, s. a. Abb. 6.9. *(Rechts)* Der Abbildungsfaktor *Zoom* $= 100\, dR/d\rho$ in Prozent startet bei 50 % und fällt parabolisch.

hierarchische Datenstrukturen: Für einen ausgewogen Baum wächst die Blätterzahl genau exponentiell mit der Verzweigungstiefe.

Superexponentielles Wachstum im Poincaré-Modell: In Abb. 6.10 werden die Darstellungsverhältnisse in der Poincaré-Scheibe P in Abhängigkeit vom Radius R analysiert. Neben den beiden Bereichen euklidisches und exponentielles Wachstum im $\mathbb{H}^2$ kann man in diesem Modell eine weitere, dritte Zone identifizieren:

- Im randnahen Kreisring mit etwa $0.85 \leq R < 1$ zeigt sich **superexponentielles Wachstum** des Platzes. Umgekehrt verschwindet die räumliche Auflösung der visuellen Darstellung.

Abbildung 6.11: Drei Zonen des radialen Flächenwachstums im Poincaré-Modell P: *innen* euklidisch, *Mitte* exponentiell, *außen* superexponentiell. Für Details siehe Kurvendiskussion Abb. 6.10.

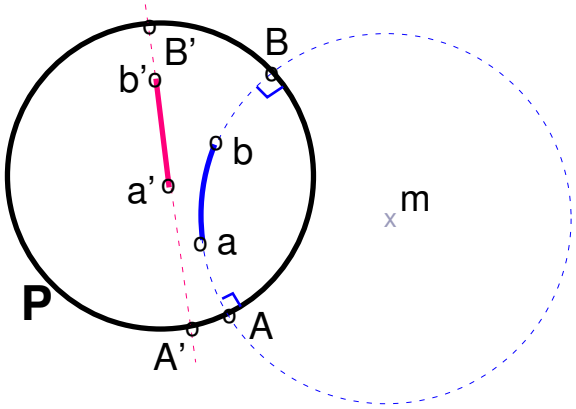


Abbildung 6.12: Der Abstand zweier Punkte im Poincaré-Modell ist über Abstandsdoppelverhältnisse definiert, siehe Gl. 6.27. Hier zwei Beispiele $d(a,b)$ und $d(a',b')$ und die Randpunkte ihrer Geodäten A, B und A', B'.

6.4.1 Längenmessung und -konstruktion im Poincaré-Modell

Den hyperbolische Abstand zweier Punkte $d(a,b)$ wird über Doppelverhältnisse mit Randabständen definiert

$$d(a,b) = \left| \ln \frac{|Aa|/|Ab|}{|Ba|/|Bb|} \right| . \tag{6.27}$$

Abb. 6.12 illustriert die Randpunkte A und B. Sie sind die ∞-Randpunkte der Geodäten, die a, b beinhalten, und $|Aa|$, $|Ab|$, $|Ba|$ und $|Bb|$ bezeichnen hierbei die euklidischen Punktabstände (Moise 1974).

Liegt der Punkt b im Ursprung, ergibt sich Gl. 6.21, und der Umkehrzusammenhang in Gl. 6.24 kann zum radialen Längenauftrag verwendet werden. Dieser Konstruktionsmechanismus ist universell einsetzbar, sobald Verschiebungen des Ursprungs zur Verfügung stehen (s. u.).

Möchte man häufig Datendistanzen im hyperbolischen Poincaré-Modell

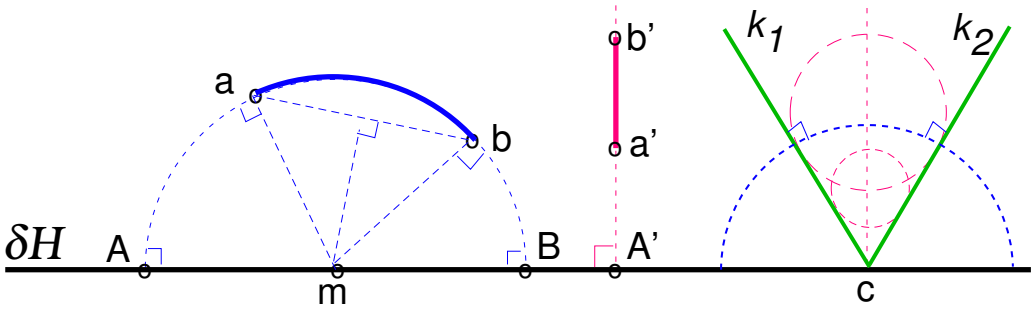


Abbildung 6.13: Das Modell der oberen Halbebene H hat einige Gemeinsamkeiten mit P: Geodäten sind entweder *(mitte:)* vertikale Halbgeraden oder *(links:)* Halbkreisbögen mit Mittelpunkt m auf dem ∞-Rand ∂H als Schnittpunkt mit der Mittelsenkrechten zweier Punkte $[ab]$. Der hyperbolische Abstand zweier Punkte $d(a,b)$ wird ebenso wie in P durch das Doppelverhältnis der Randpunktabstände (Gl. 6.27) bestimmt; *(Mitte:)* Bei Vertikalen ersetzt die 1 die Terme der fehlenden Randabstände $|Ba'|/|Bb'|$). *(Rechts:)* Hohe Symmetrie zeigen die Kurven gleichen Abstandes k_1, k_2: Zu jedem Punkt auf k_1 befindet sich der am nahe gelegenste Punkt in k_2 im festen Abstand (der mit dem Öffnungswinkel des euklidischen Kegels verknüpft ist). Im euklidischen Raum kennen wir die Parallelen als die Kurven gleichen Abstands – im $\mathbb{H}^2$ sind es (entgegen dem ersten Augenschein) jedoch keine Geraden.

berechnen, so ist eine geschlossene Form sehr nützlich (Strubecker 1969; Morgan 1993). Notiert man die P-Koordinaten x_0, x_1 als komplexe Zahl $z = x_0 + ix_1 \in \mathcal{C}$ innerhalb des Einheitskreises, ergibt sich eine elegant kompakte Formulierung

$$d(\mathbf{x}_i\,,\,\mathbf{x}_j) = 2\,\text{arctanh}\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{|1 - \mathbf{x}_i\bar{\mathbf{x}}_j|}\right)\;,\quad \mathbf{x}_i, \mathbf{x}_j \in P \in \mathcal{C}\,, \tag{6.28}$$

die später an verschiedenen Stellen nützlich sein wird.

6.4.2 Generator einer isotropen Datenverteilung im $\mathbb{H}^2$:

Um Verteilungseigenschaften von Daten im $\mathbb{H}^2$ zu studieren (s. Abs. 7.4), ist die Verfügbarkeit eines angepassten Datengenerators hilfreich.

Eine geschickte Herangehensweise ist die Erzeugung einer uniformen Verteilung auf einer zentrierten Kreisscheibe mit dem Radius R_{disk} und

der Fläche A_{disk}. Zunächst gilt es, die radiale Dichtefunktion bzw. die Verteilungsfunktion zu bestimmen und zu invertieren. Die Kreisfläche $A(R)$ in Gl. 6.23 ist proportional zu dieser Verteilungsfunktion. Mit Hilfe zweier (uniform verteilter) Zufallsvariablen $\nu_1, \nu_2 \in [0,1]$ kann der Generator für gleichverteilte Zufallspunkte innerhalb der Kreisscheibe in P wie folgt konstruiert werden:

$$\begin{aligned} x_0 &= r\cos(2\pi\nu_1) \\ x_1 &= r\sin(2\pi\nu_1) \quad \text{mit} \\ r &= \left(1+\frac{4\pi}{\nu_2 A_{disk}}\right)^{-\frac{1}{2}}, \end{aligned} \tag{6.29}$$

wobei $A_{disk} = 4\pi(R_{disk}^{-2}-1)^{-1}$ die Kreisfläche bemisst.

6.5 Die Isometrien des Poincaré-Modelles

Die form- und metrikerhaltenden Abbildungen im euklidischen Raum sind die Translation, die Rotation und Spiegelung. Wie sehen diese **Isometrien**, insbesondere im Poincaré-Kreisbild, aus? Die Spiegelung entspricht in den konformen $\mathbb{H}^2$-Modellen den Inversionen an Kugeln. Die beiden anderen lassen sich auch aus je zwei Kugelinversionen erzeugen. Isometrien müssen u.a. Geraden erhalten, d.h. die Kreise in P müssen Kreise bleiben. Auch hier erlaubt die komplexe Formulierung wieder eine eindrucksvoll kompakte Darstellung. Die Gruppe der **Möbius-Transformation**

$$T_{a,b}(z) = \frac{a\,z+\bar{b}}{b\,z+\bar{a}} \quad \text{mit} \quad |a|^2-|b|^2 = 1\,, \quad z,a,b \in \mathcal{C} \tag{6.30}$$

beschreibt die gesuchten Kreis-Automorphismen der Ebene und damit die Isometrien des Poincaré-Modells (und auch des Halbebenenmodells H). Für den Visualisierungszweck ist die Reparametrisierung in eine reine Translation und Rotation am effizientesten

$$z' = T(z;\ c,\theta) = \frac{\theta\,z+c}{\bar{c}\theta\,z+1}, \qquad z,c,\theta \in \mathcal{C}\ \wedge\ |\theta| = 1\ \wedge\ |c| < 1. \tag{6.31}$$

Das komplexe $\theta = e^{\imath\phi}$ beschreibt die Rotation in P um den Ursprung, gefolgt von einer Translation um c. Der Ursprung wird zum Punkt c und $-c$ wird das neue Zentrum $(0,0)$. Die inverse Abbildung braucht man, um die Koordinaten zurückzutransformieren:

$$T^{-1}(z;\ c,\theta) = T(z;\ -\bar{\theta}\,c,\bar{\theta}) \tag{6.32}$$

Zwei aufeinanderfolgende Transformationen von P werden beschleunigt durch

$$\begin{aligned} T(T(z;\ c_1,\theta_1);\ c_2,\theta_2) &= T(z,\ c,\theta), \qquad \text{mit} \\ c &= \frac{\theta_2 c_1 + c_2}{\theta_2 c_1 \bar{c}_2 + 1}, \\ \theta &= \frac{\theta_1\theta_2 + \theta_1 \bar{c}_1 c_2}{\theta_2 c_1 \bar{c}_2 + 1} \end{aligned} \tag{6.33}$$

berechnet, wobei natürlich $T^{-1}(T(z;\ c,\theta);\ c,\theta) = T(z;\ 0,1) = z$ gilt. Da sich in Randnähe Rundungsfehler akkumulieren können, empfiehlt sich für das neue θ eine Nachnormierung auf Länge eins $\theta \mapsto \theta/|\theta|$. Ferner sollten sämtliche Berechnungen in `double` Repräsentation ausgeführt werden.

Mathematiker klassifizieren die Isometrien gerne nach der Anzahl und Lage der Fixpunkte. *(i)* Eine *elliptische Isometrie* hat einen Fixpunkt innerhalb P, z.B. eine Rotation um den Ursprung mit $c = 0$; *(ii)* eine *loxodromische oder hyperbolische Isometrie* hat zwei Fixpunkten auf dem Rand ∂P, z.B. eine Translation, wobei die Fixpunkte die Geodätenendpunkte sind; und *(iii)* der Grenzfall der *parabolischen Isometrie*, die genau einem Fixpunkt auf dem Rand ∂P hat, um dem sich alles dreht.

6.6 Der $\mathbb{H}^2$-Browser und die Mensch-Maschine-Interaktion im Poincaré-Modell

Mit den oben beschriebenen Bausteinen sind die Fundamente eines interaktiven Navigationswerkzeuges für das Poincaré-Modell gelegt. Die Hauptcharakteristika sind:

- Der hyperbolische Raum hat eine exponentiell wachsende Nachbarschaft – an jedem Punkt (nicht nur am Ursprung!);
- Die gesamte Anzeigefläche ist wohldefiniert und kompakt: der Einheitskreis;
- Mit zunehmendem Abstand zur zentralen Fovea nimmt der für den Kontext verwendete Platz ab. Diese nichtlineare radiale Darstellung findet auch ein Vorbild in unserem eigenen retinalen Sehapparat.

Schwartz (1977) hatte die neuronale Repräsentation der Retina bestimmt. Für die zentrale Fovea sind sehr viele Neuronen zuständig und für die peripheren weniger. Die Flächendichte nimmt graduell, etwa mit dem Logarithmus des Winkels zur optischen Achse ab.

Dies wurde in Bereich des Computersehens als *log-polar mapping* vielfach nachgebildet z.B. Smeraldi et al. (2000). Im Gegensatz zur Poincaré-Darstellung leidet die *Log-polar*-Abbildung unter dem Problem einer Singularität am Ursprung;

- In P sind grob drei konzentrische Zonen unterscheidbar: ein hochaufgelöster Innenbereich euklidisch flach; ein Mittelbereich, exponentiell wachsend (oder *log-polar mapped*) und ein dünner Außenbereich, der den Rest konzentriert, statt ihn einfach abzuschneiden. Die Zonenteilung ist in Abb. 6.11 aufgezeigt.

Für den Bau eines $\mathbb{H}^2$-Browsers sind mehrere Interaktionsformen für die Mensch-Maschine-Kommunikation (*MMI, Man Machine Interface*) hilfreich. Ausgehend von einem fertigen $\mathbb{H}^2$-Layout der Objekte in P übernimmt der Browser die Aufgabe der Präsentation und effektiven Benutzerinteraktion.

Objekte sollen mit Markern, Icons und Text annotiert werden können. Aufgrund der besonderen Abbildungseigenschaften von P führt jede nichtpunktförmige Ausdehnung zu Okklusionen in Randnähe. Zum einen sollte die Darstellung, insbesondere Art, Ausdehnung und Schriftgröße, radial angepasst werden. Zum andern kann eine komplette Ausblendung jenseits eines bestimmten Grenzradius erfolgen. Bei großer Objektzahl kann dies zudem den Bildaufbau erheblich beschleunigen.

Mit der Isometrietransformation $T(z;\ c, \theta)$ kann jeder andere hyperbolische Punkt c in den „Fokus“ gerückt werden (Gl. 6.31). Hierzu werden im GUI-Fenstersystem (*Graphical User Interface*) Mausereignisse gebunden. Neben den verfügbaren Tastern (links, Mitte, rechts) können Doppelklicks und Modifier (z.B. Shift, Ctrl) mit ausgewertet werden, um ein reiches Interaktionsrepertoir anzubieten:

- Aktion *drag-point* (z.B. linke Maustaste): Die anfänglich (*mouse down event*) unter dem Mauszeiger sichtbare Position s wird während der Mausbewegung in die aktuelle (und letzlich *mouse release*) Mausposition e überführt. Dazu muss der anfängliche Mausklickort s im ursprünglichen $\mathbb{H}^2$-System s' festgestellt werden und die gegenwär-

tige Transformation durch die Transformation $T(z;\ c_{new}, \theta)$

$$c_{new} = \frac{re[(e-s')(1+\bar{e}\bar{s}')] + im[(e-s')(1-\bar{e}\bar{s}')]\imath}{1-|es'|^2} \quad (6.34)$$

$$s' = T^{-1}(s;\ c_{current}, \theta) \quad (6.35)$$

ersetzt werden.

Diese Interaktionsform ist der mausbewegten Drehung einer Kugel ähnlich und damit dem Erstnutzer schnell vertraut;

- Aktion *click-to-center* (z.B. rechte Maustaste): Die unter dem Mauszeiger sichtbare Position s wird neues Zentrum, d.h. $e = 0$.

  Der Benutzer signalisiert hiermit dem Browser seinen neuen Interessensfokus. Der Browser transferiert den Fokus mit orientierungserhaltender Animation (s. u.);

- Aktion *click-set-center* (z.B. mittlere Maustaste): führt den ursprünglichen Nullpunkt an die aktuelle Mausposition.

  Stellt sich der Benutzer einen Hebel oder Griff an diesem originalen Wurzelpunkt vor, kann er ihn hiermit per Klick unter die Maus holen und verschieben. Dies ist unabhängig von der Verschiebungshistorie und erlaubt eine absolute Lagekontrolle und Rückkehr zur Urposition (*homing*);

- Aktion *Rotation*: Eine Rotation von P ist z.B. sinnvoll, wenn länglich ausgedehnte Objektbeschriftungen sich in gleicher Höhe gegenseitig verdecken. Durch Drehung von P mittels θ (Gl. 6.31) verändern sie ihre vertikale relative Lage.

  Als Benutzeraktion ist eine Kopplung einer Modifiziertaste mit der Winkelauswertung der Mausbewegung möglich (Konzept „Steuerrad“). Einfacher und leichter kommunizierbar sind ein θ-Inkrement und Dekrement an Schaltflächen zu binden;

- *Darstellungsvarianten etc.*: Je nach Verwendungsziel sind ggf. weitere interaktive Darstellungsveränderungen sinnvoll: u.a. das Skalieren der Schriftgröße, Ein-/Ausblenden von Annotation, radiales Zoomen (s. u.). Bewährt hat sich die Anzeige von Zusatzinformationen via *tooltips*, i.e. Informationseinblendungen beim Verharren mit der Maus über einem Objekt. Ferner Kontextmenüs mit Selektion, Deselektion, Aktionsauswahl etc.

6.6.1 Animation

Um dem Betrachter die Orientierung zu erleichtern, werden größere Transformationssprünge bei Einfachmausklicks vermieden. Sie werden stattdessen in eine lineare Folge von Anzeigeframes zerlegt. In n diskreten Zeitschritten t_ν konstruiert man die Animation mit

$$e_\nu = e(t) = s + t_\nu(e - s); \qquad t_\nu \in \left\{ \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1 \right\}. \tag{6.36}$$

Gl. 6.36 beschreibt eine gerade Bewegung in der euklidischen Ebene, die für kleine und mittlere Bewegungen geeignet ist. Für große Entfernungen ist die Trajektorie entlang der $\mathbb{H}^2$-Geraden optimal. Dies verschiebt den Pfad auf das entsprechende P-Kreisbogensegment und teilt die Länge in n in $\mathbb{H}^2$ gleichlange Bogenabschnitte. Je nach aktueller Rechenleistung kann n adaptiert werden, so dass eine gewünschte Zeitspanne bzw. Transfergeschwindigkeit erreicht wird. Das Zeichnen sollte natürlich im Doublebuffering-Modus erfolgen.

6.6.2 Zeichnen von $\mathbb{H}^2$-Verbindungen

$\mathbb{H}^2$-Verbindungsgeraden sind euklidische Kreisbögen in P. Um Rechenzeit zu sparen, kann man während einer flüssigen Animation eine gerade Line zwischen den Endpunkten a und b ziehen. Der Eindruck der Krümmung geht dabei jedoch verloren. Besser ist, insbesondere für ruhende Darstellungen, das Zeichnen des korrekten Poincaré-Kreisbogens. Dessen Kreismittelpunkt m ist durch

$$m = \frac{\imath}{2} \frac{b(1 + |a|^2) - a(1 + |b|^2)}{re(b)im(a) - re(a)im(b)} \tag{6.37}$$

bestimmt und wird von a nach Punkt b gezeichnet (s. Abb. 6.12). Am Vorzeichen des Nenners erkennt man den Drehsinn: Positiv bedeutet Gegenuhrzeigersinn und umgekehrt. Wird der Nenner null, liegt ein verallgemeinerter Kreis, d.h. eine euklidische Gerade durch den Ursprung vor.

6.6.3 Nonkonforme Vergrößerungsabbildung: Zooming

Will man den Blickwinkel in P verändern, ist eine simple Umskalierung (globale Streckung) nicht möglich. Korrekterweise bedeutet dies ein wie-

derholtes Layout der $\mathbb{H}^2$-Koordinaten mit geeignet veränderten Parametern. Einen ähnlichen, aber nicht identischen Effekt kann man in Grenzen durch eine geeignete, nicht-lineare Radialtransformation erreichen. Betrachtet man die fertig berechneten Displaykoordinaten in Polardarstellung, dann verschiebt man letztlich den Ort durch eine Radialstreckung von r auf r':

$$ZF\,\frac{r}{1-r^2} = \frac{r'}{1-r'^2}. \tag{6.38}$$

Die implizite Gleichung ist durch Umskalierung des hyperbolischen Kreisumfanges (Gl. 6.22) motiviert. ZF bezeichnet den gewählten Zoomfaktor. Unter Preisgabe der Konformität hat dies einen einfach berechenbaren Linseneffekt und vermeidet die Deplatzierung des Fokus.

Kapitel 7

Datenvisualisierung im hyperbolischen Raum

„Linien aus Licht,
aufgereiht im Nichtraum des Verstandes,
Ballungen und Anordnungen von Daten.
Wie die Lichter einer Stadt, die sich langsam entfernen ...“

(William Gibson Neuromancer)

Wenden wir uns nun dem Problem zu, wie darzustellende Daten in unseren Darstellungsraum abgebildet werden. Im Folgenden werden drei Lösungen für dies *Layout*-Problem vorgestellt:

- HTL – ein Baum-Layout-Verfahren für hierarchische, azyklische Graphdaten (Trees),
- HSOM – Hyperbolic Self-Organizing Map – für vektorielle Daten,
- HMDS – Hyperbolic Multi-Dimensional Scaling – für Ähnlichkeitsdaten.

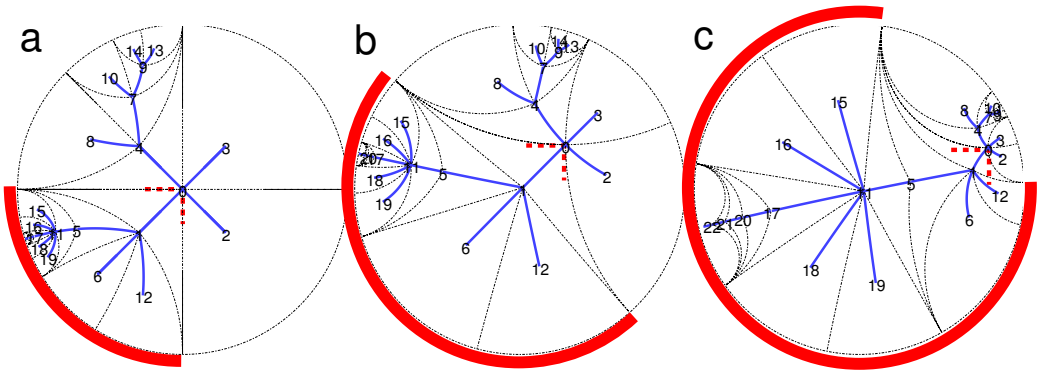


Abbildung 7.1: Das Layout von baumartigen Daten (breite, blaue Linien) und Winkelaufweitung in exponentiell wachsender Nachbarschaft. Die gestrichelten Hilfslinien verdeutlichen den rekursiven Konstruktionsprozess durch die Begrenzungsstrahlen der Sektoren. Durch eine radiale Translation werden die Knoten rekursiv zentriert. Dabei stellt sich ihr Platzangebot jeweils als euklidisches Tortenstück dar und kann damit einfach konstruiert werden. *Äußere Markierungsbogen: (a)* Diesen 90°-Sektor teilt der Wurzelknoten "0" dem Knoten "1", wie jedem seiner 4 Kinder zu. *(b)* Wird Knoten "1" zentriert, stellt sich derselbe Sektor jetzt mit einem Öffnungswinkel von über 180° dar. *(c)* Zwei Schritte weiter weitet er sich auf über 270°. *Knoten "0" mit 90°-Winkelmarkierung*. Die Translation (Gl. 6.31) ist konform und der originale 90°-Schnittwinkel bleibt erhalten. Die gestrichelten Sektorgrenzbögen, die sich am Rande treffen, sind ein gutes Beispiel für asymptotische Parallelen.

7.1 Hyperbolic Tree Layout (HTL) – Layoutkonzept für hierarchische Daten

Für graph-basierte, hierarchische Daten wurde die Aufgabe von Lamping und Rao bei Xerox Parc (Lamping und Rao 1994; Lamping, Rao und Pirolli 1995) gelöst und patentiert.

Gegeben sei ein azyklischer, baum-artiger Graph mit einer definierten Wurzel. Gesucht ist ein kreuzungsfreies Layout in P, i.e. jedem Baumknoten wird ein P-Punkt zugewiesen (ein Blatt wird im Layout als „kinderloser" Knoten gehandhabt).

Der Algorithmus teilt den zur Verfügung stehenden Platz folgendermaßen auf:

1. Der Wurzelknoten wird im Ursprung platziert. Der Platz wird wie eine

Torte in der Kinder-Anzahl entsprechend viele, gleichbreite Sektoren geteilt;

2. Ein Kindknoten i wird aufgerufen unter Angabe der Elternposition, der Zweiglänge l und der Sektorstrahlgrenzen sich und seine Kinder ausgewogen zu platzieren. Als erstes wird die Position mit Hilfe einer relativen, radialen Translation T_i (Gl. 6.31, in Richtung Sektormitte und Länge l) definiert und absolut festgelegt;

3. Sofern Knoten i keine Kinder hat, ist er fertig und returniert – andernfalls zentriert sich der Knoten;

4. Unter T_i werden die Sektorstrahlgrenzen neu bestimmt. Der Effekt ist sehr bemerkenswert: Der Öffnungswinkel weitet sich auf, was in Abb. 7.1 illustriert wird;

5. Der i-eigene Kreissektor wird nun unter den direkten Kindern von i gleich aufgeteilt und jedes Kind rekursiv mit Schritt 2 aufgerufen.

Der Abstand zwischen Eltern- und Kindknoten l kann konstant oder als Funktion des Subsektorwinkels $l(\phi)$ gewählt werden. Je länger l, desto größer ist der Aufweitungseffekt. Die Auswirkungen des „intensiveren Unendlichen“ des hyperbolischen Raumes wird hier nochmal in besonderer Weise veranschaulicht.

Abb. 7.2 zeigt die Implementierung eines Java-basiertes H2-Baumbrowser-Applets, das als Navigations- und Auswahlwerkzeug im Data-Mart Auswertungsportal dient (s. Abs. 9). Neben den beschriebenen Rotations- und Zoominteraktionselementen (hier als Schaltfächen) wird das Auswählen durch Doppelklicken möglich und zusätzlich durch einen optionalen Suchdialog unterstützt.

Tamara Munzner (1997, 1998) entwickelte in ihrer Disseration ein Baumlayoutverfahren für den dreidimensionalen hyperbolischen Raum. Ihr *H3Viewer* kann die dabei entstehenden distelartigen Strukturen zügig manipulieren, s. Abb. 7.3. Der Vorteil des zusätzlichen Platzes in der dritten Dimension ist allerdings mit zwei Nachteilen verbunden:

- Okklusion tritt auf und das Ausblenden von Objekten außerhalb einer Fokussphäre wird wesentlich kritischer, da es auch im Zentralbereich zwangsläufig zu Verdeckungen kommt;

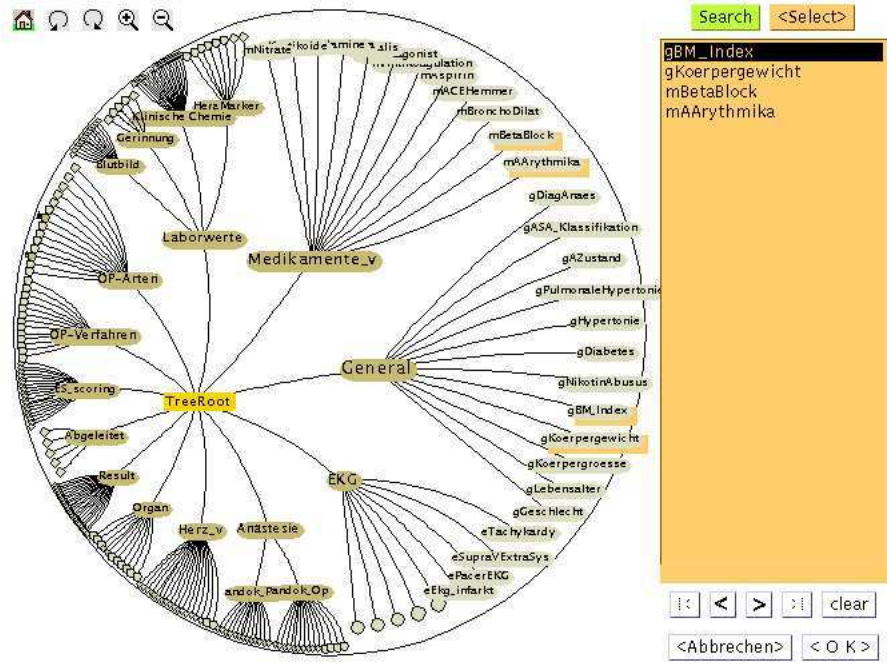


Abbildung 7.2: $\mathbb{H}^2$-TreeBrowser-Applet zur Objektselektion aus einer baumartigen Hierarchie für die Merkmalsselektion. Im webbasierten Datenexport aus einer Datenbank (Data-Mart, s. Abs. 9) ist die Auswahl aus mehr als 250 Merkmalen übersichtlicher, wenn die Hierarchie der Merkmale aus dem Kontext ersichtlich ist. Die interaktiven Auswahlmöglichkeiten werden mit einem Textsuchdialog (*„search“*) und dem Traversieren (*„>“*) der (*rechts*) Auswahl- bzw. Suchliste ergänzt.

- Die zusätzliche Navigationsfreiheit von 6 statt 3 Orientierungsfreiheitsgraden ist mit konventionellen Computerinputgeräten, z.B. der Maus, nicht einfach zu gestalten. Die Tiefenwahrnehmung ist nur für bekannte Objekte ohne spezielle Stereodisplays möglich.

7.2 Die Hyperbolic Self-Organizing Map (HSOM)

Im Folgenden wird ein Erweiterung von Kohonen's selbst-organisierender Merkmalskarte (s. Abs. 5.9) für nicht-euklidische Räume vorgestellt (Ritter 1999). Die Kernidee ist die Verwendung einer angepassten Nachbarschaftsfunktion $h(\mathbf{a}, \mathbf{a}^*) = h(d_{\mathbf{a},\mathbf{a}^*})$, s. Gl. 5.72. Statt eines regelmäßigen

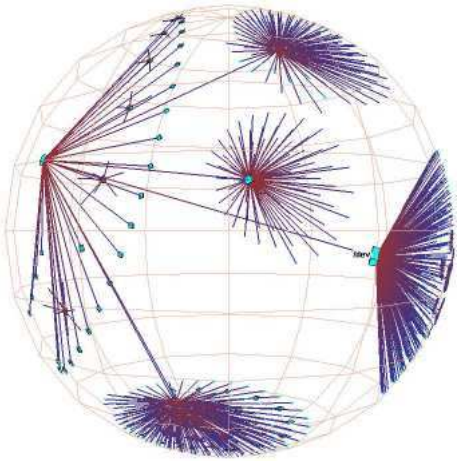


Abbildung 7.3: „H3-Viewer“-Projektion des H^3 nach T. Munzner.

2D- oder 3D-Gitters wird nun ein $\mathbb{H}^2$-Gitter zugrunde gelegt. Eine besonders geeignete Wahl ist ein Ausschnitt eines regulären Tesselationsgitters (s. Abb. 7.4).

Reguläre Dreiecksgitter können mittels eines geeigneten Anfangsdreiecks durch wiederholte Anwendung von Generatoren der Symmetriegruppe erzeugt werden, siehe auch Abs. 6.4. Abb. 7.5 zeigt anhand zweier Exemplare, wie eine $\mathbb{H}^2$-Tesselation durch Verbindung von jeweils n kongruenten, gleichschenkligen Dreiecken an jedem Knotenpunkt entsteht. Das äquilaterale hyperbolische Basisdreieck wird allein durch n bestimmt:

$$\text{Innenwinkel} \quad \alpha = 360^\circ/n \tag{7.1}$$

$$\text{Poincaré-Kantenlänge} \quad l = \sqrt{2\cos\alpha - 1}\,. \tag{7.2}$$

Gl. 7.1 gilt offensichtlich aus Symmetriegründen. Wegen des Defizites der $\mathbb{H}^2$-Dreiecksinnenwinkelsumme $3\alpha < 180^\circ$ ist 7 die kleinstmögliche Zahl n. Gl. 7.2 ergibt sich am einfachsten aus den hyperbolischen Eckdistanzen (s. u. Gl. 7.5) in einem Dreieck mit den komplexen Eckpunkten $0, l, le^{\imath\alpha}$ in P.

Wie kann die Knotenzahl $K_{n,R}$ berechnet werden? Das Gitter kann als in Ringen zuwachsender Baum betrachtet werden. Im R-ten Jahresring werden $\Delta K_{n,R}$ Knoten angebaut:

$$\text{Gesamtzahl} \quad K_{n,R} = 1 + \sum_{i=1}^{R} \Delta K_{n,i} \tag{7.3}$$

$$\text{und marginale Anzahl} \quad \Delta K_{n,R} \;=\; K_{n,R} - K_{n,R-1}\,.$$

Bei genauem Betrachten des konzentrischen Aufbaus in Abb. 7.5 erkennt man, dass einige Knoten (aus dem R-ten Ring) mit genau *einem* Knoten des nächst inneren Ringes $R-1$ verbunden sind – und einige, wie sich herausstellt, genau $\Delta K_{n,R-1}$ Knoten mit *zweien*. Betrachtet man die Anzahl von vakanten Bindungen im Ring R, ergibt sich unter Berücksichtigung der tangentialen Ringnachbarn

$$\Delta K_{n,R+1} = (n-4)\Delta K_{n,R} - \Delta K_{n,R-1} \tag{7.4}$$

mit den Basisfällen $\Delta K_{n,1} = n$ und $\Delta K_{n,0} = 0$.

Tabelle 7.1 gibt einige Beispiele für die daraus resultierende Knotenzahl $K_{n,R}$. Wie erwartet und in Abb. 7.6 gezeigt, wächst sie (asymptotisch) exponentiell mit der Anzahl der Dreiecksringe R um den Ursprung.

Anpassung der Nachbarschaftsfunktion: Bei der Gitterkonstruktion wird jedem dieser $K_{n,R}$ Knoten (Neuron) $\mathbf{a}$ der Gitterort $\mathbf{g}_a$ in der $\mathbb{H}^2$-Ebene zugeordnet. Der Zwischenknotenabstand wird dann in der richtigen

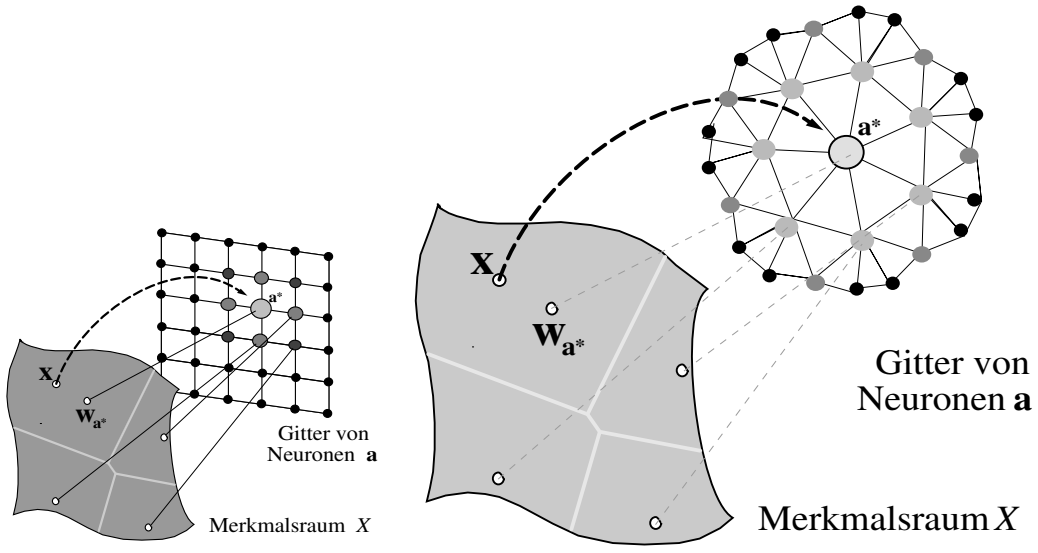


Abbildung 7.4: Die „Self-Organizing Map“ („SOM“) wird durch ein Gitter von verarbeitenden Knoten bzw. „Neuronen“ gebildet. *(a, links:)* Zum Vergleich: Die Standard-SOM nutzt ein reguläres 2D-Gitter, s. Abb. 5.19, S. 137. *(b, rechts:)* Die hyperbolische SOM nutzt dagegen ein reguläres, hyperbolisches Gitter, hier durch einen Ausschnitt eines $\mathbb{H}^2$ illustriert (s.a. Abb. 7.5).

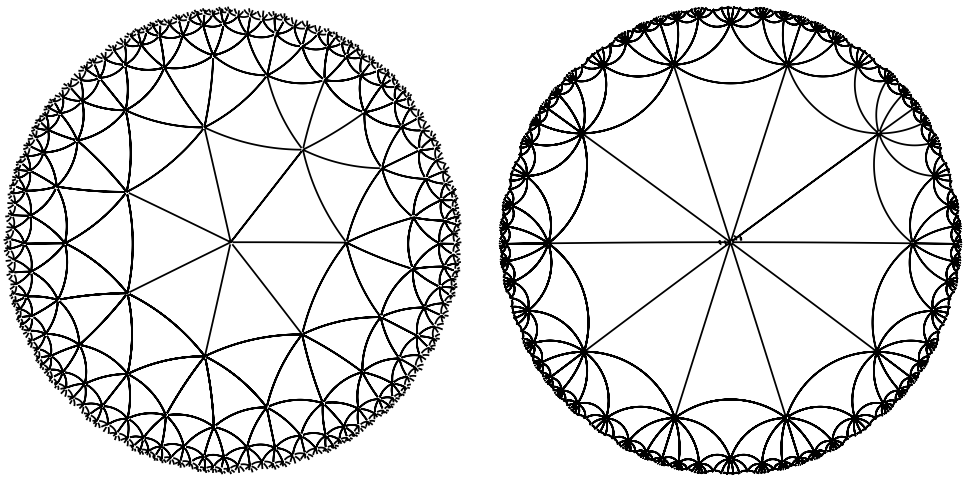


Abbildung 7.5: Reguläre $\mathbb{H}^2$-Tesselationen. Die Ebene wird vollständig und überdeckungsfrei mit kongruenten, äquilateralen Dreiecken gekachelt. *(Links:)* Hier treffen sich je $n = 7$ Dreiecke an den Vertexen und bilden n-Gonen (i.e. Polygone mit n Ecken). *(Rechts:)* Für $n = 10$ sind die Dreiecksseiten länger. Gut erkennbar ist die Erscheinung der $\mathbb{H}^2$-Geraden als Kreisbogen, die den „∞-Rand“ senkrecht schneiden.

R	1	2	3	4	5	6	7	8	9
$K_{7,R}$	8	29	85	232	617	1625	4264	11173	29261
$K_{8,R}$	9	41	161	609	2281	8521	31809	118721	443081
$K_{9,R}$	10	55	271	1306	6265	30025	143866	689311	
$K_{10,R}$	11	71	421	2461	14351	83651	487561	...	
$K_{12,R}$	13	109	865	6817	53677	422605			

Tabelle 7.1: Knotenzahl $K_{n,R}$ von hyperbolischen Dreiecken in einem Dreiecksgitter aus n-Gone und R „Ringen“ um den Ursprung. Der erste $R = 1$-Ring enthält natürlich $K_{n,1} = n + 1$ Knoten. (s.a. Gl. 7.3ff und Abb. 7.6).

Metrik, hier der Poincaré-Metrik, nach Gl. 6.28 gemessen

$$d_{\mathbf{a},\mathbf{a}^*} = 2 \operatorname{arctanh} \left(\frac{|\mathbf{g}_a - \mathbf{g}_{a^*}|}{|1 - \mathbf{g}_a \bar{\mathbf{g}}_{a^*}|} \right). \tag{7.5}$$

Bemerkenswert ist, dass alles weitere am SOM-Verfahren unverändert bleiben kann. Gl. 7.5 ist der Schlüsselpunkt, an dem die hyperbolische Struktur des Raumes via Regularisierung der Karte eingeprägt wird.

Abb. 7.7 zeigt eine typische Entfaltung der HSOM in drei Stadien (Ritter

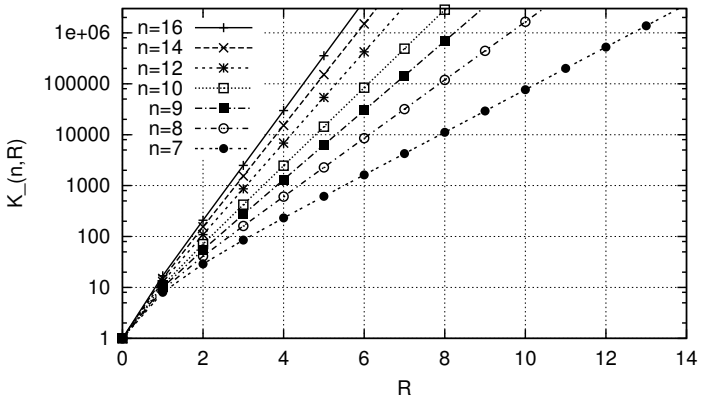


Abbildung 7.6: Die Knotenzahl $K_{n,R}$ eines regulären hyperbolischen Dreiecksgitters versus der „Ringzahl“ R für verschiedene Vertexordnung n („n-Gon“). Die semilogarithmische Darstellung zeigt das exponentielle Anwachsen mit der Ringzahl R (s.a. Tab. 7.1).

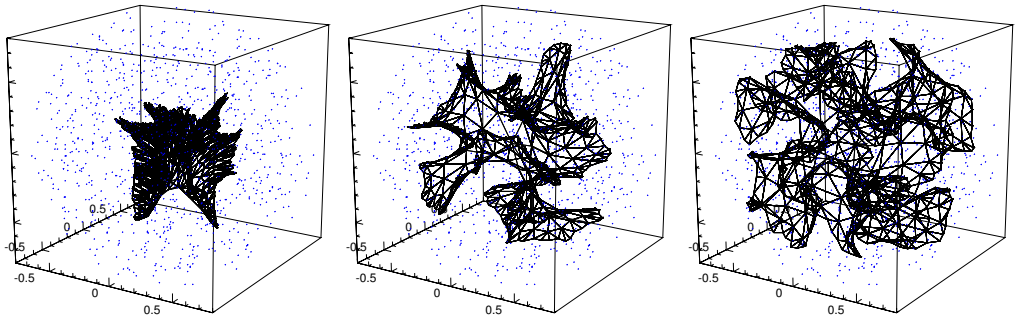


Abbildung 7.7: Entfaltung einer 2D-HSOM in einer uniformen Datenverteilung, d.h. einer 3D-Kugel: *(links)* in einer frühen Ordnungsphase; *(Mitte)* die Sattelpunktstruktur wird sichtbar; *(rechts)* die Endkonfiguration zeigt, wie das $\mathbb{H}^2$-Gitter den Raum der Einheitskugel ausfüllt.

1999). Die Endkonfiguration illustriert, wie sich das $\mathbb{H}^2$-Gitter raumfüllend ausgebreitet hat.

An den Stellen, an denen sich (auf dem Gitter) entfernte Falten annähern, entstehen in der Kartenabbildung Unstetigkeiten, die auch als **topologische Defekte** bezeichnet werden. Ihr Auftreten ist unvermeidlich, wenn die Struktur der Datenverteilung nicht der Kartentopologie entspricht. Für Visualisierungszwecke ist häufig die Dimensionsreduktion auf

eine planare Mannigfaltigkeit gewünscht, aber man sollte die Möglichkeit von topologischen Defekten nicht außer Acht lassen.

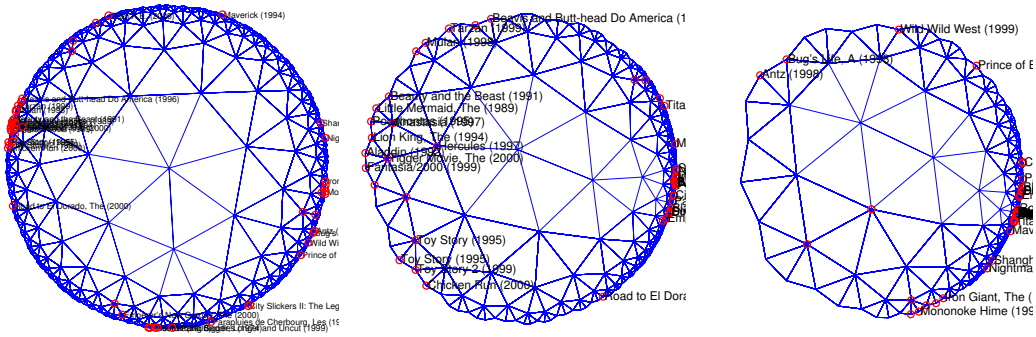


Abbildung 7.8: Drei Fokuspositionierungen der HSOM (von mittig bis links oben). Die Linien verbinden nächste Nachbarn im Knotendreiecksgitter ($n = 7$, $R = 5$, $K_{n,R} = 617$). Die Beschriftung nennt den in dem Textdokument beschriebenen Film bzw. den am häufigsten beschriebenen Film, falls mehrere Dokumente repräsentiert werden. Es bilden sich Gruppen von semantisch zusammengehörigen Filmen, z.B. Disney-Filmproduktionen im 10-Uhr-Bereich.

Die Eignung der HSOM zur automatischen Kategorisierung von Textdokumenten demonstrierte Jörg Ontrup u.a. für die Filmkritikdaten, die in Abs. 8.1 beschrieben sind (Ontrup und Ritter 2001). Anhand des Extra-Kategorisierungslabels konnte die HSOM auf der separierten Testdatenmenge Benchmarkergebnisse erzielen, die mit anderen *State-of-the-art*-Methoden vergleichbar sind. Abb. 7.8 zeigt an drei Projektionen mit einer HSOM mit $K_{7,5} = 617$ Knoten, wie eine semantisch bedeutungsvolle Ordnung entsteht (weitere Details in Ontrup und Ritter 2001).

Durch die Einbettung in den $\mathbb{H}^2$ gewinnt das Verfahren zum einen die Navigationsfähigkeit im Poincaré-Modell und zum anderen das besondere Wachstumsverhalten des Platzes in der Nachbarschaft: Es wächst lokal exponentiell um **jeden** Punkt anstatt quadratisch (solange Gitterrandeffekte keine Rolle spielen).

Nachteilig ist die Fixiertheit auf das Knotengitter:

- Zum einen führt der Wettstreit um den best-match-Knoten zuweilen zu einer unausgewogenen Zuständigkeit, einige Knoten bekommen keine, andere viele Dokumente. Legt man Wert auf deutliche Clusterbildung auf dem Netz, müssen genügend viele Knoten angeboten werden – bis zu einem Mehrfachen der Datenanzahl N. Dies führt

mitunter zu großen Netzen, die einen Großteil ihrer Knoten in den entfernteren „Außenbereichen“ besitzen und damit weite Navigationsstrecken beim Browsen bedeuten. Die Gesamtstruktur erscheint damit weit verstreut und die Trainingsphase ist lange.

- Wählt man ein kompaktes Netz mit $K << N$, werden starke Aggregationen häufig, d.h. ein Knoten ist häufig für zahlreiche Datentupel zuständig. Die Einzelobjekte, hier Dokumente, können somit nicht mehr separiert räumlich aufgelöst werden.

Eine Mischform sind *additive* HSOMs, die kompakt starten und dynamisch Knoten ein- oder anfügen. Im hyperbolischen Dreiecksgitter bietet sich das Konzept des *sub-meshing* an: Durch (drei) Schnittführungen durch je zwei der drei Kantenmitten wird ein gewähltes Dreieck konstruktiv gevierteilt. Der rekursiv ausführbare Prozess erzeugt damit auch die Visualisierungskoordinaten auf einfache Weise. Problematisch ist die zeitliche Teilungssteuerung während des Trainings. Je früher Teilungen vorgenommen werden, desto schneller differenziert das Gitter, aber umso weniger ist eine visuelle Ausgewogenheit der Darstellung zu garantieren. Die Gitterfixiertheit ist hier nicht aufgehoben, sondern auf eine kleinere Skala verschoben.

In den folgenden Abschnitten werden Strategien aufgezeigt, die hierbei Abhilfe schaffen.

7.2.1 Interpolationsansätze

Sucht man einen Ausweg aus dem Dilemma der Layout-Allokation von großen Datenmengen mit möglichst wenig Knoten, stellt sich die Frage nach möglichen Interpolationsverfahren. Ausgehend von einer gröberen, dimensionsreduzierenden Abbildung des Merkmalsraumes in den Layout-Raum, stellt sich die Frage, welche gut skalierbaren Ansätze sich eignen, um große Mengen von Dokumenten, oder allgemein Datenobjekten, sinnvoll im Layoutraum zu platzieren? Neben der vorangestellten Shepard-Methode werden insbesondere Verfahren beleuchtet, die topologische Informationen generieren. Möglichkeiten, dies zur schnellen projektiven Visualisierung zu verwenden, werden diskutiert.

Shepard's Methode ist ein topologiefreies Verfahren, das eine Linearkombination von Stützpunkten ermittelt (Shepard 1968). Der Ort jedes

Stützpunktes i im hochdimensionalen Merkmalsraum $\mathbf{x}_i$ (z.B. hier der Referenzvektoren) korrespondiert mit dem Ort u_i im Layout-Raum (z.B. $\mathbf{g}_i \in$P):

$$u(\mathbf{x}) = \sum_i \phi_i(\mathbf{x} - \mathbf{x}_i) u_i \tag{7.6}$$

Die Gewichtungsfaktoren ϕ_i entstehen durch eine Zerlegung der 1

$$\phi_i(\mathbf{x} - \mathbf{x}_i) = \frac{d(\mathbf{x} - \mathbf{x}_i)}{\sum_i d(\mathbf{x} - \mathbf{x}_i} \tag{7.7}$$

und sind hier rein abstandsbasiert

$$d(\mathbf{x} - \mathbf{x}_i) = \frac{1}{||\mathbf{x} - \mathbf{x}_i|| + \epsilon} \; . \tag{7.8}$$

Die kleine Konstante ϵ verhindert die Divergenz an den Stützstellen. Diese Form der Interpolation ist sehr leicht zu implementieren, nutzt aber keinerlei Strukturinformation. Die Extrapolation jenseits der Stützstellen gelingt nicht: Das Ergebnis wird dann zur bloßen Mittelwertbildung, was für entfernte Daten zu einer Klumpung im Schwerpunkt führt.

PSOM – die *Parametrisierte SOM* nutzt die topologische Struktur der SOM. Sie ermöglicht extrem *schnelles Lernen* mit einem sehr kleinen Trainingsdatensatz, wenn die zugrunde liegende Mannigfaltigkeit kontinuierlich und glatt ist (eine ausführliche Darstellung findet sich z.B. in Ritter 1993; Walter und Ritter 1996; Walter 1996). An jeder Stützstelle $\mathbf{a}$ des Gitters $\mathbf{A}$ wird eine Basisfunktion $H_{\mathbf{a}}(\mathbf{s})$ geknüpft, was eine kontinuierliche Verallgemeinerung des Gitters $\mathbf{A}$ zu einer kontinuierlichen Abbildungsmannigfaltigkeit S mit $\mathbf{s} \in S$ gestattet. Die strikte Trennung von Ein- und Ausgaberaum wird aufgehoben durch die Bildung des kartesischen Produktes derselben. Die Basisfunktionen $H_{\mathbf{a}}(\mathbf{s})$ fungieren dann als Gewichtungsfaktoren ϕ_i

$$\mathbf{w}(\mathbf{s}) = \sum_{\mathbf{a} \in \mathbf{A}} H_{\mathbf{a}}(\mathbf{s}) \mathbf{w}_{\mathbf{a}} \tag{7.9}$$

ähnlich Gl. 7.6 und müssen bestimmte Bedingungen erfüllen (i.e. Orthonormalität und Zerlegung der 1). Eine geschickte Wahl sind die Lagrange-Polynome, die sich auf ein mehrdimensionales, rechteckiges Gitter verallgemeinern lassen. Nachdem die PSOM trainiert ist, wird die beste Position $\mathbf{s}^*$ in der Abbildungsmannigfaltigkeit S bestimmt, die den kleinsten Abstand

$$\mathbf{s}^* = \mathbf{s}(\mathbf{x}) = \underset{\forall \mathbf{s} \in S}{\text{argmin}} \; dist\left(\mathbf{w}(\mathbf{s}), \mathbf{x}\right) \tag{7.10}$$

zum gegebenen Vektor $\mathbf{x}$ hat. Die Abstandsmetrik $dist(\cdot)$ berücksichtigt dabei nur die als Eingabe gewählten Komponenten. Die Ausgabe $\mathbf{w}(\mathbf{s}^*)$ kann als assoziative Komplettierung der Eingabe bezeichnet werden. In der Anwendung ergeben sich daraus vielseitige Möglichkeiten, z.B. im Kontext der Robotik und des Computersehens (Walter 1998; Walter et al. 2000). Der PSOM gelingt es in einer außergewöhnlich kleinen Trainingsdatenmenge, Krümmungsinformation aus deren topologischer Ordnung zu extrahieren. Der Preis dafür ist die iterative Bestimmung von $\mathbf{s}^*$ mittels rekurrenter Dynamik (Gl. 7.10).

Bei knotenreichen PSOM sind die Skaliereigenschaften hochdimensionaler Interpolationspolynome ungünstig. Als Lösung bieten sich Local-PSOM (L-PSOM) an, die dynamisch stückweise Polynome verwenden (ähnlich zu Splines, siehe Walter und Ritter 1995).

Interpolating SOM (ISOM): Der Local-PSOM-Ansatz enthält als Spezialfall die Verwendung von Polynomen erster Ordnung, die den von J. Göppert (1997) untersuchten *ISOM*s äquivalent sind. Für einen neuen Eingabevektor wird ein lokales Koordinatensystem zentriert am *Best-match*-Knoten des SOM-Gitters gelegt und die affinen Koordinaten werden iterativ gesucht.

Hauptflächen *(principal surfaces)* sind eine mehrdimensionale Verallgemeinerung von Hauptkurven. **Prinzipale Kurven (*Principal Curves*)** (PC) generalisieren die PCA nichtlinear auf gekrümmten Kurven (Duchamp und Stuetzle 1996). Sei $f(\lambda) \in \mathbb{R}^p$ eine Funktion, die die glatte Kurve entlang ihres Wegs in $\mathbb{R}^p$ parametrisiert und $\lambda_f(x)$ die Menge aller Punkte, die zu x auf der Kurve am nächsten sind. Für eine Datenverteilung X wird die Hauptkurve oder *principal curve* durch den lokalen Erwartungswert

$$f(\lambda) = E\{X|\lambda_f(X) = \lambda\} \tag{7.11}$$

in selbstkonsistenter Weise definiert. Auch für einen gegebenen Datensatz sucht man die Hauptkurve, die „mitten“ durch die Daten geht. Gefunden wird sie iterativ, meist mit dem EM-Verfahren. In der Regel gibt es unendlich viele Lösungen (Duchamp und Stuetzle 1996) und es bedarf einer geeigneten Regularisierung mit der die Glattheit der Verteilung gesteuert wird. Die mehrdimensionale Verallgemeinerung führt zu den Hauptflächen (*Principal Surfaces*, PS). Existenz- und Konvergenzbeweise wurden für 1D-PCs vorgeschlagen, sind aber nicht allgemein auf mehrere Dimensionen übertragbar (s. z.B. Tibshirani 1992).

Probabilistic Principal Surfaces (PPS) und Generative Topograhic Map (GTM): Chang und Ghosh (2001) schlugen das PPS als vereinheitlichtes, probabilistisches Modell vor. PPS nutzt ein flächenorientiertes Rauschmodell und baut auf dem GTM-Verfahren (*Generative Topographic Map*) auf, einer parametrisierten Variante von Kohonen's SOM (Bishop, Svensen und Williams 1998). Ein einzelner Modellparameter α kontrolliert die Entstehung einer PPS, GTM oder flächenorientierten GTM. Die Konvergenzeigenschaften von PPS sind äquivalent zu GTM. Der Hauptvorteil von GTM relativ zur SOM ist die bessere theoretische Fundierung (u.a. Existenz einer Kostenfunktion) und praktisch das Wegfallen von Randeffekten des Gitters. Im Effizienzvergleich ist PPS leider noch langsamer als GTM, welches selbst deutlich langsamer ist als das SOM-Verfahren.

Topologische Rück-Projektion: Alle vorgenannten Verfahren lassen sich mit Hilfe von trilinearen oder so genannten barizentrischen Koordinaten (Möbius, s. Coxeter 1969) lokal auf die Dreiecksgitterstruktur der HSOM anpassen. Problematisch ist die Frage nach der geeigneten Interpolation an den Kanten und Ecken. Wie Abb. 7.7 (S. 184) illustriert, ist die Entfaltung einer HSOM mitunter mit starken Unebenheiten verbunden.

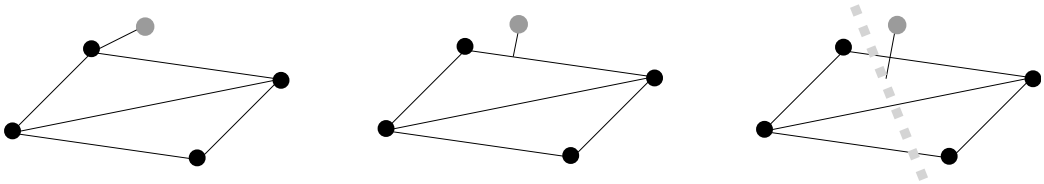


Abbildung 7.9: Drei Rückprojektionsverfahren eines neuen Datenpunktes (grau) auf das Gitter im Merkmalsraum (vier Punkte dargestellt): (*a, links:* knotenbasiert) Das *Best-match*-Verfahren wählt in der Menge aller Knoten den nächsten; (*b, Mitte:* kantenbasiert) oder den nächsten Kantenort; (*c, rechts:* flächenbasiert) oder den nächsten Ort in den Dreiecksflächen. Bei den beiden letzten Verfahren treten Unstetigkeitsprobleme auf, die im eindimensionalen Fall in Abb. 7.10 illustriert werden (die gepunktete Linie zeigt eine mögliche Schnittführung).

Abb. 7.9 illustriert verschiedene Rückprojektionsverfahren im Dreiecksgitter. Ihnen allen ist das Problem der Unstetigkeiten gemeinsam, das in Abb. 7.10 im Schnittbild demonstriert wird. Die Gebietsteilung durch die Dreiecksflächenstücke findet auch im m-dimensionalen Vektorraum statt und besteht aus: *(i)* Dreiecksprismen, die senkrecht auf den Gitterflächen stehen und sich gegenseitig begrenzen (sie entsprechen Gebiet $3,5$ in

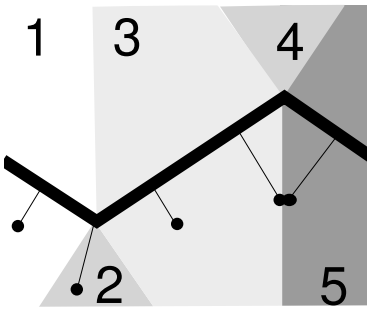


Abbildung 7.10: Probleme der Rückprojektion im 2D-Schnittbild (s. Abb. 7.9c). In den Gebieten 1, 3 und 5 werden die Punkte durch senkrechte Projektion eindeutig auf die stückweise lineare Projektionsfläche (Linienzug) abgebildet. Alle Punkte im Gebiet 2 (4) werden auf den Eckpunkt (ggf. Kanten) abgebildet. An den Trennflächen 3/5 (1/3) entstehen zwangsläufig Unstetigkeiten im Projektionsergebnis.

Abb. 7.10) und *(ii)* kappenartige Polyeder, die den Ecken und Kanten zugeordnet sind (sie entsprechen Gebiet 2, 4; z.T. mit einer Seite offen). Der relative Volumenanteil dieser Kappen wird größer,

- je höher die Dimension m ist und
- je unebener die Einbettung der Gitterfläche in den $\mathbb{R}^m$ ist.

In beiden Fällen werden die Unstetigkeitsprobleme größer.

7.2.2 Unebenheiten bei hochdimensionalen Gittereinbettungen

Wie stark nimmt die Rauigkeit der Dreiecksgitterfläche in höheren Dimensionen zu? Um diese Frage zu beleuchten, wurde exemplarisch eine HSOM mit einer uniformen Datenverteilung trainiert und das sich ergebende Gitter statistisch auf Unebenheit und Asymmetrie analysiert. Zwei Indikatoren wurden bestimmt, deren geometrische Motivation in Abb. 7.11 erläutert wird.

Verbiegungswinkel von Dreieck zu Nachbarkanten: Zum einen ist dies

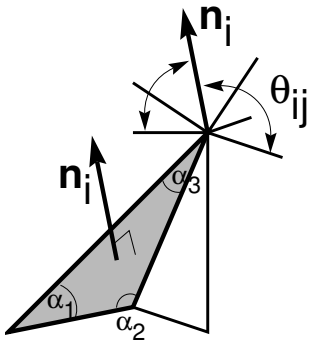


Abbildung 7.11: Die Bestimmung der lokalen Gitterrauigkeit an einem Dreieck i erfolgt über die Elevationswinkel $\phi_{ij} = |90° - \theta_{ij}|$ der angrenzenden Gitterkanten j über der Ebene i. Vorteilhafterweise werden sie über die Schnittwinkel θ_{ij} zur Flächennormalen $\mathbf{n}_i$ errechnet. Der minimale Innenwinkel $\alpha_{min} = \min(\alpha_1, \alpha_2, \alpha_3)$ gibt Aufschluss über die Spitzheit des Dreiecks. Je näher α_{min} an seiner oberen Schranke $60°$ liegt, desto symmetrischer ist das Dreieck.

der Elevationswinkel ϕ_{ij}, der zwischen der Dreiecksfläche i und einer angrenzenden Kante j liegt. Die Mittelung dieses Verbiegungswinkels $\langle\phi\rangle$ erfolgt über alle Dreiecke des Gitters und alle angrenzenden Kanten j jedes einzelnen Dreiecks i.

Dreiecksspitzheit: Die Innenwinkelsumme jedes Dreiecks im $\mathbb{R}^m$ ist $180°$ (s. Abb. 7.11). Damit gilt für den kleinsten Innenwinkel $0 < \alpha_{min} \leq 60°$. Ferner ist ein Dreieck desto gleichwinkliger und damit symmetrischer, je näher α_{min} an der oberen Schranke liegt.

Die Ergebnisse von Simulationsexperimenten zeigt Abb. 7.12. Ein HSOM mit $N_{7,3} = 85$ Gitterknoten, 112 -dreiecken und 196 -kanten wurde je 10 mal auf drei m-dimensionale Gaußverteilungen trainiert ($\lambda_i = 1.5, \lambda_f = 0.2$ mit je 1000 Trainingsschritten). Aufgetragen sind die gemittelten charakteristischen Winkel $\langle\alpha_{min}\rangle$ und $\langle\alpha_{min}\rangle$ versus der Einbettungsdimension m. Die Resultate bestätigen die Hypothese, dass von einer einigermaßen ebenen Gittereinbettung in hochdimensionalen Räumen nicht die Rede sein kann. Für eine kugelförmige Gaußverteilung erreicht die mittlere Gitterverbiegung etwa $40°$ schon bei $m = 12$. Je dünner (Dickenparameter s) die kreisscheibenförmige elliptische Gaußverteilung ist, desto später werden diese Werte erreicht. Bemerkenswert ist, dass es dem Gitter bei höheren Dimensionen gelingt, die meisten Dreiecke recht symmetrisch auszuformen, wie man an $\alpha_{min} \approx 53°$ ablesen kann.

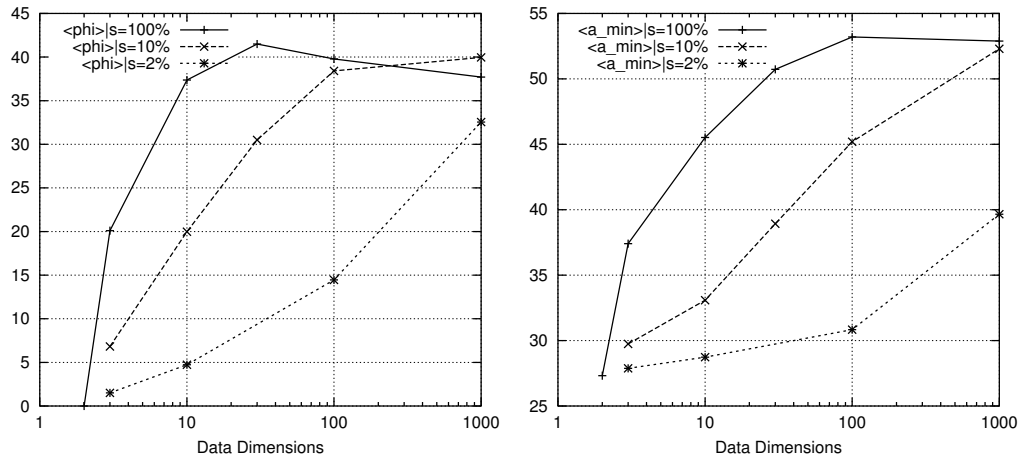


Abbildung 7.12: Gitterrauigkeit versus Einbettungsdimension m: *(links:)* der mittlere Verbiegungswinkel $\langle\phi\rangle$; *(rechts:)* der mittlere kleinste Dreieckswinkel $\langle\alpha_{min}\rangle$ (in Grad °). Die Datenverteilung besteht aus scheibenförmigen, elliptischen Gaußverteilungen mit Radius 1 und Dicke s, d.h. mit zwei Komponenten $\sigma_1 = \sigma_2 = 1$ und den restlichen $m-2$ Komponenten $\sigma_i = s \leq 1$. Die drei Graphen sind für eine dünne ($s = .02$) und eine flache ($s = .1$) Scheibe sowie für eine Kugel ($s = 1$) gezeichnet.

Zwischenbilanz: Die Interpolation mit topologischen Karten erfordert ein Rückprojektionsverfahren. Bei hochdimensionaler Einbettung sind die Einbettungsflächen durchaus sehr uneben, was zu ungewünschten Unstetigkeiten in der Visualisierungsanwendung führt. Für stückweise ebene Interpolationsflächen, die sich durch lineare Splines darstellen lassen, erfolgt eine Fokussierung auf den Bildraum der Knoten- bzw. Kantenmenge (Abb. 7.10). Verfahren höherer Ordnung weisen unvorteilhafte Skalierungseigenschaften bezüglich der Einbettungsdimension auf (Walter und Ritter 1995).

Im nächsten Abschnitt wird nun ein weiteres Visualisierungsverfahren eingehender vorgestellt, das eine nicht-euklidische Verallgemeinerung erfahren hat.

7.3 HMDS – Multidimensionale Skalierung im hyperbolischen Raum

Wie kann man Daten, die durch Unähnlichkeiten beschrieben sind (Fall F2 in Abs. 2.2), unter bestmöglichem Erhalt ihrer Abstandsstruktur in den hyperbolischen Raum einbetten? Wie kann man das in Abs. 5.10 dargestellte MDS-Verfahren generalisieren? Ist die daraus entstehende nichteuklidische Einbettung per se ein Vorteil oder ist allein der Nutzen des Visualisierungspotentials in der hyperbolischen Ebene der Gewinn?

Die hyperbolische Verallgemeinerung von MDS heißt konsequenterweise **hyperbolische Multidimensionale Skalierung** (*hyperbolic Multi-Dimensional Scaling* **HMDS**, Walter und Ritter 2002; Walter 2004). Die Kernidee stellt sich als strukturell sehr einfach heraus: Anstatt die MDS-Lösung in einem niedrig-dimensionalen euklidischen Raum $\mathbb{R}^L$ zu suchen und in den $\mathbb{H}^2$ zu transferieren (was nicht direkt möglich wäre), modifiziert man das MDS-Verfahren so, dass es von Anfang an im hyperbolischen Raum operiert. Die Schlüsselstelle ist die Distanzfunktion in Gl. 5.81. Die euklidische Distanz wird durch die geeignete Metrik ersetzt, hier die Distanzmetrik im Poincaré-Modell (s. Gl. 6.28)

$$d_{ij} = 2 \operatorname{arctanh}\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{|1 - \mathbf{x}_i \bar{\mathbf{x}}_j|}\right) \ , \quad \mathbf{x}_i, \mathbf{x}_j \in \mathsf{P}. \tag{7.12}$$

Einige Aspekte müssen berücksichtigt werden: Die P-Scheibe enthält den gesamten $\mathbb{H}^2$. Daher muss sichergestellt werden, daß Update-Schritte für den Knoten i^* nicht zum Verlassen des Einheitskreises führen können. Einfache Längenbegrenzungen von $\eta\Delta_{i^*}$ in Gl. 5.85 sind unzureichend. Stattdessen regelt die Möbiustransformation (Gl. 6.31) die Schrittweitenreduktion in Kreisrandnähe:

$$\mathbf{x}_{i^*}^{(new)} = T(\mathbf{x}_{i^*}^{(old)};\ \eta\Delta_{i^*}, 1)\ . \tag{7.13}$$

Während die Gradientenberechnung $\partial d_{ij,q}/\partial x_{i,q}$, die in Gl. 5.86 benötigt wird, für den euklidischen Fall einfach zu berechnen ist ($= (x_{i,q} - x_{j,q})/d_{ij}$), wird der Fall für Gl. 7.12 (in mehrfacher Hinsicht) komplex:

$$\frac{\partial}{\partial x_{i,1}} d(\mathbf{x}_i, \mathbf{x}_j) = \frac{2t}{1-t^2}\left(\frac{v_1}{v_1^2 + v_2^2} - \frac{x_{j,1}v_3 + x_{j,2}v_4}{v_3^2 + v_4^2}\right) \tag{7.14}$$

$$\frac{\partial}{\partial x_{i,2}} d(\mathbf{x}_i, \mathbf{x}_j) = \frac{2t}{1-t^2}\left(\frac{v_2}{v_1^2 + v_2^2} + \frac{x_{i,1}v_4 - x_{j,2}v_3}{v_3^2 + v_4^2}\right) \tag{7.15}$$

mit

$$\begin{aligned}
\mathbf{x}_i &= (x_{i,1} + \mathsf{i}\, x_{i,2}) \in \mathsf{P} \subset \mathbb{C} \\
\mathbf{x}_j &= (x_{j,1} + \mathsf{i}\, x_{j,2}) \in \mathsf{P} \subset \mathbb{C} \\
v_1 &= x_{i,1} - x_{j,1} \\
v_2 &= x_{i,2} - x_{j,2} \\
v_3 &= x_{i,1}x_{j,1} + x_{i,2}x_{j,2} - 1 \\
v_4 &= x_{i,1}x_{j,2} - x_{j,1}x_{i,2} \\
t^2 &= \frac{v_1^2 + v_2^2}{v_3^2 + v_4^2}.
\end{aligned} \tag{7.16}$$

Dies enthält zwei Spezialfälle:

- Der Nenner $v_1^2 + v_2^2$ wird nur 0, wenn die beiden Punkte $\mathbf{x}_i$ und $\mathbf{x}_j$ identisch sind;
- Der Nenner $v_3^2 + v_4^2$ wird innerhalb des Einheitskreises P nie 0 – ist also auch unkritisch.

Wegen der Komplexität dieses Ergebnisses ist eine zweite Ableitung schwer darstellbar. Das Levenberg-Marquardt-Verfahren (1963) verbessert den Minimalisierungsschritt (Gl. 5.86) mit einem Ansatz, der als eine Kombination eines Verfahrens erster und zweiter Ordnung betrachtet werden kann. Der Argumentation von Press et al. (1988) folgend, kann man dennoch auf die Berechnung einer zweiten Ableitung verzichten. Die Implementierung nutzt für jeden Datenpunkt i einen individuellen und eingeschränkten Parameter $\lambda \in [10^{-4}, 10^4]$ (für weitere Details s. Press et al. 1988).

7.3.1 Vorverarbeitung der Unähnlichkeiten

Wegen der Nichtlinearität in Gl. 7.12 hat die Transferfunktion $T_{disp}(\cdot)$ in Gl. 5.73 mehr Einfluß im $\mathbb{H}^2$. Betrachtet man zum Beispiel die Reskalierung der Unähnlichkeiten

$$\mathbf{D}_{ij} = \mathbf{D}(\delta_{ij}) = \alpha\, \delta_{ij} \qquad \alpha > 0, \tag{7.17}$$

hat dies im Euklidischen keinen visuellen Struktureffekt – es erfolgt lediglich eine Vergrößerung um den Faktor α. Im Gegensatz zum $\mathbb{H}^2$, wo

eine α-Skalierung dazu führt, dass die Daten „mehr Krümmung“ in einem vergrößerten (exponentiell wachsenden) Raumbereich „spüren“. Das optimale α hängt von der Datenverteilung, ihrer Unähnlichkeitsstruktur und der Visualisierungsaufgabe ab.

7.3.2 Beispiel: der *Iris*-Datensatz

Als erstes Beispiel dient ein Klassiker, Fischers Iris-Datensatz, der 150 Blumen aus drei Klassen beschreibt: *iris setosa* („Δ“), *iris versicolor* („$\times$“), und *iris virginica* („+“). δ ist hier die euklidische Paardistanz in den vier Beschreibungsgrößen Länge und Breite jedes Blattes und der Blüte. Abb. 7.13 (a-d) zeigt deutlich die Separation der drei Klassen in der hyperbolischen Ebene. Einige Navigationsschnappschüsse illustrieren den Fokus-und-Kontext-Effekt. Ferner kann man den α-Skaliereffekt anhand zweier Werte (*a* und *b,c,d*) erkennen.

7.3.3 Beispiel: der *Animals*-Datensatz

Abb. 7.14 präsentiert den zweiten Datensatz und demonstriert die Rekonstruktion von semantischer Nähe. Der Datensatz beschreibt 13 binäre Merkmale verschiedener Tierklassen (s. Bildtext und *UCI Machine Learning Repository*). Je mehr Eigenschaften Tierarten teilen, desto verwandter sind sie und desto näher werden sie in der Regel von HMDS-Verfahren platziert. Es entsteht ein semantisch sinnvolles Artenbild.

Abb. 7.15 zeichnet den verbleibenden Reststress $E(\alpha)$ in Abhängigkeit zum Skalierfaktor α (Gl. 7.17).

7.3.4 Beispiel: Zufallsbäume in 200 Dimensionen

Abb. 7.16 zeigt einen hierarchischen Clusterdatensatz mit 280 Punkten. Sie stellen die Knoten eines dreistufigen Zufallsbaumes dar, der in den $\mathbb{R}^{200}$ ragt. Die Richtungen sind zufällig gewählt mit einem Teilungsfaktor (*branching factor*) von 5-5-10 und Zweiglängen von 2, 1, 5 und 0.5 je Stufe.

Zur Identifikation können die Knoten optional beschriftet werden. Mit Hilfe der Fokus+Kontext-Darstellung im $\mathbb{H}^2$ lassen sich bequem auch die

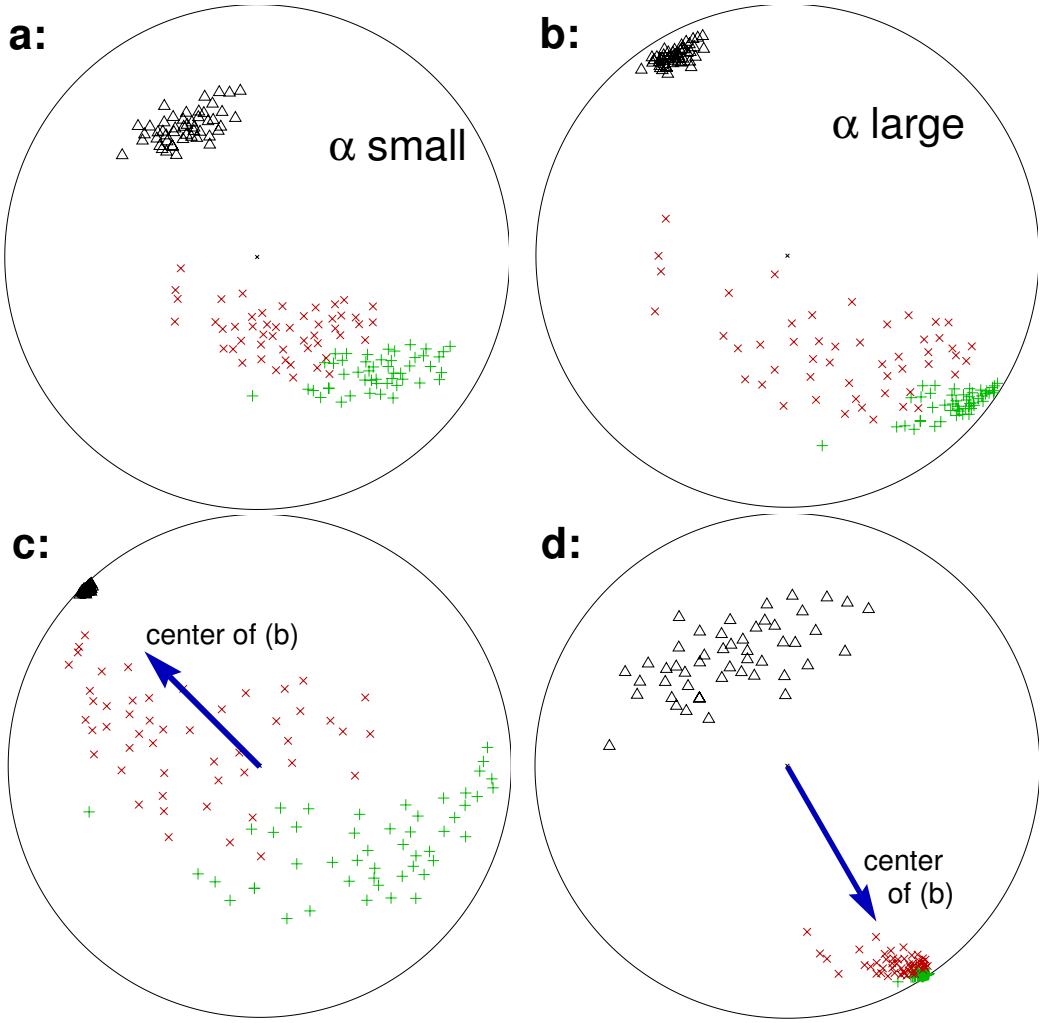


Abbildung 7.13: *(a–d) Iris-Datensatz*: Die Cluster der drei Irisarten sind deutlich an den Markern erkennbar. *(a)* hat einen kleineren α Wert als *(b)*. Während *(a)* einem konventionellen, euklidischen MDS ähnelt, nimmt *(b–d)* einen breiteren (exponentiell wachsenden) Raum im $\mathbb{H}^2$ ein, der durch Fokusverschiebung mittels Mausinteraktion exploriert werden kann. Die Pfeile markieren die Verschiebung des anfänglichen *(b)*-Ursprungs.

Details in Außenbereichen inspizieren. Die statischen Ausnahmen in Abb. 7.16 können die interaktive Dynamik leider nur andeuten.

Wie vorteilhaft ist die hyperbolische Einbettung im Vergleich zu einer euklidischen? Um diese Frage zu beantworten, lässt sich der Stress $E(\{\mathbf{x}_i\})$ (Gl. 5.82) studieren. Mit wachsendem Skalierfaktor α (Gl. 7.17) er-

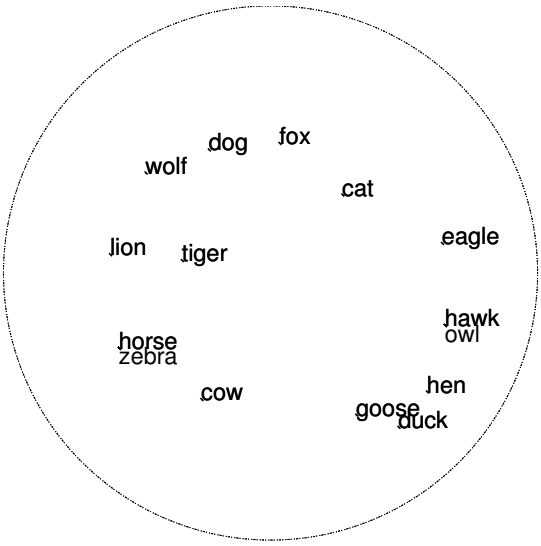


Abbildung 7.14: Der ***animals dataset*** beschreibt 16 Tierarten mit 13 binär kodierten, biologischen Merkmalen (klein, mittel, groß, zweibeinig, vierbeinig, hat Haare, hat Hufe, hat Mähne, jagt, läuft, fliegt, schwimmt). Zwei Paare habe identische Beschreibungen und sind daher auf denselben Ort abgebildet: *owl+hawk* in 4 Uhr-Richtung und *horse+zebra* in 8 Uhr-Richtung.

fahren die Daten mehr und mehr gekrümmten Raum. In Abb. 7.17 *oben links* erkennt man ein Minimum $E_{I\!H^2}\text{min} = 0.0095$ für $\alpha = 2.9$, das etwa fünfmal kleiner ist als für konventionelles *Sammon-mapping* $E_{\mathbb{R}^2}\text{min} = 0.057$ im flachen $\mathbb{R}^2$.

Für den optimalen α-Punkt sind in Abb. 7.17 *rechts* die Disparitäten vor und nach der HMDS-Abbildung (D_{ij}, d_{ij}) als Punktwolke geplottet. Die Regressionsgerade zeigt einen Pearson's Koeffizienten $R_{H^2} = 0.901$, welcher deutlich besser ist als der des konventionellen MDS Ergebnisses mit $R_{R^2} = 0.702$. *Unten links* wird die Entwicklung des Korrelationskoeffizienten $R(\alpha)$ aufgezeigt. Die Quintessenz ist, dass der Datensatz klar von der nicht-euklidischen Einbettung profitiert.

Vergleich mit anderen Visualisierungsmethoden: Abb. 7.18a zeigt denselben Datensatz als Zufallsprojektion *links* und Abb. 7.16b *rechts* als PCA-Projektion in den beiden ersten Achsen. Während die Clusterfeinstruktur im $I\!H^2$ deutlich wird, ist dies in den beiden Darstellungen in Abb. 7.18 durchaus nicht der Fall. Die PCA-Darstellung lässt die Grobstruktur der Cluster erkennen, allerdings sind die visuellen Interclusterabstände strikt projektionsbedingt und damit in hochdimensionalen Räumen relativ arbi-

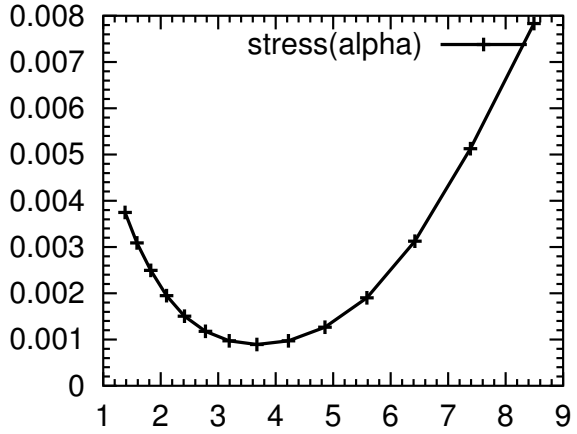


Abbildung 7.15: Der Reststress E_{H^2} für den Animal-Datensatz versus dem Unähnlichkeits-Skalierfaktor α (Gl. 7.17): Das Minimum von $E_{H^2} = 0.00089$ bei $\alpha = 3.7$ ist kleiner als ein Fünftel des Reststresses für die flache, euklidische Einbettung $E_{R^2} = 0.0048$.

trär.

7.4 Verteilungen in hochdimensionalen Räumen

Sehr merkmalsreiche Datensätze sind inhärent hochdimensional. Damit sind einige Eigentümlichkeiten verknüpft. Eine weithin bekannte ist die „Volumenanreicherung“ in den Außenbereichen einer Verteilung. Betrachtet man eine uniforme Kugelverteilung im $\mathbb{R}^m$, so ist die Oberfläche einer Kugelschale mit Radius r gleich dem Volumen einer $\mathbb{R}^{m-1}$-Kugel und damit der relative Volumenanteil gleichdicker Kugelschalen proportional zu r^{m-1}. Dies führt bei großen m dazu, dass praktisch das gesamte Volumen in einer dünnen Außenschale enthalten ist. Welche Konsequenzen hat dies für die Verteilung von Punktabständen im Allgemeinen? Mit Hilfe der Methode der Monte-Carlo-Simulation wurden verschiedene $\mathbb{R}^m$ Zufallsverteilungen daraufhin untersucht: *(i)* i.i.d. Gauß'sche Radialverteilungen, *(ii)* uniforme Verteilungen in der Einheitskugel, *(iii)* auf deren Oberfläche, *(iv)* innerhalb eines m-dimensionalen Hyperwürfels und *(v)* auf dessen Ecken.

Abb. 7.19 illustriert den ersten Fall als Histogramm von Paarabständen. Mit wachsendem m verschiebt sich die Verteilung zu größeren δ Werten.

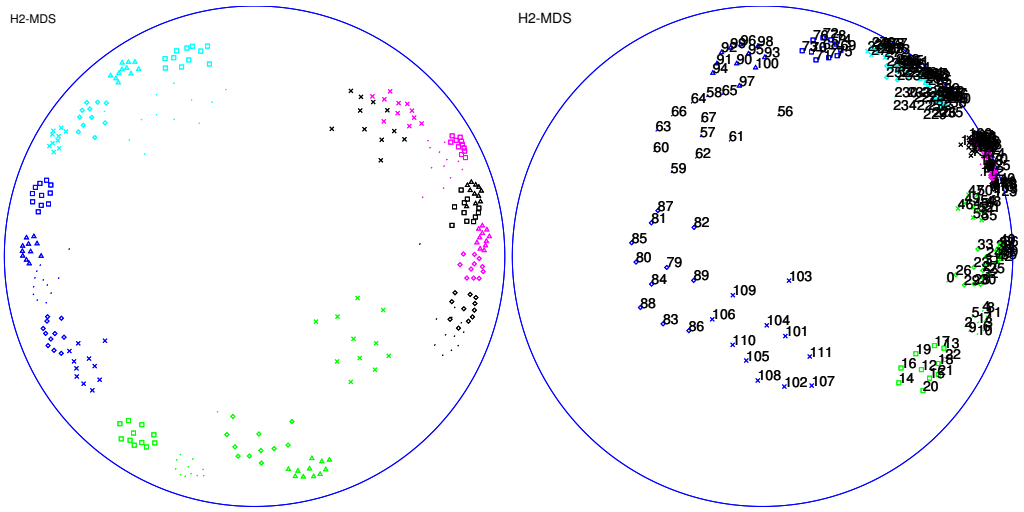


Abbildung 7.16: Der dreistufige „Zufallsbaum“ im $\mathbb{R}^{200}$ HMDS projiziert in den $\mathbb{H}^2$. Dargestellt sind zwei Navigationsschnappschüsse mit und ohne Beschriftung (systematisch dreiziffrig nach Zweignummer). Die einzelnen Blattgruppen im Baum sind klar erkennbar.

Bemerkenswert ist, dass die Form und Breite im Wesentlichen erhalten bleibt. Für die Fälle *(ii–v)* verhält es sich, bis auf Diskretisierungseffekte, bei kleinen m ganz analog. Dies legt eine Verallgemeinerung in hochdimensionale Datenverteilungen nahe, die nicht in niedrigdimensionalen Unterräumen liegen.

Die Konsequenz ist, dass in hochdimensionalen Verteilungen nahe Datenpunkte extrem selten sind. Wie in Abb. 7.19 an der Modenverschiebung deutlich wird, zeigen sie einen gewissen „Grundabstand“, der eine stark repulsive Wirkung entwickelt (da in Gl. 5.82 quadratisch wirkend), wenn sich Paare zu nahe kommen. Die Konsequenz ist mehrfach beschrieben worden (z.B. deLeeuw und Stoop 1986; Klock und Buhmann 1997): Das optimale MDS-Einbettungsergebnis zeigt mit steigender m-Dimensionalität die **Tendenz ringförmige Strukturen** auszubilden. Unglücklicherweise birgt diese von Repulsion bestimmte Situation auch noch sehr viele lokale Minima, die es dem Standard-Sammon-Algorithmus schwer machen, ein optimales Ergebnis zu liefern. Manche Standardimplementationen haben hier ernste numerische Probleme (z.B. *sompak*). Spezielle Verfahren, wie deterministisches *Annealing*, können sich hier profilieren (Klock und Buhmann 1997).

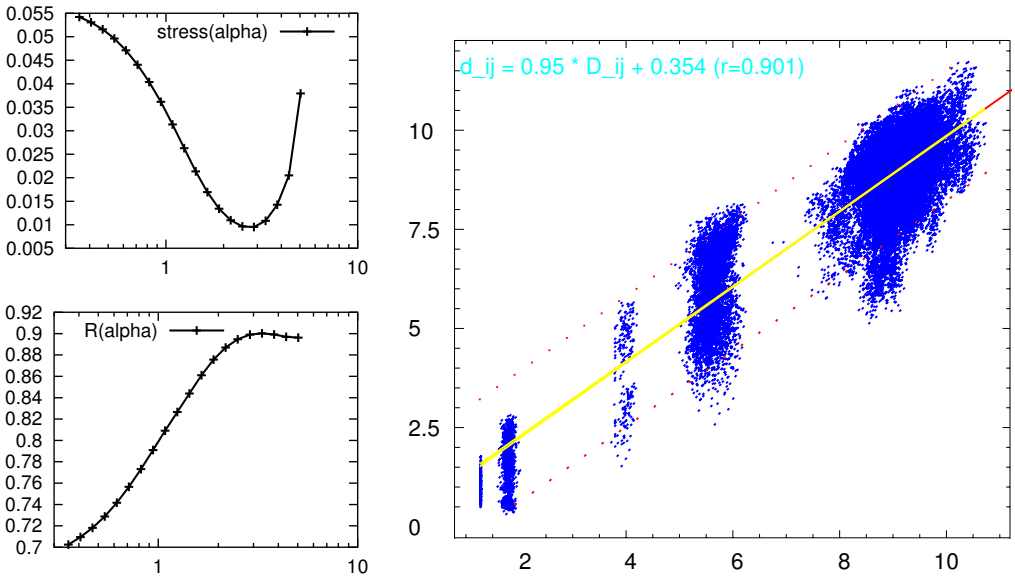


Abbildung 7.17: HMDS-Beispiel „Zufallsbaum“: *(a, links oben:)* Der Reststress $E_{\mathbb{H}^2}$ versus Skalierfaktor α zeigt ein Minimum bei $\alpha = 2.9$. *(b, rechts:)* Scatterplot der Paarabstände d_{ij} im $\mathbb{H}^2$ versus den (mit $\alpha = 2.9$ skalierten) *dissimilarities* D_{ij}. Die Regressionsgerade wurde mit $R = 0.901$ angepasst. *(c, unten links:)* Der Korrelationskoeffizient $R(\alpha)$ zeigt ein etwa mit $E(\alpha)$ korrespondierendes Maximum.

Wie nimmt der $\mathbb{H}^2$ einen solchen Datensatz mittels HMDS auf? Abb. 7.20 zeigt die beiden MDS-Abbildungen, *links* im $\mathbb{R}^2$ und *rechts* im $\mathbb{H}^2$, zusammen mit den Disparitäts-Distanz-Scatterplots $\{\mathbf{D}_{ij}, d_{ij}\}$. Der euklidische Reststress ist hier um 40 % höher ($E_{\mathbb{R}^2} = 0.40$) als bei hyperbolischer Einbettung ($E_{\mathbb{H}^2} = 0.285$ bei $\alpha = 0.33$), was in Abb. 7.20 (c,d) nachvollziehbar ist.

Es stellt sich heraus, dass der $\mathbb{H}^2$ durchaus nicht frei von lokalen Minima ist, aber er bietet im Vergleich zum $\mathbb{R}^2$ offenbar mehr Wege, diese lokalen Minima zu vermeiden. Damit kann im hyperbolischen Raum eine gute Lösung einfacher gefunden werden.

Im Kontext des nächsten Anwendungsbeispiels (Abs. 8.1.2) werden Kontrastmodulations-Verfahren vorgestellt, die Unterschiede von Disparitäten in hochdimensionalen Daten besser hervortreten lassen.

Wie sieht die Distanzverteilung einer hyperbolischen Gleichverteilung aus? Diese Frage ist in gewisser Weise die Umkehrung der vor-

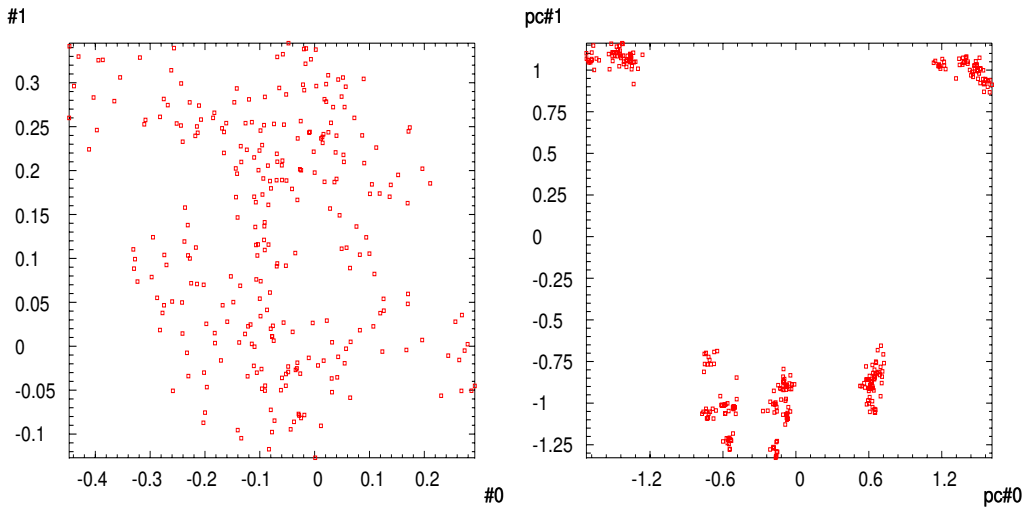


Abbildung 7.18: *(a, links:)* 2D-Zufallsprojektion des „Zufallsbaums“ in $\mathbb{R}^{200}$. *(b, rechts:)* PCA-Projektion in die Ebene, die durch die beiden stärksten Eigenvektoren aufgespannt wird. Die genauere Struktur bleibt verborgen. Beide Verfahren setzen – im Gegensatz zum HMDS – eine Vektorrepräsentation der Daten voraus.

herigen Fragestellung. Nun ist eine uniforme hyperbolische Verteilung das Ziel, die mittels Gl. 6.29 innerhalb eines Kreises mit dem Poincaré-Radius r erzeugt wird. Abb. 7.21 gibt Aufschluss darüber, welche Dissimilaritätsverteilung δ dazugehört. Mit zunehmendem Poincaré-Radius r (Scharparameter) verschieben sich die δ zu höheren Werten und die Verteilung wird linksschief. Die Verschiebung des Mittelwertes und der Mode von δ wird *rechts* in Abb. 7.21b deutlich. Der semi-logarithmische Plot zeigt zusätzlich den hyperbolischen Durchmesser $2\rho(r)$ und die exponentiell wachsende Fläche $A(r)$ in $\mathbb{H}^2$ (Gl. 6.23).

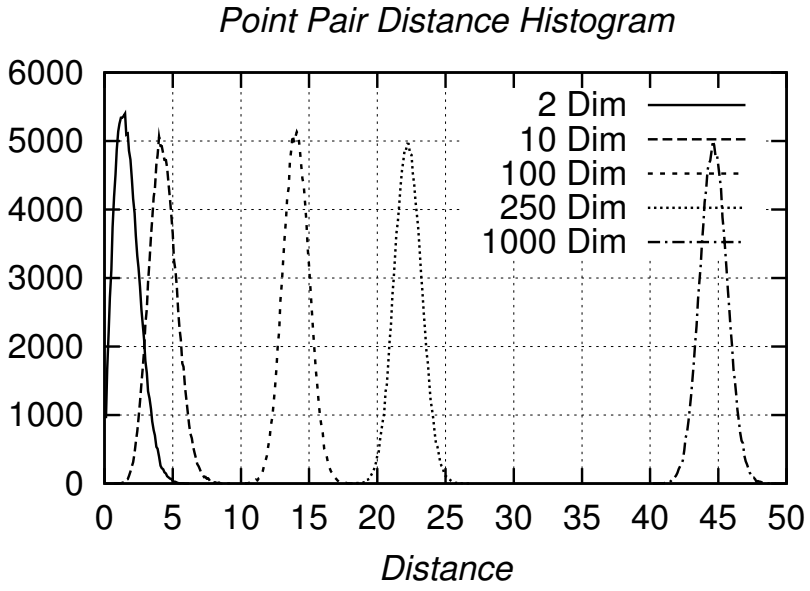


Abbildung 7.19: Histogramm von euklidischen Distanzen δ zwischen Punktpaaren einer sphärischen Gaußverteilung der Dimension $m \in \{2, 10, 100, 250, 1000\}$ (124 750 Paare von 500 Zufallspunkten mit Einheitsvarianz; Topfbreite 0.1).

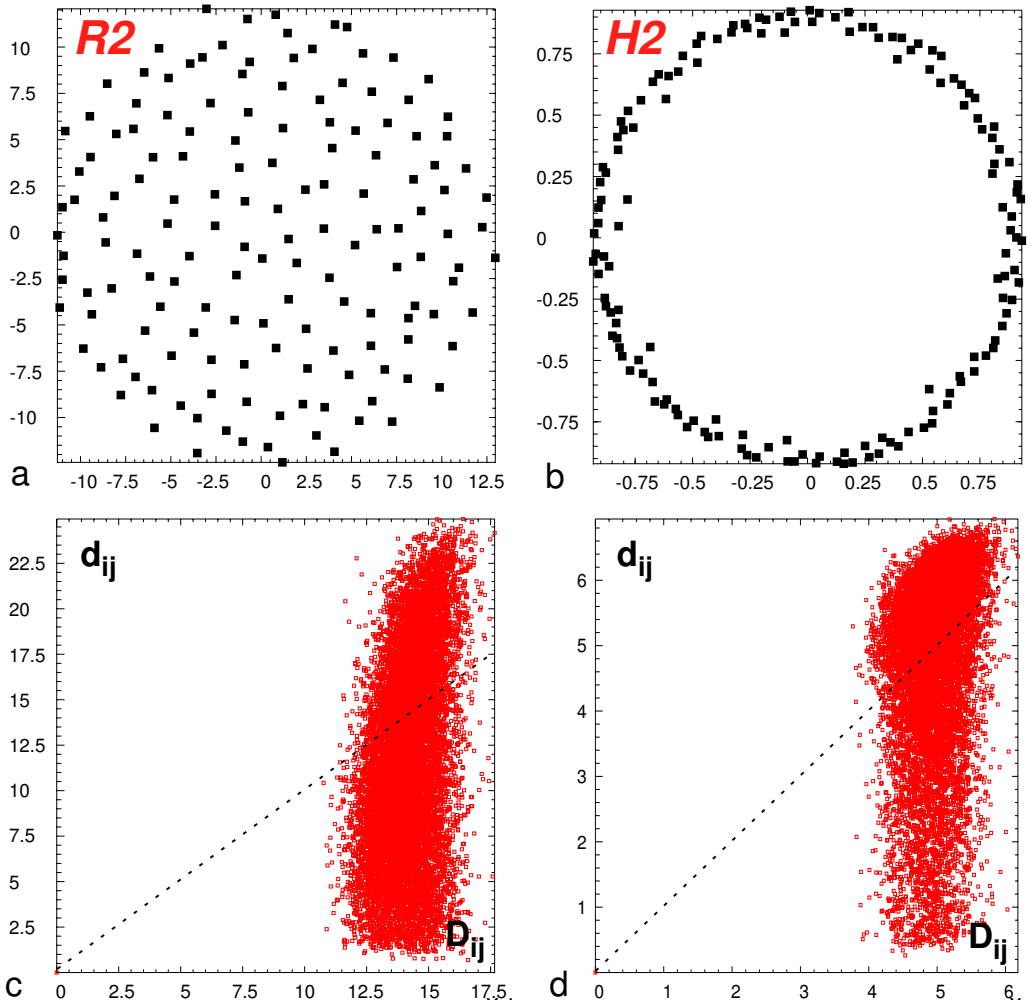


Abbildung 7.20: *(Top)* Multidimensionale Skalierung von 150 Gauß verteilten Punkten in $m = 100$ Dimensionen in den *(a)* $\mathbb{R}^2$ und *(b)* in den hyperbolischen $\mathbb{H}^2$. *(Oben:)* MDS im euklidischen Raum *(a)* fällt viel leichter in ein lokales Minimum. Die rechte Seite *(b)* zeigt die, wie erwähnt, optimale, ringförmige Struktur. Offensichtlich bietet der exponentiell wachsende Raum im $\mathbb{H}^2$ den kritischen Zusatzraum, um dem iterativen MDS-Prozess zu ermöglichen, die lokalen Minima zu vermeiden. *(Unten:)* Die beiden zugehörigen Scatterplots zeigen die Unähnlichkeiten (*dissimilarities*) $\mathbf{D}_{ij}$ und die Bilddistanzen d_{ij}. Die optimale Abbildung ist stressarm und approximiert die gestrichelte Linie. *(d)* Im $\mathbb{H}^2$ liegt die Mehrheit näher an der Optimalform als im linken Diagramm, *(c)* wo die d_{ij} tendenziell unter der Diagonalen liegen.

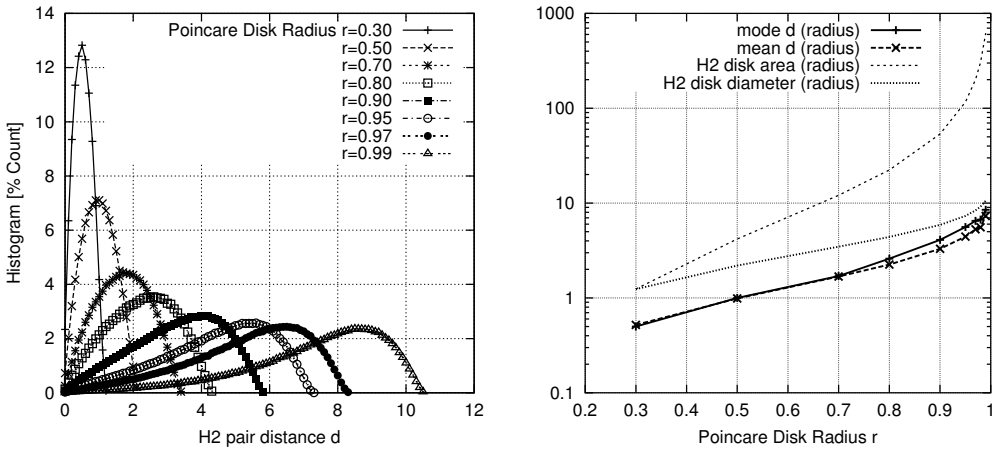


Abbildung 7.21: *(a, links:)* Histogramm der hyperbolischen Distanzen δ zwischen Punktpaaren, die aus einer kreisförmigen uniformen $\mathbb{H}^2$-Verteilung gezogen wurden (s. Abs. 6.4.2). Die Ordinatenskala gibt den Prozentwert der einen Million Werte an, die in Töpfe der Breite 0.1 gruppiert wurden. *(b, rechts:)* Die unteren Linien verfolgen die Mode und den Mittelwert der δ Verteilung in Abhängigkeit des Kreisradius r (in P). Wegen der Linksschiefe ist die Mode größer als der Durchschnitt, aber kleiner als der hyperbolische Durchmesser des Kreises $2\rho(r)$, wobei der relative Abstand mit r schrumpft.

Kapitel 8

$\mathbb{H}^2$-Navigation in Dokumentkollektionen mit hybrider Architektur

Das nächste Anwendungsbeispiel führt in das Gebiet des *information retrival*, des *text minings* und der integrierten Präsentation von Textdokumenten. Kulturhistorisch ist die Erfindung des geschriebenen Wortes als Wissensspeicher eine kaum zu überschätzende Errungenschaft. Schrift und somit auch Text repräsentiert Wissen und Information, macht sie haltbar, transportierbar, studierbar, nachschlagbar und, dank Gutenberg, effizient kopier- und verbreitbar. Die junge Verbreitung der Informationstechnologie und des Internets setzen hier historisch nochmals neue Maßstäbe: Information zu speichern, zu duplizieren, in Windeseile und über beliebig weite Strecken zu transportieren wurde so günstig wie noch nie. Abgesehen von Grundinvestitionen, rangieren die betriebswirtschaftlich marginalen Kosten nahe Null – was die technische Seite betrifft. Umgekehrt nimmt die relative Bedeutung der Ressource Zeit zu und damit die Wichtigkeit, mit großen und rapide wachsenden Informationsbeständen so umzugehen, dass die kognitive Leistungsfähigkeit des Benutzers optimal unterstützt wird.

Unstrukturierter Text steht neben Bilddatenbanken in den folgenden beiden Anwendungsbeispielen im Vordergrund. Wie kann der Benutzer in effektiver Weise in Dokumentbeständen den Überblick gewinnen und darin navigieren?

Im Verlauf dieses Kapitel werden u.a. Skalierungseigenschaften für

sehr große Dokumentkollektionen diskutiert und daraus wird eine Hybridarchitektur entwickelt und vorgestellt.

8.1 Anwendungsbeispiele: Navigation im *Space of Movies*

Der Dokumentkorpus im ersten Beispiel besteht aus textuellen Filmkritiken aus einer Internetquelle, i.e. der Newsgroup *news://rec.art.movies.reviews*, in der jedermann seine Meinung über einen Film darlegen kann. Etwa dreizehntausend wurden von Jörg Ontrup (2001) aufbereitet und mit strukturierter Hintergrundinformation aus der *internet movie database* verknüpft. Dies erschließt die Möglichkeit, zu einem Filmtitel u.a. das Erscheinungjahr, den Produzenten und eine offizielle Filmkategorisierung abzurufen.

Das Anwendungsszenario ermöglicht die Navigation und das *browsen* im „Raum“ der Filme. Ziel ist eine Anordnung zu finden, die ähnliche Filme auch in räumlicher Nachbarschaft anzeigt. Die einzelne Filmkritik tritt in den Hintergrund.

8.1.1 Repräsentation der Filme

Mittels des *Bag-of-word*-Modells, das in Abs. 2.4.2 erläutert ist, wurde ein Wörterbuch mit 5084 Termen generiert, woraus ein 5084-dimensionaler Merkmalsraum nach dem TFIDF-Schema resultiert (s. Abs. 2.4.2).

Bei genauerer Inspektion stellt sich heraus, dass etliche Beschreibungstexte nur wenige Worte enthalten und daher nur spärliche Informationen liefern. Erschwerend kommt hinzu, dass die Texte von völlig verschiedenen Autoren stammen und keine Systematik im Dokumentaufbau vorliegt.

Daher wurde eine Beschränkung auf all jene Filme vorgenommen, die von mehr als fünf *reviews* beschrieben wurden und die den Kategorien *animation* oder *science-fiction* angehörten. Im nächsten Schritt wurde für jeden der 132 Filme der *mittlere Filmmerkmalsvektor* ermittelt, indem alle beschreibenden Texte zusammengefasst wurden. Durch die Mittelwertbildung

$$\vec{f}_m = \sum_{t \in A} \vec{f}_t\,, \quad A = \{t |\, t \text{ beschreibt Film } m\}. \tag{8.1}$$

der normierten Merkmalsvektoren (Gl. 2.17) wird eine Informationsaggregierung erreicht, die gleichzeit die Autorenunabhängigkeit verbessert.

8.1.2 Modulation des Ähnlichkeitskontrastes: zwei Beispiele

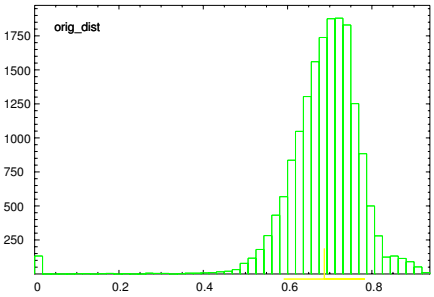


Abbildung 8.1: *(Oben links:)* Histogramm der Unähnlichkeiten δ_{ij} für die ausgewählten Filme.

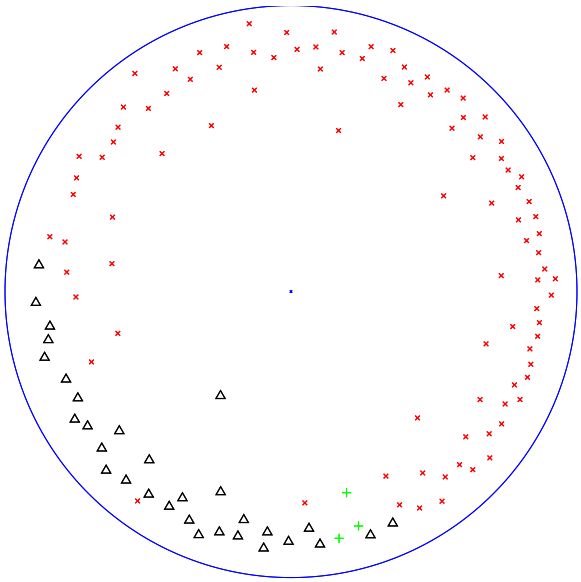


Abbildung 8.2: *(Oben rechts:)* Hyperbolisches MDS-Ergebnis für die lineare Disparitätstransferfunktion (Gl. 2.16 und Gl. 7.17). Die typische „Ringstruktur“ wird durch Hochdimensionalität der Daten verursacht (vgl. Abb. 7.20). Die (roten) „$\times$“ markierten *Science-fiction*-, die (schwarzen) „Δ“ *Animation*- und die (grünen) „+“ markierten Filme, die beiden Genres angehören. Natürlich ist die Genreinformation nur in der Evaluationsphase eingesetzt worden und stand dem HMDS-Verfahren nicht zur Verfügung.

Wie verhält sich die Unähnlichkeitsstruktur der selektierten Filme zu den hochdimensionalen Gaußverteilungen, wie in Abb. 7.19 dargestellt? Vergleicht man das Histogramm der Unähnlichkeiten in Abb. 8.1, so erkennt man, dass die effektive Dimension wesentlich kleiner ist als die 5084 Dimensionen des Filmmerkmalsvektors $\vec{f}$.

In Abb. 8.2 wird die typische Ringstruktur deutlich, die schon Abb. 7.20b geprägt hat. Die Markertypen erlauben bereits einen Überblick der Abbildungsqualität. Die Genreinformation wurde hier für die nachträgliche Markierung genutzt: die „Δ“-Marker für *Animation*-File sind wohlsepariert von

der „$\times$“-markierten *Science-fiction*-Gruppe. Die drei „+“-markierten Filme gehören zu beiden Kategorien und liegen im Übergangsbereich.

Während die Ringkonfiguration die Struktur ist, die die Distanzstruktur am besten erhält, ist sie nicht unbedingt die beste vom Standpunkt der Visualisierungsaufgabe.

Wie bereits in Abb. 7.19 deutlich wurde, ist die Distanzverteilung mit steigender Dimensionalität systematisch zu größeren Werten verschoben. Inspiriert durch Arbeiten von Klock und Buhmann (1997) wurde mit einigen Gegenstrategien experimentiert. Die Kernidee ist der Einsatz einer nichtlinearen Transferfunktion $T_{disp}(\cdot)$, die die Mode der Unähnlichkeitsstruktur effektiv in einen niedrigdimensionalen Bereich transformiert. Damit wird der relative Unähnlichkeitsabstand von Datenpaaren verstärkt, was die Bezeichnung „Kontrastverstärkung“ motiviert. Im Folgenden werden zwei Beispiele untersucht.

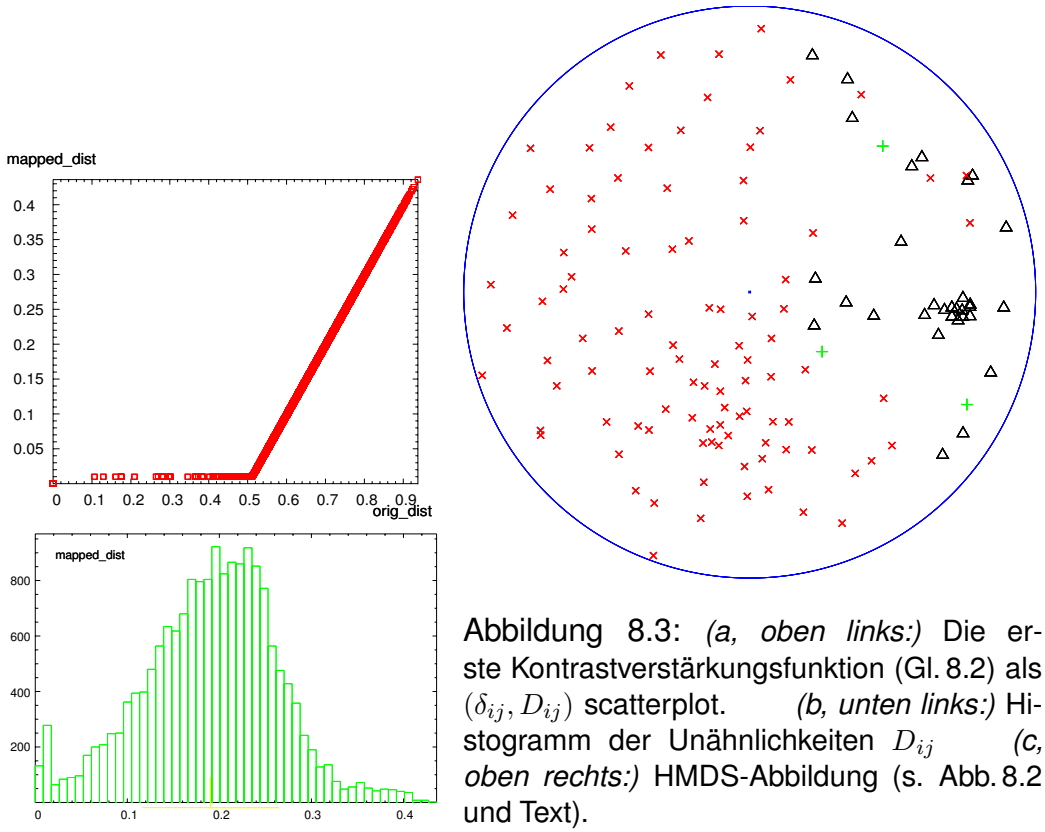


Abbildung 8.3: *(a, oben links:)* Die erste Kontrastverstärkungsfunktion (Gl. 8.2) als (δ_{ij}, D_{ij}) scatterplot. *(b, unten links:)* Histogramm der Unähnlichkeiten D_{ij} *(c, oben rechts:)* HMDS-Abbildung (s. Abb. 8.2 und Text).

Die erste Transferfunktion ist eine einfache lineare Verschiebung zu kleineren Unähnlichkeitswerten mit zusätzlicher Minimalbegrenzung auf

D_A.

$$\mathbf{D}_{ij} = \alpha \max(\delta_{ij} - \delta_A\,,\, D_A). \tag{8.2}$$

Abb. 8.3a exponiert die Transferfunktion für

$$\delta_A = Q_{\delta^+}(0.01)\ .$$

Durch diese Parameterwahl der 1 %-Perzentile der δ_{ij}^+-Verteilung gewinnt man eine Selbstanpassung der Transferfunktion an die auftretenden δ_{ij}^+-Werte (wobei alle $\delta_{ii} = 0$-Werte ignoriert werden).

Das *clipping* auf einen Minimalwert (>0), hier $D_A = 0.01$, sorgt für eine minimale repulsive Wirkung zwischen Datenpaaren. Die resultierende Unähnlichkeitsverteilung zeigt in Abb. 8.3b eine deutlich sichtbare Spitze (die 1 % aller Nicht-Null-Paarungen enthält). Das HMDS-Bild (Abb. 8.3c) hat die Ringgestalt (Abb. 8.2) aufgegeben und exponiert einige dichte Cluster.

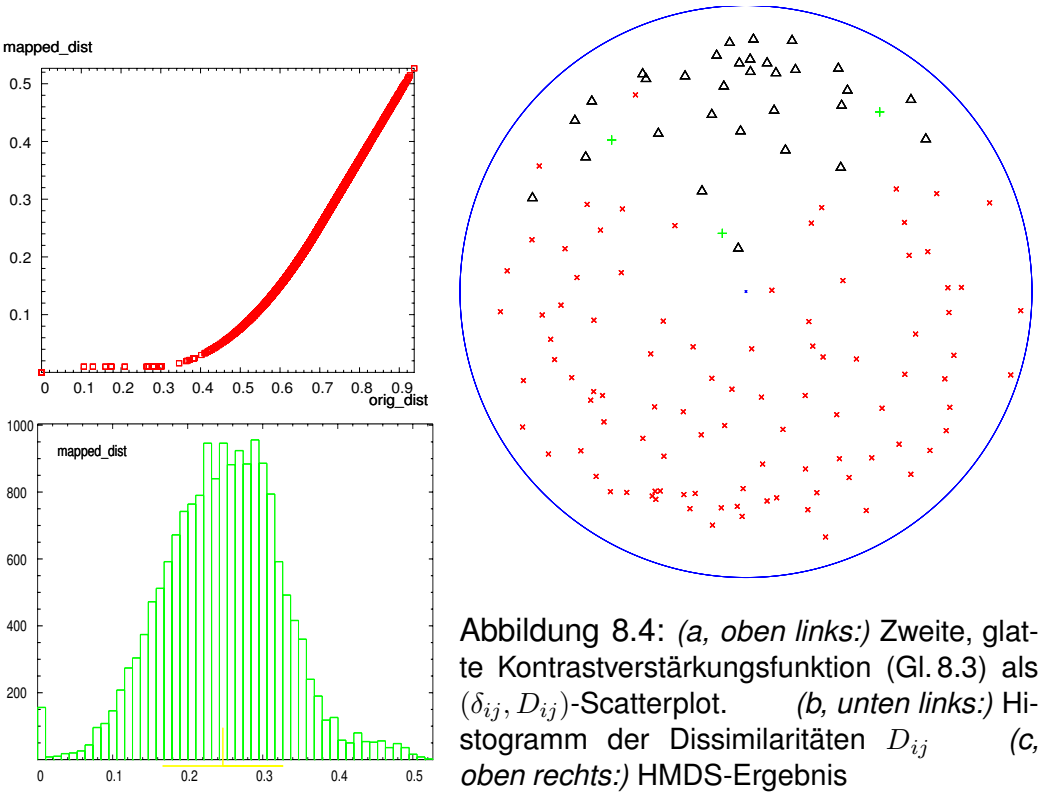


Abbildung 8.4: *(a, oben links:)* Zweite, glatte Kontrastverstärkungsfunktion (Gl. 8.3) als (δ_{ij}, D_{ij})-Scatterplot. *(b, unten links:)* Histogramm der Dissimilaritäten D_{ij} *(c, oben rechts:)* HMDS-Ergebnis

Die zweite Transferfunktion zielt auf eine glatte, knickfreie Kontrastver-

stärkungsfunktion. Die Funktion ist als stückweises Polynom

$$D_{ij} = \alpha \left\{ \begin{array}{ll} 0 & \text{für } i = j \\ D_A & \text{für } \delta_{ij} \leq \delta_A \\ D_A + c_1(\delta_{ij} - \delta_A)^2 & \text{für } \delta_A < \delta_{ij} \leq \delta_B \\ c_2 + c_3\,\delta_{ij} & \text{für } \delta_B < \delta_{ij} \end{array} \right\} \tag{8.3}$$

so entworfen , dass es an den Nahtpunkten $A = (\delta_A, D_A)$ und $B = (\delta_B, D_B)$ auch eine stetige Ableitung hat. Dies wird durch

$$c_1 = \frac{D_B - D_A}{(\delta_B - \delta_A)^2}\,, \tag{8.4}$$

$$c_2 = D_B - \delta_B\, c_3 \quad \text{und} \tag{8.5}$$

$$c_3 = 2\frac{D_B - D_A}{\delta_B - \delta_A} \tag{8.6}$$

gewährleistet.

Abb. 8.4a zeigt die beiden Geradenstücke und das quadratische Mittelteil für $A = (Q_{\delta^+}(0.001), 0.01)$ und $B = (Q_{\delta^+}(0.5), 0.25)$. Wie in Abb. 8.4b zu sehen, ergibt dies hier eine glatte D_{ij}-Verteilung mit einem geringen Anteil kleiner Abstände. Das HMDS-Abbild ist ein guter Kompromiss zwischen ausgewogener Verteilung und optimaler Navigierbarkeit. Im Vergleich zu Abb. 8.3 erzeugt die glatte Funktion breitere Kohäsion und höhere Reinheit der „Δ“-markierten *Animation*-Filme. Der dichte Cluster, meist Disney-World-Produktionen, wird nun weiter und leichter unterscheidbar (s. vollbeschriftete Darstellungen Figs. 8.6ab, 8.7ab). Trotzdem sind Mikrocluster klar erkennbar und exponieren besonders starke Filmbindungen, so z.B. die beiden Δ markierten „Toy Story 1+2“-Filem (11 Uhr), das $\times$-cluster erweist sich als „Alien“-Filmfamilie (etwa 7 Uhr).
Figs. 8.6ab, 8.7ab bieten einige Schnappschüsse mit Filmtitelbeschriftung.

8.1.3 Ist die hyperbolische Einbettung letztlich vorteilhaft?

Abb. 8.5 beantwortet die Frage mit einem klaren Ja. Vergleichbar mit Abb. 7.15 und 7.17), hat die Reststresskurve ein klares Minimum > 0, d.h. bei $\alpha = 9.4$. Denn wenn die Unähnlichkeitsstruktur eine euklidische Einbettung im Sinne der E-Minimierung bevorzugen würde, hätte dies ein Minimum in der Nähe von $\alpha = 0$ erwirkt.

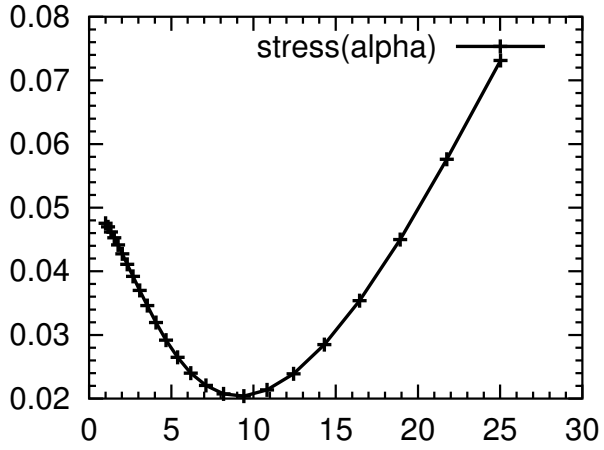


Abbildung 8.5: Der Reststress $E_{\mathbb{H}^2}$ versus dem Unähnlichkeitsskalierfaktor α in Gl. 8.3 besitzt ein Minimum bei $\alpha = 9.4$. Wäre eine euklidische Einbettung optimal, würde sich das Reststressminimum in der Nähe von $\alpha = 0$ finden. So zeigt sich eindeutig der Vorteil einer HMDS vor einer konventionellen MDS-Einbettung. Der dafür Grund ist, dass im hyperbolischen Raum mehr Platz ist, exponentiell mehr Platz um jeden Punkt.

Diskussion: Die Navigationsschnappschüsse in Abb. 8.6, 8.7 geben einen flüchtigen Eindruck der explorativen Möglichkeiten, die durch HMDS, der hyperbolische multi-dimensionale Skalierung, eröffnet werden.

Dichte Cluster sind deutlich erkennbar, z.B. die „Star Trek"-Gruppe, oder die Nähe von „Bug's Life" und „Ants"; beide handeln von Insekten. Cineasten können aus dem HMDS noch so manchen Zusammenhang entnehmen – am einfachsten durch Interaktion. Dabei sei nochmal daran erinnert, dass keinerlei Hintergrundinformation verwendet wurde. Die räumliche Anordnung wurde allein durch die Paarabstände der Filmrepräsentationen erzeugt. Diese beruhten nur auf den Worthäufigkeitseigenschaften der Texte von völlig unzusammenhängenden Autoren.

8.2 Anwendungsbeispiele: Navigation in Bildsammlungen

Im vorhergehenden Beispiel wurde versucht, einen Eindruck von den interaktiven Navigationsmöglichkeiten in einem *space of movies* zu vermitteln

Abbildung 8.6: Vier Navigationsschnappschüsse (Fortsetzung in Abb. 8.7) durch den entwickelten „space of movies“ – hier mit Titelbeschriftung. *(a oben:)* Die Ursprungslage. *(b unten:)* Durch Fokustransfer in die obere *Science-fiction* Region („×"´ Marker) kann man einige interessante Filmpaarungen entdecken, z.B. ein Band von „Star Trek Filmen in Richtung 11 Uhr. In 2 Uhr kann man die vier „Alien“-Streifen in ihrem dichten Cluster inspizieren.

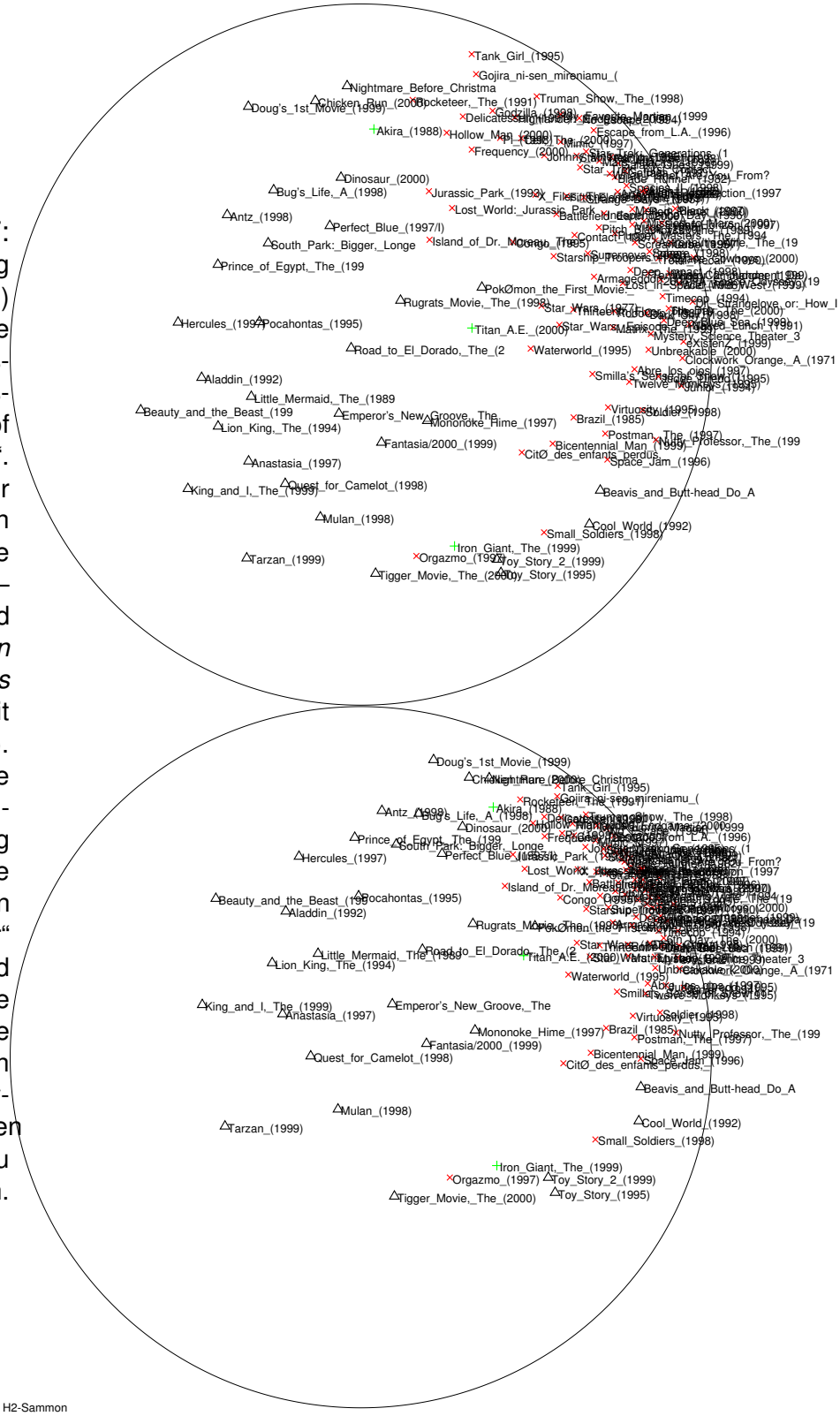


Abbildung 8.7: (Fortsetzung von Abb. 8.6) Zwei weitere Navigationsschnappschüsse im „space of movies“. *(a oben:)* Der Fokus wurde in die linke obere Ecke bewegt – in die Gegend der *animation movies* (markiert mit „Δ"´). *(b unten:)* Eine weitere Radialverschiebung lässt diese Außenregion „näher“ kommen und erlaubt eine ganze Gruppe von Wald-Disney-Animationsfilmen zu identifizieren.

(limitiert durch die Papierform). Die Darstellungen basierten auf der Textrepräsentation von im Internet verfügbarer Co-Information über die Filme. Im Prinzip kann aber jede Form von Distanzmasse integriert werden. Zum Beispiel können dies für Filmsequenzen auch Korrelationsmaße verschiedenster Art sein (s. Abs. 2.4.3).

In Abb. 8.8 wird dies anhand von Farbähnlichkeiten an Digitalbildern illustriert. Zum Einsatz kommen die in Abs. 2.4.3 erläuterten *Earth-Mover-Distance*-Farbmetriken (EMD, s. S. 24), die auf den 100 hier verwendeten Bilden berechnet werden. Deutlich wird das Clustering in Gruppen ähnlicher Farben und Motive. In einem weiteren Schritt könnten Struktur- und Textur-basierte Merkmale ermittelt und in der Distanzberechnung integriert werden.

8.3 Eigenschaftsvergleich der Layouttechniken

Wie lassen sich die drei vorgestellten Techniken zum Erzeugen eines Datenlayouts im $\mathbb{H}^2$ vergleichen? Was sind ihre Eigenschaften, Vor- und Nachteile bezüglich verschiedener Betrachtungswinkel?

8.3.1 Zulässige Typen von Eingabedaten

HTL: Die *Hyperbolic-Tree-Layout*-Technik (Abs. 7.1) erfordert, wie ihr Name schon sagt, azyklische Graphdaten als Eingabe. Bevorzugt sind ausgewogene Hierarchien mit Verzweigungsfaktoren im Bereich 4–12.

HSOM: Die *Hyperbolic Self-Organizing Map* (Abs. 7.2) verarbeitet ausschließlich Vektorrepräsentationen, im Gegensatz zum HMDS.

HMDS: das *Hyperbolic-Multi-Dimensional-Scaling*-Verfahren (Abs. 7.3) verwendet Unähnlichkeitsdaten. Da eine geeignete Distanzfunktion im Prinzip sämtliche andere Datentypen in Unähnlichkeitsdaten überführen kann und sich auch fehlende Werte gut behandeln lassen (s. Abs. 2.3), kann man Unähnlichkeitsdaten als den allgemeinsten Datentyp betrachten.

Abbildung 8.8: Navigationsschnappschuss für eine Sammlung von 100 Bildern, deren Ähnlichkeit mittels der EMD-Farbmetrik bewertet wurde. Deutlich wird die natürlich wirkende Strukturierung der Bilder, die die Navigation und Bildsuche beschleunigt. Die Bildvergrößerung ändert sich radial entsprechend der Auflösung im Poincaré-Modell Gl. 6.20. Siehe auch folgende Abb.

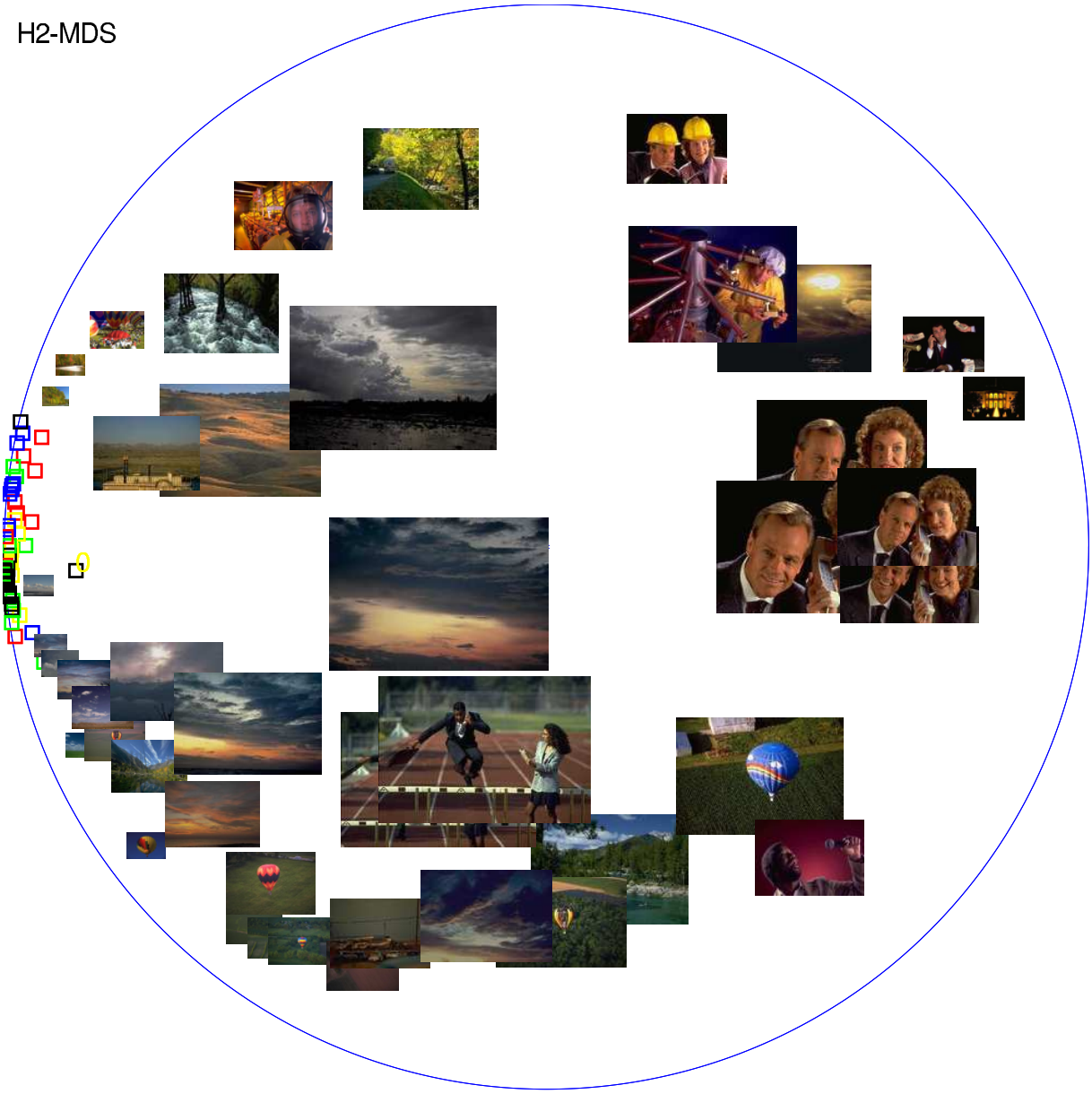


Abbildung 8.9: Wie vorherige Abb. 8.8. Hier mit stärkerer Entfaltung der Bilder durch größeres α.

8.3.2 Skalierverhalten bezüglich der Datenanzahl N

HTL und HSOM teilen den Vorteil einer linearen Skalierung in der Anzahl N der Objekte. Die HSOM skaliert auch linear in der Dimensionalität des Eingaberaumes m und der Anzahl von Knoten K.

HMDS skaliert nicht so gut und benötigt $N(N-1)/2$ Unähnlichkeitswerte aller Objektpaare. Wenn N auf mehr als wenige hundert Objekte ansteigt, wird der Layoutprozess ist für die interaktive Datenexploration zu langsam und das Ergebnis manchmal auch nicht ganz überzeugend. Die Vorausberechnung des HMDS-Layouts kann dann eine gute und responsive Lösung bieten.

8.3.3 Layoutresultat

HTL returniert die $\mathbb{H}^2$-Positionen aller Objekte, die durch rekursive Platzteilung festgelegt wurden.

Die HSOM returniert die festen $\mathbb{H}^2$-Gitterpositionen (z.B. Knoten des Tesselationsgitters Abb. 7.5). Ein Objekt wird auf denjenigen Knoten abgebildet, dessen Prototypenmerkmalsvektor dem Objektmerkmalsvektor am ähnlichsten ist. Jeder Knoten ist zum einen durch den Prototypen, den Gruppenrepräsentanten gekennzeichnet, zum anderen kann er mit deskriptiven Informationen ausgerüstet werden. Dazu kommt prinzipiell jede Information in Frage, die die Objektmenge charakterisiert, die dem Knoten zugewiesen wurde. Sie beschreibt also den „Zuständigkeitsbereich“ und kann in verschiedenste Arten graphischer und textueller Attribute bedarfsgerecht transformiert und dargestellt werden.

HMDS returniert (wie HTL) die $\mathbb{H}^2$-Positionen jedes Objektes. Die räumliche Position repräsentiert dabei Nachbarschaft und Ähnlichkeit auf der individuellen Objektebene. Damit verwirklicht die HMDS das metaphorische Konzept einer Landschaftskarte auf dem höhsten Detaillierungsgrad.

8.3.4 Neue Objekte

HTL benötigt ein partielles Neulayout aller kleinsten Teilbäume, welche die neuen Objekte enthalten.

Die HSOM bildet ein neues Objekt einfach auf den bestpassenden Knoten im Gitter ab. Die benötigte Zeit skaliert mit der Knotenzahl K, da K Vergleichsoperationen involviert sind.

Die HMDS erfordert eine erneuerte Minimierung der Kostenfunktion. Zur Beschleunigung kann das vorherige Layout als Ausgangslage verwendet werden.

8.4 Ein hybrider Ansatz zum Navigieren in großen Datenkollektionen

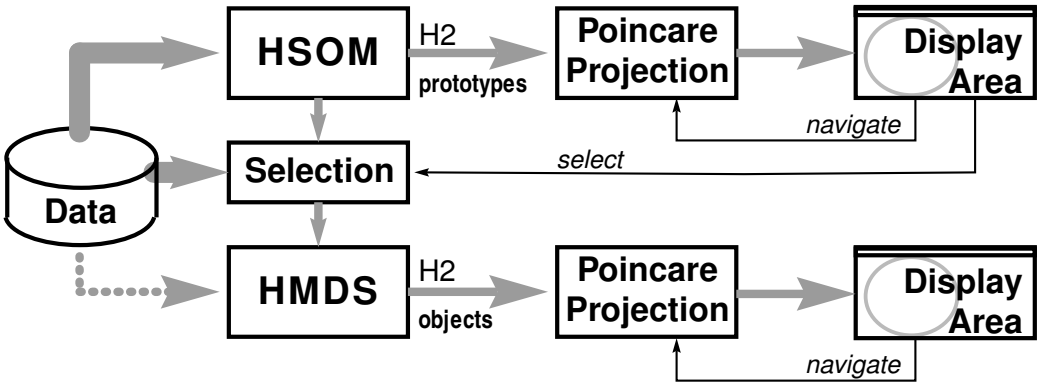


Abbildung 8.10: Die vorgeschlagene Architektur kombiniert die Vorteile der zwei $\mathbb{H}^2$-Layouttechniken: *(oben:)* Die HSOM ermöglicht eine Überblickskarte der gesamten Datenbasis; *(unten:)* Die HMDS bildet kleinere Datensätze auf eine räumlich kontinuierliche Weise ab und spiegelt dabei die Ähnlichkeitsstruktur der individuellen Objekte wider. Das Anzeigekonzept ist vereinheitlicht: beide nutzen die außergewöhnlichen Visualisierungs- und Navigationsmöglichkeiten des $\mathbb{H}^2$.

Der vorangegangene Vergleich zeigte Vor- und Nachteile der verschiedenen Layouttechniken und motiviert zu einer Synthese zum Navigieren in großen Datenkollektionen. Sie besteht aus drei Hauptkomponenten:

- der HSOM zur selbstorganisierten Bildung einer groben Themenkarte;
- der HMDS zur detaillierten Inspektion von Datenteilmengen. Hierbei wird etwaige räumliche Nähe durch die Ähnlichkeit der Objekte reflektiert;
- Das Anzeigekonzept ist vereinheitlicht und nutzt in beiden Teilbereichen die Vorteile des hyperbolischen Raumes $\mathbb{H}^2$ und der Fokus-&-Kontext-Technik.

Die hybride Architektur wird von Abb. 8.10 illustriert und im folgenden Anwendungsbeispiel demonstriert.

8.5 Anwendungsbeispiele: Navigation im Nachrichtenstrom von Reuters

Der Benchmarkdatensatz *Reuters-21578* ist eine Kollektion von Textdokumenten aus dem Reuters-Newsticker in der Zeit vom 26.02.1987 bis 20.10.1987 und wurde von David Lewis von ATT Research Labs aufbereitet (1997). Die meisten Nachrichtenartikel wurden manuell kategorisiert, d.h. einer oder mehreren von insgesamt 132 Kategorieren zugeordnet. Die zehn zahlenstärksten Kategorien wurden zu Test- und Annotationszwecken weiterverwendet (s. u. Abb. 8.13). Es wurden zwei Teilmengen gebildet, die sich an der Datumsgrenze 7.04.1987 trennen: Der erste Trainingsdatensatz ($\leq$7.4.) enthält 9603 und der zweite Testdatensatz 3299 Artikel. Dieser so genannte „ModApte“-Split wurde von mehreren Autoren verwendet (Joachims 1998; Lodhi et al. 2001). Die Textrepräsentation folgt wieder dem Konzept des Bag-of-words- oder Vektorraummodells (s. Abs. 2.4.2). Nach dem Trimmen des Wörterbuches wurden 5561 verschiedene Wortstämme herauskristallisiert (Walter et al. 2003).

8.5.1 Textkategorisierung

Zunächst wird die HSOM mit dem ersten Datensatz traininert und die Qualität der Kartenbildung anhand des zweiten Testdatensatzes unter-

sucht. Wie beurteilt man das gefundene $\mathbb{H}^2$-Layout objektiv? Die Kernidee ist die Bestimmung der dominanten Kategorien jedes Knotens anhand aller ihm zugeordneten Trainingsdokumente. Als nächstes wird die Trefferhäufigkeit für neue Dokumente aus der Testmenge ausgewertet. Dabei wird angenommen, sie gehören jeweils auch zu den Kategorien des Best-match-Prototypen. In einer detaillierteren Analyse wurden kategoriespezifische Kontingenztabellen geführt und der Break-even-Punkt der *Precision-recall*-Kurve bestimmt (s. Abs. 5.7.4). Dieser Wert erlaubt einen Vergleich mit Arbeiten aus der Literatur. Es stellt sich heraus, dass sich die HSOM hier knapp unterhalb der auf Klassifikationsleistung optimierten Verfahren ansiedelt (Näheres in Ontrup und Ritter 2001; Yang 1999) – im Gegensatz zu den Filmdaten mit vergleichbar guten Ergebnissen (Ontrup und Ritter 2001). Als Baustein des Navigationssytems (Abb. 8.10) ist die Erzeugung einer zweidimensionalen Darstellung aber eine sehr entscheidende Eigenschaft und kompensiert vielfach den geringen Leistungsnachteil.

Abb. 8.11 bildet eine HSOM mit inspesamt $K_{9,4} = 1306$ Knoten ab (s.a. Tab. 7.1). Die Gitterstruktur ist auf der $\mathbb{H}^2$-Poincaré-Kreisscheibe gezeichnet. Die perspektivische 3D-Projektion erlaubt, weitere Informationen in der dritten Dimension unterzubringen. Die Nadel, die jeden Knoten markiert, ermöglicht die datengetriebene Attributgestaltung, insbesondere:

- Nadellänge = Höhe;
- Kopfform = Glyphart;
- Kopfgröße = Glyphskalierung;
- Kopffarbe = Ton und Sättigung;
- Schaftzeichnung, komplexere Glyphen und Facettencolorierung.

In Abb. 8.11 wird die jeweils gruppendominierende Kategorie durch den Glyphen und seine Farbe markiert, die Höhe und Größe korrespondieren mit der Zahl der Dokumente, die vor und nach dem 7.04.1987 erschienen sind.

Vorteilhaft erscheint der Überblick über den Inhalt, auch in den Randgebieten, denn es findet hier kein Ausblenden des ∞-Randes statt und die entlegeneren Knoten umzäunen die innere $\mathbb{H}^2$-Fläche. An der regionalen Sammlung gleicher Glypharten und -farben – also gleicher Kategorien – wird bereits das Ordnungsvermögen der HSOM deutlich. Nachteilig ist die

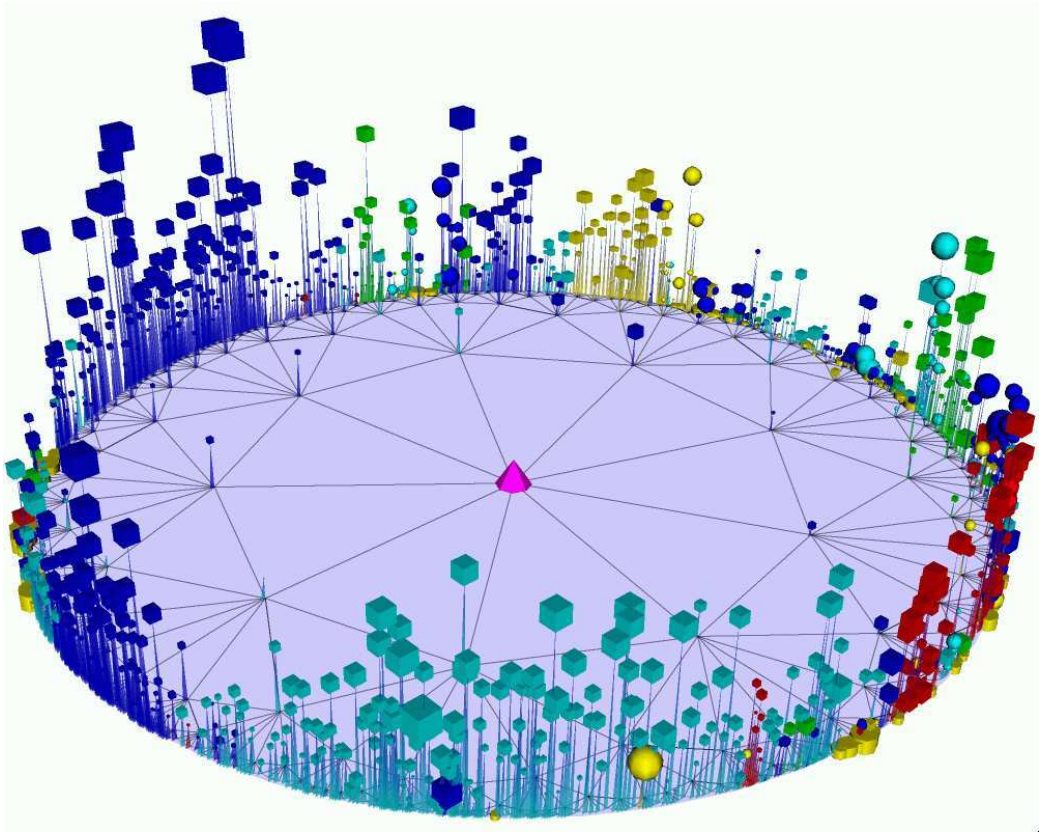


Abbildung 8.11: Das HSOM-Gitter in Poincaré-Projektion, hier perspektivisch und mit Markern dargestellte dominante Nachrichtenkategorie aus dem Reuters-21578-Korpus, wird durch den Glyph ausgedrückt und zeigt, dass semantisch sinnvolle Themencluster entstanden sind.

weite Streuung der Knoten, die auch mit weiten Navigationsstrecken verbunden ist.

Abb. 8.12a zeigt die Ursprungsstellung einer deutlich „kleineren“ HSOM mit $K_{7,3} = 161$ Knoten. Jeder Knoten erweitert nun seinen Zuständigkeitsbereich und enthält im Mittel 80 Dokumente der insgesamt 12.902. Diese Anzahl von Dokumenten kann von der nachgeordneten HMDS in Echtzeit flüssig abgebildet werden. Zunächst werden interaktiv interessante Bereiche exploriert. In 5 Uhr-Richtung steht ein „?“-bezeichneter, gelb gefärbter Kugelglyph in Umgebung von grünen Markern. Dieser soll nun exemplarisch inspiziert werden. Die Knotenwahl ist mit einer Teilmengenselektion verknüpft und bewirkt, dass die Gruppe von zugeordneten Dokumenten dem HMDS-Verfahren zugeführt wird und die Paarabstände berechnet werden.

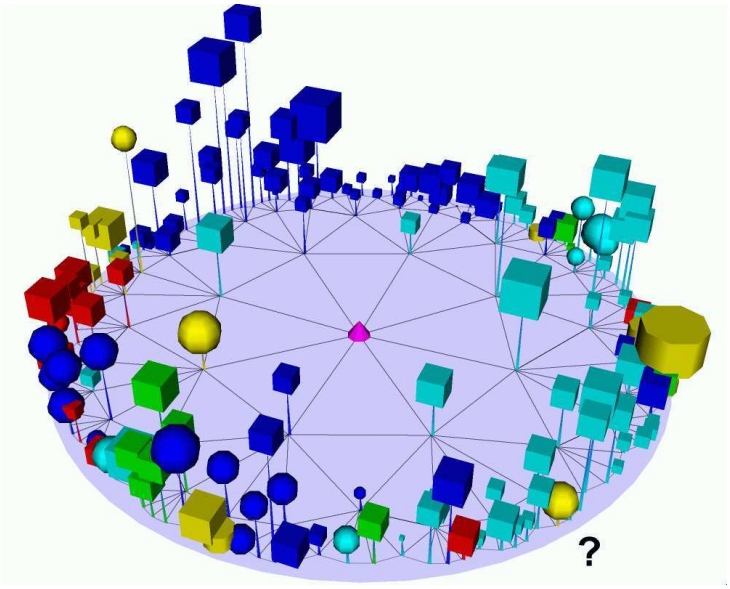
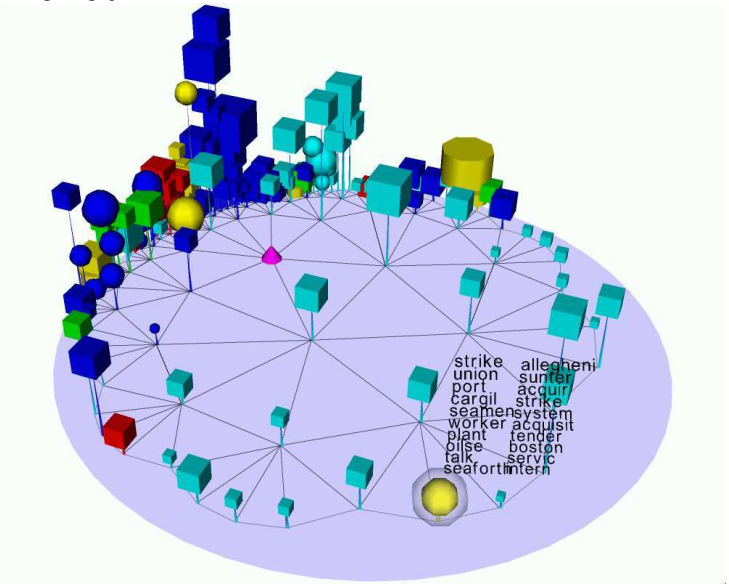


Abbildung 8.12: (a, oben:) Dieselben Daten wie in Abb. 8.11, aber mit einem kompakteren HSOM-Gitter mit $K_{7,3} = 161$ Knoten. Die Glyphgröße und -höhe korrespondiert jeweils mit der Gruppenstärke vor und nach dem Stichtag 7.04.1987. (b, unten:) Weiterer Navigationsschnappschuss: Der vorher mit „?“ markierte Knoten ist hier näher im Blickfeld und farblich hervorgehoben. Die *Keyword-probe*-Technik schreibt die wichtigsten Schlüsselwörter auf eine virtuelle Flagge über den gewählten Knoten. Hier zeigen sie an zwei benachbarten Knoten deren thematische Ähnlichkeit.



Die Abb. 8.12b zeigt noch ein weiteres Inspektionswerkzeug, den ***keyword probe***. Die gewünschten Knoten werden textuell beschriftet. Hier mit den zehn Schlüsselwörten, die den Knoten kennzeichnen. Technisch sind dies die Wörterbucheinträge (Wortstämme), die mit den zehn stärksten Komponenten des betreffenden Prototypenmerkmalsvektors korrespondieren. Der *keyword probe* für den Nachbarknoten lässt wieder die thematische Nachbarschaft erkennen.

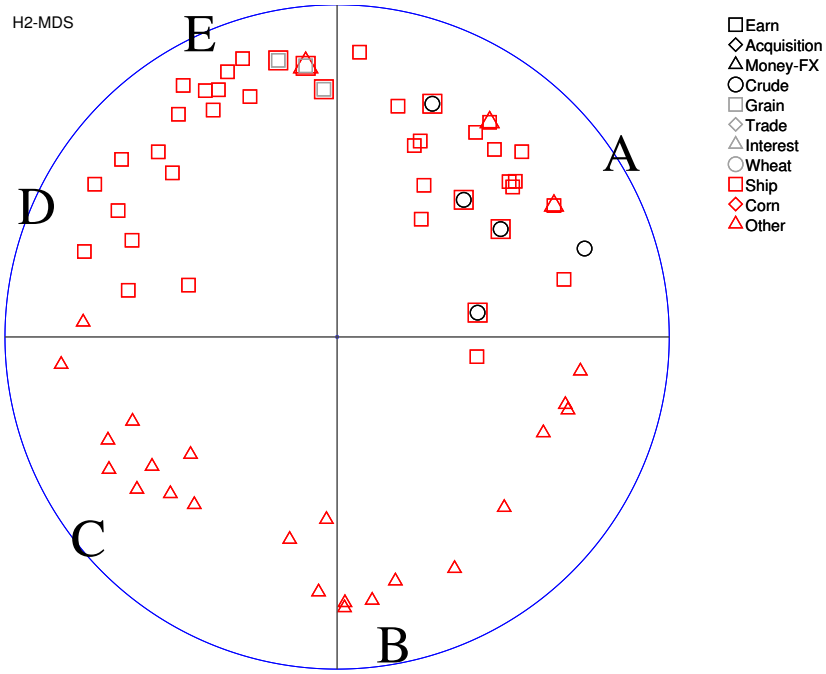


Abbildung 8.13: Die HMDS-Visualisierung aller Dokumente, die dem in Abb. 8.12a mit „?“ identifizierten Knoten zugeordnet sind. Die Legende *rechts* gibt Aufschluss über die Klassifizierung in die zehn wichtigsten Nachrichtenkategorien. Durch konzentrisches Zeichnen (mit wachsender Größe) können hier auch Mehrfachzuordnungen auf Dokumentebene sichtbar gemacht werden. Das Kreuz im Ursprung teilt den $\mathbb{H}^2$ in vier Quadranten und macht in den Folgebildern die Fokusverschiebung deutlich.

Abb. 8.13 zeigt das detaillierte HMDS-Resultat aller Dokumente im augenfälligen „?“-Knoten (Abb. 8.12a). Da die relative räumliche Position der individuellen Dokumente eine wichtige Rolle einnimmt, ist hier auf eine perspektivische Verzerrung verzichtet worden. Die 2D-Glyphen sind so gewählt, dass sie auch konzentrisch gezeichnet werden können, um eine Mehrfachkategorisierung einer Nachricht klar Rechnung zu tragen (s. Legende in Abb. 8.13.

Mehrere Cluster können deutlich unterschieden werden. Die „A“-markierte-Gruppe ist eine Kategoriemixtur, während die anderen „B“–„E“ wesentlich homogener sind.

Aktiviert man die Datenbeschriftung (hier für eine Teilmenge), kann man anhand der prägnanten (aber langen) Nachrichtentitel die semantische Stimmigkeit der Nachbarschaft verifizieren. Abb. 8.14 zeigt den „C“-Cluster zentriert (Abb. 8.13 in 8 Uhr-Richtung). Die Nachrichten drehen sich alle um einen mehrwöchigen Streik, der in der Ölsaatenfabrik Cargill (Seaforth, UK) Anfang 1987 wiederholt Anlass für Nachrichten gab.

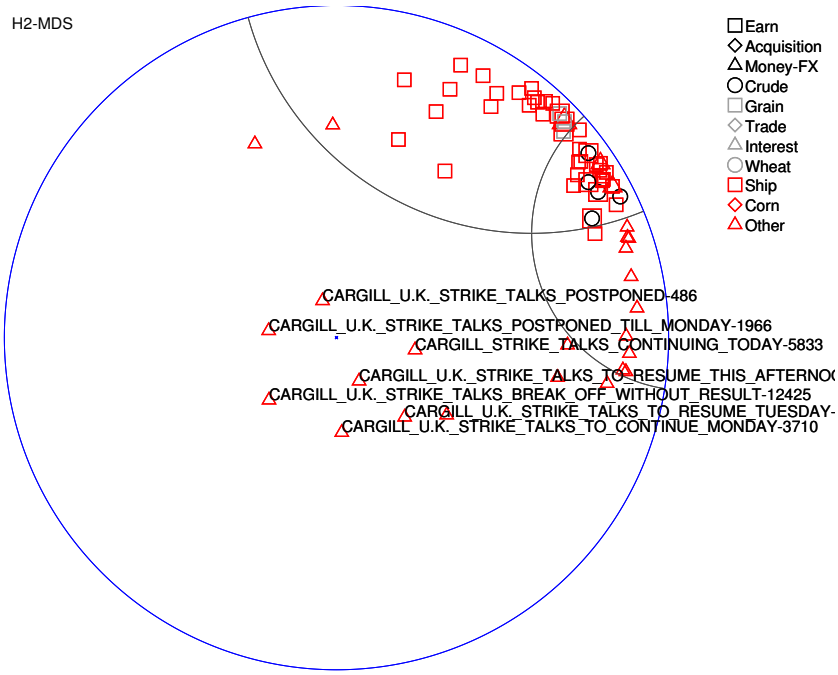


Abbildung 8.14: HMDS-Navigationsschnappschuss mit dem „C“-Dokumentcluster (Abb. 8.13) im Fokus und mit Einblendung von Nachrichtentiteln. Alle befassen sich mit einem Thema: einem Arbeiterstreik in einer britischen Ölsaatenfabrik. Das Orientierungskreuz aus $\mathbb{H}^2$-Geraden erscheint in Abb. 8.13 gerade und nun als Kreisbögen, die senkrecht zueinander und auch zum ∞-Rand stehen.

Als visuelle Orientierung sind in allen drei HMDS-Bildern zwei sich im Ursprung senkrecht kreuzende $\mathbb{H}^2$-Geraden eingezeichnet. Das Fadenkreuz in Abb. 8.13 vierteilt den $\mathbb{H}^2$ symmetrisch. In Abb. 8.14 ist es isometrisch nach rechts oben gewandert und erscheint erwartungsgemäß als gekreuzte Kreisbögen. Die Schnittwinkel sind rechtwinklig erhalten – zueinander und zum ∞-Kreisrand. Alternativ kann auch ein $\mathbb{H}^2$-Gitter (Abb. 7.5)

als Orientierungstütze eingeblendet und isometrisch mitgeführt werden.

Abb. 8.15 rückt den „E“-Nachrichtencluster in den Fokus: Dieser hat auch mit Streiks zu tun, aber im Rotterdamer Hafen im März 1987, wie man an den Labels ablesen kann. Man erkennt die thematische Zusammengehörigkeit sowohl auf Clusterebene als auch auf Knotenebene, denn alle Dokumente dieser Ansicht gehören ja zum selben HSOM-Knoten (d.h. dem mit „?“ markierten in Abb. 8.12a).

Die Beschriftung mit der vollen Nachrichtenüberschrift ist Dank deren Prägnanz sinnvoll, aber auch sehr raumnehmend. In den Bildern wurde daher interaktiv nur eine Auswahl aktiviert. Es wird nochmals deutlich, wie hilfreich es ist, eine interessante Region fokussieren zu können und damit in intuitiver Weise für den aktuellen Interessensschwerpunkt mehr Anzeigeplatz zu allokieren. Das Umfeld wird dabei graduell komprimiert, ohne dass es abrupt ausgeblendet werden muss.

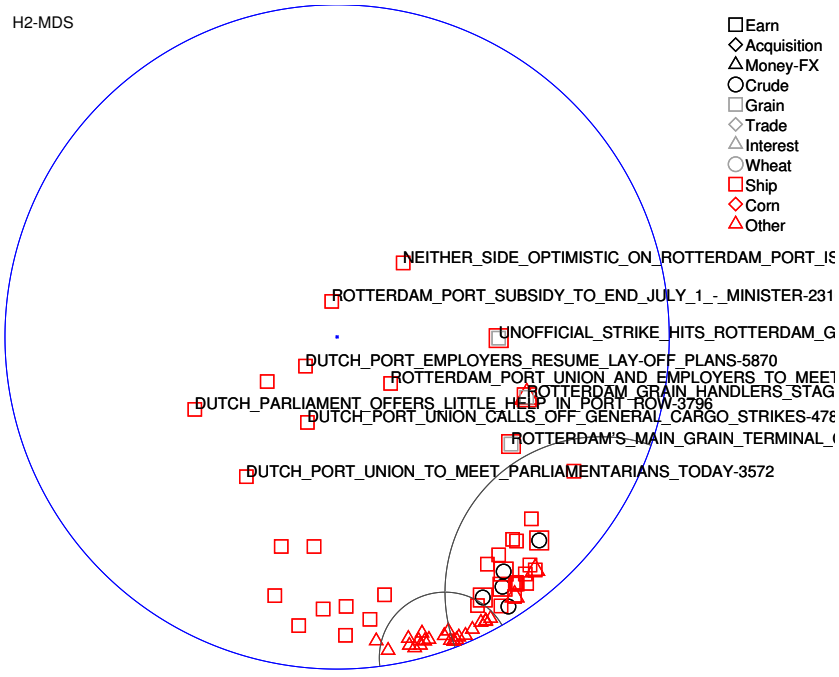


Abbildung 8.15: HMDS-Navigationsschnappschuss – nun mit dem „E“-Dokumentcluster im Fokus. Auch hier zu einem recht spezifischen Thema, d.h. einen Streik im Rotterdamer Hafen. Die Ursprungslage von Abb. 8.13 wird auch hier mit dem $\mathbb{H}^2$-Geradenkreuz markiert.

8.5.2 Ähnlichkeitssuche anhand einer Suchanfrage oder eines Vorlagedokuments

Der hybride Navigationsansatz kann auch eingesetzt werden, um ähnliche Dokumente in einer sehr großen Datenkollektion Φ zu lokalisieren. Das Konzept unterstützt dabei die Vorlage

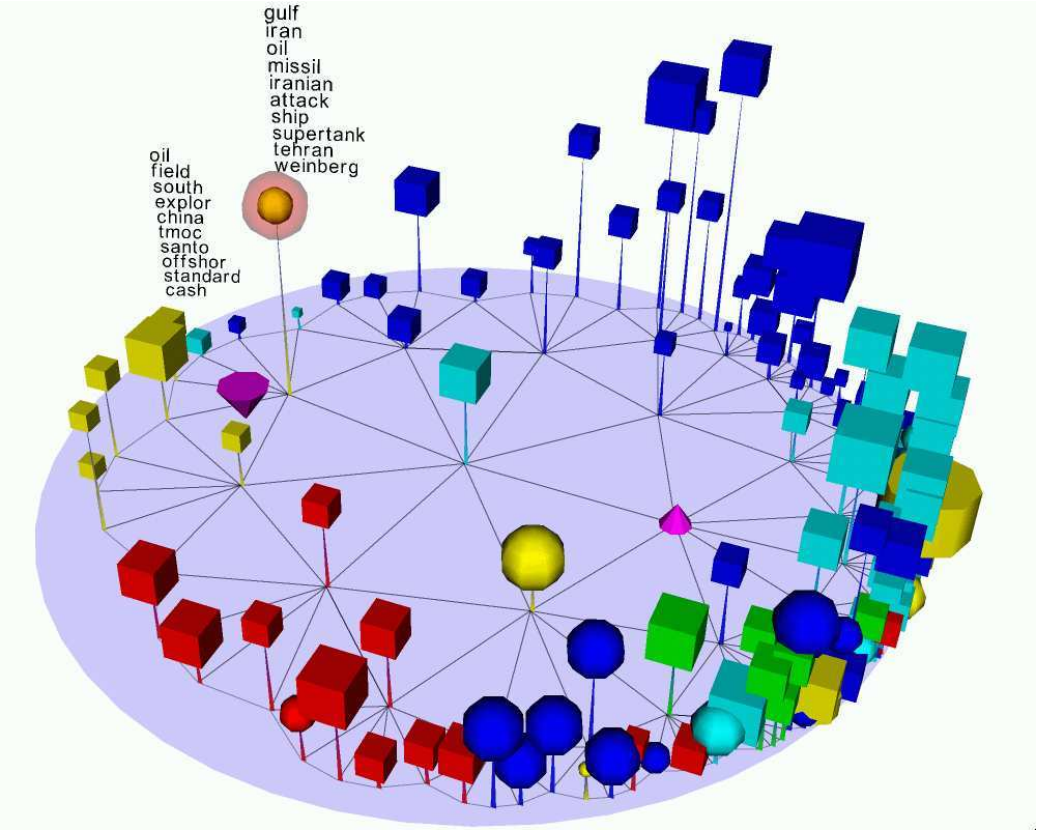


Abbildung 8.16: Ein weiterer Navigationsschnappschuss der HSOM (gleich der u.a. in Abb. 8.12a) in Richtung 10 Uhr und Markierung der für den Suchtext (*„USA leading the strike in a Gulf war against Iraq?“*) zuständigen Knoten a^*. Der *keyword probe* offenbart Einsichten in die thematische Verortung des Knotens und eines seiner Nachbarn.

- einer textuellen Suchanfrage, mittels eines oder mehrerer Stichworte, oder
- eines weiteren Textdokuments.

Der Suchaufwand skaliert hervorragend, da nur Knotenzahl-viele Vergleiche mit den K Prototypen der HSOM durchgeführt werden müssen. Insgesamt werden folgende Schritte ausgeführt:

1. Für eine neue Vorlage wird zuerst der Vektorraummodellmerkmalsvektor $f'_{Vorlage}$ bestimmt;
2. der ähnlichste HSOM-Prototypenvektor f_{a^*} wird gesucht und ggf. angezeigt;
3. Die Dokumentteilmenge Θ für den HMDS-Schritt wird selektiert. Sie besteht aus der Vorlage und den Dokumenten, die dem Best-match-Knoten a^* zugeordnet sind;
4. Die Paardistanzen zwischen den Dokumenten Θ werden berechnet;
5. Eine HMDS Einbettung aller Dokumente Θ wird ermittelt und
6. interaktiv navigierbar im $\mathbb{H}^2$ angezeigt.

Zwei Beispiele sollen dies verdeutlichen.

Das erste Beispiel *(i)* ist der Suchtext *„USA leading the strike in a Gulf war against Iraq?“*. In Abb. 8.16 ist der HSOM-Knoten farblich hervorgehoben, der für diesen Suchtext zuständig ist. Das zweite Beispiel *(ii)* ist auch ein Nachrichtenartikel, diesmal aus der Webedition von CNN vom 27.02.2003 mit dem vielversprechenden Titel *„Bush: Ending Saddam's regime will bring stability to Mideast“*[1]. Beide Texte werden auf denselben HSOM Knoten abgebildet, der in Abb. 8.16 farblich hervorgehoben ist. Der Text hat bemerkenswerterweise auch mit *strike* zu tun. Es geht dabei jedoch nicht um Arbeitskämpfe, wie im vorher inspizierten Knoten (Abb. 8.12b), sondern um Militärschläge.

Dies wird nicht nur bei der Analyse der Schlüsselwörter im HSOM-Knotenüberblick (Abb. 8.16) deutlich, sondern insbesondere in der Detailansicht. Abb. 8.17 zeigt das HMDS-Ergebnis der Dokumententeilmenge Θ zusammen mit den beiden Suchvorlagen. Die Texte liegen zeitlich weit auseinander und werden vom Gesamtsystem semantisch sinnvoll und hilfreich verortet – zunächst grob mittels der HSOM, dann auf Detailebene mittels der HMDS – und auf beiden Ebenen konzepttreu und interaktiv navigierbar dargestellt.

[1] http://www.cnn.com/2003/WORLD/meast/02/27/sprj.irq.bush.speech/index.html, Stand März 2003

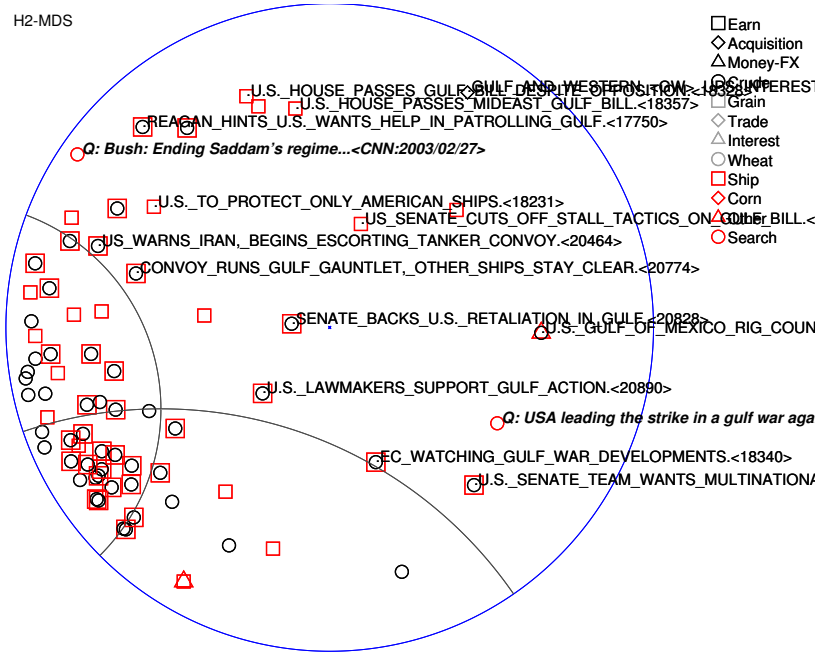


Abbildung 8.17: Ein HMDS-Mapping der Dokumententeilmenge Θ zum Knoten a^*, der für beide Suchvorlagen zuständig wurde. Der Suchtext (*Query: USA leading the strike in a gulf war against iraq?*) ist unter 4 Uhr, das CNN-Suchdokument (*Query: Bush: Ending Saddam's regime ...*) in 10 Uhr-Richtung zu finden. Die Titel der Umgebungsdokumente zeigen die Themenverwandtschaft: Spannungen in der Golfregion (s.a. Abb. 8.18).

8.5.3 Weitere Schritte in der Ergebnispräsentation

Abhängig von der Anwendung kann im nächsten Schritt das Dokument durch Mausselektion dem formatspezifischem Dokument- oder Objektbrowser zugeführt werden.

Ist der Titel noch nicht aussagekräftig, kann der *keyword probe* auch auf Dokumentebene aktiviert werden, um eine Einschätzung der wichtigsten Worte zu erhalten. Hierbei werden interaktiv durch Mausklick einzelne Dokumente auf ihre dominanten Wortmerkmale hin untersucht. Abb. 8.18 visualisiert einige *Keyword-probe*-Ergebnisse, hier jeweils die zehn stärksten Vektorkomponenten des Dokumentenmerkmalsvektors. Damit offeriert dieses Verfahren den Titel flankierende Informationen, bevor auf das Volldokument zugegriffen wird.

Handelt es sich um eine Suchanfrage, so hat sich die *Keyword-in-context*-Darstellung bewährt und wird in vielen Suchmaschinen praktiziert.

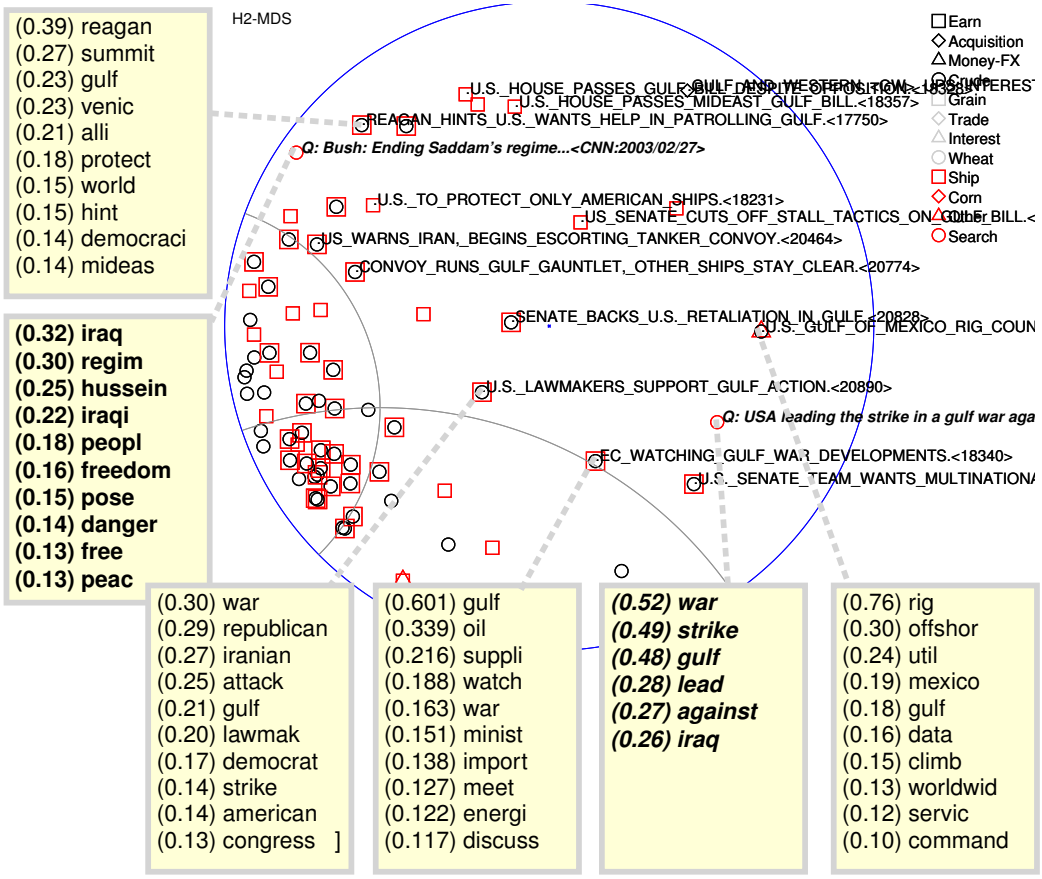


Abbildung 8.18: Integrierte Darstellung der *keyword probes* für einige Dokumente in Abb. 8.17. Die jeweils zehn wertstärksten Komponenten der normierten Dokumentenmerkmalsvektoren $\vec{f}_i'$ (mit Wertangabe in Klammern) sind in den Seitenboxen aufgelistet. Die beiden Suchvorlagen sind hervorgehoben („Q:“). Auf diese Weise kann Einblick in die (gewichteten) Worthäufigkeiten der Einzeldokumente gewonnen werden, und die Themenstruktur der Dokumente können auf einem hohen Abstraktionsniveau überprüft werden.

Die Suchwörter werden in den in Frage kommenden Dokumenten gesucht, typographisch hervorgehoben und gemeinsam mit Umgebungstextausschnitten präsentiert. Die manuelle Inspektion kann unterstützt werden, indem alle Suchworte im Volltext eine geeignete Markierung finden, wie zum Beispiel im *Cache*-Mode bei Google.

Möchte man ganze Textdokumente vergleichen, ist die Hervorhebung aller Worte nicht mehr sinnvoll. Dafür ist eine Analyse der übereinstimmenden Worte und der Gruppierung in Wortfelder möglich. Eine kompakte graphische Annotation in kleine *Tile Bars* schlug Hearst (1995) vor.

8.5.4 Wahl der HSOM-Gittergröße

Die Größenwahl des HSOM-Gitters ist generell ein Kompromiss, bei dem folgende Einflüsse eine Rolle spielen:

- Die Trainingszeit für die HSOM skaliert etwa mit der Anzahl Knoten K des Gitters (siehe auch Abs. 8.6);
- Ein sehr großes Netz erfordert lange Navigationstrecken in der $\mathbb{H}^2$-Interaktion (vgl. Abb. 8.11 und Abb. 8.12a). Die dann sehr zahlreichen, entfernten Randknoten reihen sich nahe dem Einheitkreis – oder werden aus Gründen der Fokussierung und Performanz nicht dargestellt oder substituiert;
- Je größer der Gesamtkorpus, desto größer die mittlere Anzahl selektierter Dokumente $\langle|\Theta|\rangle = |\Phi|/K$;
- In der Praxis bewährt sich eine HMDS Objektanzahl $N = |\Theta|$ im Bereich 40 bis 150.

Glücklicherweise ist die genaue Wahl der HSOM-Größe unkritisch. Durch eine Auswahloptimierung der Vergleichsdokumente, die im Folgenden vorgestellt wird, ist zudem die HMDS-Objektanzahl N gänzlich von der HSOM-Gitterwahl entkoppelbar.

8.5.5 Auswahloptimierung für die Ähnlichkeitssuche

Eine bessere Feinsteuerung der Θ-Selektion lässt sich durch Modifikation des Ablaufschrittes 3 von S. 227 implementieren.

(i) Eine Reduktion der Menge $|\Theta|$ ist durch Auswahl der N Dokumente möglich, die der Vorlage am ähnlichsten sind. Dieser Vorgang schließt $|\Theta|$ Vergleiche und $|\Theta| \log |\Theta|$ Sortierschritte ein.

(ii) Eine Erweiterung der Menge $|\Theta|$ erreicht man durch Hinzunahme der Dokumente, die dem zweit-, dritt-, ..., -ähnlichsten HSOM Prototypenvektor $a^{*2}, a^{*3}, \ldots$ zugeordnet sind (Gl. 5.70).

Kombiniert man die beiden Verfahren (*ii* dann *i*), indem man gezielt den Vergleichsraum über die Bereichsgrenzen des Best-match-Knotens (BMU)

erweitert und anschließend definiert reduziert, kann man das Selektionsverhalten des Gesamtsystems an diesen Voronoi-Grenzen verbessern. Dies ist bedeutsam für Suchvorlagen, die nicht eindeutig zu einer bestimmten Themengruppe gehören und damit zu verschiedenen Dokumentenclustern (und damit den Prototypenvektorn) ähnlich weit entfernt sind.

Zusätzlich wird damit eine Immunisierung gegen Auswirkungen topologischer Defekte in der HSOM erreicht. Faltungen bedeuten Objektnachbarschaft ggf. weit jenseits der Gitternachbarschaft. Durch die distanzabhängige Erweiterung *(ii)* werden die Dokumente aus der Nachbarschaft erreichbar, auch wenn sie zu nichtbenachbarten Knoten gehören. Diese Vorzüge sind besonders bei sehr hochdimensionalen Datensätzen von Interesse, da hier topologische Defekte kaum ausgeschlossen werden können.

Durch den hybriden Ansatz skaliert das Verfahren auch in Bereichen von Millionen von Dokumenten sehr gut, da strukturell die Best-match-Suche $O(K)$ durch eine Rangbildung $O(K \log K)$ ersetzt wird. Der restliche Ablauf skaliert quadratisch mit dem Kontrollparameter N.

8.6 Jumpstarting

Zur Beschleunigung des HMDS-Verfahrens kann die geeignete initiale Vorstrukturierung der Objektlokation $\{\mathbf{x}_i\}$ einen nicht unwesentlichen Zeitgewinn bedeuten. Hierzu eignet sich das 2D-Fastmap-Verfahren, das in Abs. 5.10.4 vorgestellt wurde. Ausgehend von den Disparitäten $\mathbf{D}_{ij}$ findet es eine hilfreiche Anfangskonfiguration $\{\mathbf{x}_i''\} \in \mathbb{R}^2$. Drei Dinge müssen dabei beachtet werden:

Identität: Identische Paare aus $\{\mathbf{x}_i''\}$ müssen verhindert werden. Wenn die Dreiecksungleichung häufig verletzt ist, können mehrere Datenpunkte auf die gleiche Position abgebildet werden. In diesem Fall würde das HMDS-Verfahren in Gl. 7.14 divergieren. Addiert man einen geringen Rauschterm η zu den $\mathbf{x}_i$, ist die Gefahr gebannt.

Zentrierung: Die Daten werden zusätzlich zentriert:

$$\mathbf{x}_i' = \mathbf{x}_i'' - \bar{\mathbf{x}}' + \eta \; . \tag{8.7}$$

Radialskalierung: Das Fastmap-Verfahren liefert Ergebnisse in $\mathbb{R}^2$ – gebraucht wird aber eine Startkonfiguration in der Poincaré-Scheibe im $\mathbb{H}^2$. Die probate Abbildung ist eine radiale Reskalierung

$$|\mathbf{x}_i| = \frac{\exp|\mathbf{x}'_i| - 1}{\exp|\mathbf{x}'_i| + 1} \tag{8.8}$$

in den Einheitskreis. Diese Wahl ist durch den hyperbolischen Radialabstand motiviert und von Gl. 6.24 abgeleitet.

Dieses Starthilfeverfahren ermöglicht es, schneller eine gute HMDS Lösung zu finden.

Kapitel 9

Fallbeispiel: Datamining in der Herzchirurgie

Anhand dieses Fallbeispiels soll das Zusammenspiel von mehreren Komponenten der integrierten Präsentation von komplexen Daten in einem Datamining-Projekt demonstriert werden. Es handelt sich um Arbeiten, die in einem vom Autor initiierten Kooperationsprojekt mit dem Herzzentrum Lahr/Baden entstanden sind. Zunächst werden die Anwendungsdomäne und der Aufbau eines Data-Mart-Systems erläutert, Modelle zur Risikoadjustierung vorgestellt und es wird von Anwendungen berichtet.

9.1 Anwendungsdomäne Herzchirurgie in Lahr

Durch die gestiegene Lebenserwartung und wohlstandsbedingte Lebensführung haben kardiovaskuläre Erkrankungen stark an Bedeutung zugenommen. Dazu gehören Blutgefäßverengung durch harte oder weiche Plaques, die den Herzmuskel insbesondere unter Belastung nicht mehr genügend mit Blut versorgen und aufgrund der damit einhergehenden Schmerzen die Lebensqualität des Betroffenen stark einschränken können. Der so genannte Herz- oder Myokardinfarkt ist eine akute Herzmuskelunterversorgung, die ausgedehnte Muskelareale unter Umständen dauerhaft schädigen kann. Er wird durch lokalen Thrombusverschluss eines Herzkranzgefäßes verursacht. Thromben sind Blutgerinnsel, die durch Blutgerinnungsaktivierung im Bereich aufgebrochener weicher Plaques oder in strömungsarmen Zonen von Blutgefäßen, z.B. in pathologischen Gefäßwandaus-

sackungen, ihren Ursprung haben. Während beim Myokardinfarkt lokal entstandene Thromben den Schaden verursachen, sind es beim Hirninfarkt (Schlaganfall, Stroke oder auch Apoplex genannt) meist losgelöste und angeschwemmte Thromben. Da die Lunge als großer Thrombenfilter wirkt, ist durch die beiden Blutkreisläufe des Herzkreislaufsystems das Hirn vor Thromben aus dem Torus geschützt, nicht aber vor Thromben aus dem Bereich der beiden linken Herzkammern (Vorhof und Ventrikel) und der fünf Versorgungsarterien des Hirns. Dieser Umstand begründet das besondere Operationsrisiko, einen Schlaganfall während oder kurz nach einer Herzoperation (perioperativ) zu erleiden.

9.1.1 Tätigkeitsspektrum

Kardiologen haben sich auf Diagnostik und nicht-chirurgische Behandlung dieser Erkrankungen spezialisiert. Sie führen ferner minimal-invasive Untersuchungen und Eingriffe durch Gefäßkatheder durch, z.B. Kontrastmittelangiographien (Gefäßbeurteilung mit Röntgenbildern), Ballondilatationen (Aufdehnungen von innen) bei Gefäßverengungen und ggf. Einbau von Stents (Metallstützgittern) .

Die Herz-, Thorax- und Gefäßchirurgie befasst sich, wie der Name schon sagt, mit großen Operationen, die die Öffnung des Brustraumes und Operationen am Herzen einschließen. Die besondere Kunst liegt darin, die lebenswichtige Körperversorgung mit sauerstoffreichem Blut auch während langer operativer Eingriffe sicher aufrechtzuerhalten. Typische Operationen umfassen das Aufnähen von Hilfsgefäßen für die Versorgung der Herzkranzgefäße („bypass“), die Rekonstruktion von undichten und defekten Herzklappen (auch Ersatz von Aorten-, Mitral- und/oder Trikuspidalklappe) und die Restaurierung von großen Gefäßwänden.

Das Herzzentrum Lahr in Südbaden ist eine hochspezialisierte Klinik, die jährlich etwa 2.000 solcher großen herzchirurgischen Operationen ausführt (Detailzahlen finden sich im Jahresbericht Walter, Arnrich, Rosendahl und Ennker 2001). Im Vergleich zu einem Allgemeinkrankenhaus ist das Fallspektrum schmal – dafür aber hochfrequent. Der daraus resultierende Datensatz ist für medizinische Verhältnisse sehr umfangreich und damit aus Sicht des Dataminings und der medizinischen Forschung besonders interessant.

Von Seiten der Klinik besteht eine weitere Motivation für das Datamining-

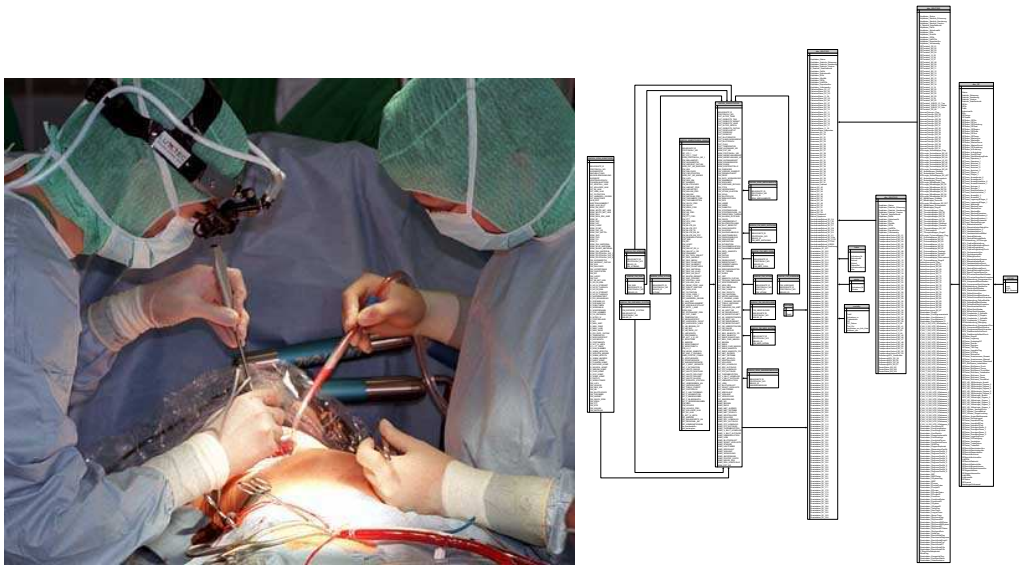


Abbildung 9.1: *(a, links:)* Ein Blick in den Operationssaal während einer Herzoperation. *(b, rechts:)* Die Ansicht eines Informatikers auf die wichtigsten, zusammengeführten Daten über eine OP in einem verkleinertem *entity-relation-diagram*.

Projekt in der Unterstützung und Förderung der Qualitätssicherung (QS) und des Klinikmanagements. Seit Bestehen der Klinik war es von seiten des Klinikdirektoriums (Ennker 1998) stets höchste Priorität, sowohl die Qualitätsstandards als auch die Prozesstransparenz deutlich zu erhöhen, nicht nur innerhalb des eigenen Hauses, sondern gerade auch im interklinischen Vergleich. Dies wird auch daran deutlich, dass Lahr als erstes deutsches Herzzentrum 1999 eine Zertifizierung nach ISO 9001 erlangte.

9.1.2 Risikoadjustierung und EuroSCORE

Ein Vorreiter für Transparenz war das New York *State Department of Health*. Es veröffentlichte 1991 die ersten Klinikvergleichszahlen für Bypassoperationen (*Coronary Artery Bypass Grafting*, CABG). Es waren Leistungszahlen, die die relative Versterbenshäufigkeit (Letalität) bis auf Operateursebene landesweit publik machte. Es stellte sich heraus, dass eine geeignete Berücksichtigung des individuellen Risikos einer Operation nötig ist, um nicht groben Fehlinterpretationen Vorschub zu leisten. Anfänglich führten nämlich die blanken Letalitätszahlen dazu, dass bei den sehr erfahrenen Kapazitäten, die die schwersten Fälle operierten und hohe Mor-

talitätszahlen verzeichneten, die Patienten ausblieben und umgekehrt völlig unerfahrene Operateure mit positiver Erfahrung einfacherer Fälle nun bestürmt wurden, komplizierte Fälle zu behandeln. Heute werden solche Zahlen risikoadjustiert veröffentlicht, um diesem Umstand Rechnung zu tragen (CSRS 2002). Die angewandten Risikomodelle versuchen anhand von präoperativen Merkmalen (solche, die vor der Operation bekannt sind) eine Eintrittswahrscheinlichkeit eines bestimmten Ereignisses abzuschätzen (z.B. Versterben innerhalb einer Zeitspanne). Sie unterscheiden sich hinsichtlich den verwendeten Merkmale und der Patientenkollektive, die in den Studien Eingang gefunden haben, z.B. Wolf et al. (1991) Parsonnet et al. (1996), Heston et al. (1997) und Shaw et al. (2000). Ihnen allen gemein ist die Modellierungsform der logistischen Regression, in die alle Merkmale linear eingehen, s. Gl. 5.50 und Abs. 5.7.2. Dies ist natürlich eine gewisse Simplifizierung der Realität, denn das Risiko jeder einzelnen Operation setzt sich aus vielen Einflussfaktoren zusammen, die durchaus komplexer interagieren. Das Risikomodell ist die Quintessenz ausgedehnter statistischer Modellierungversuche auf der Grundlage von verfügbaren, standardisierten Messungen und Beurteilungen.

Eine auf Europa zugeschnittene, umfangreiche Studie entwickelte ein Risikomodell, das ***Euro****pean* ***S****ystem for* ***C****ardiac* ***O****perative* ***R****isk* ***E****valuation* (EuroSCORE), das auf insgesamt 19.030 Operationsdaten aus 132 Herzzentren in acht europäischen Ländern fußt (Nashef et al. 1999; Roques et al. 1999). Es berücksichtigt 18 Einflussfaktoren, darunter allgemeine Patientenmerkmale (Altersgruppe, Geschlecht, Vorerkrankungen), herzbezogene Daten (Reduktion der Auswurfrate, Bluthochdruck) und operationsbezogene Parameter (u.a. Notfallumstände, Aortenoperationen, s. Abb. 5.7.4). In früheren Kapiteln wurden EuroSCORE-verbundene Daten mehrfach exemplarisch präsentiert, siehe Seite 47, 126, 130, 132 und 139.

Zwei Versionen mit den gleichen Merkmalen, aber unterschiedlichen Faktoren wurden publiziert:

Der *Simple-additive*-EuroSCORE mit ganzzahligen Multiplikationszahlen ist in manueller Bearbeitung noch übersichtlich zu handhaben und ergibt eine ganze Zahl zwischen 0 und $\approx$22;

Der *Logistic*-EuroSCORE mit fixen Faktoren, die direkt aus einer logistischen Regression bzgl. der Letalität (*outcome*) stammen. Dank der vollständigen Parametrisierung ist eine absolute Vergleichsletalitätsrate berechenbar. Sie ist als Online-Rechner unter www.euroscore.org leicht erreichbar.

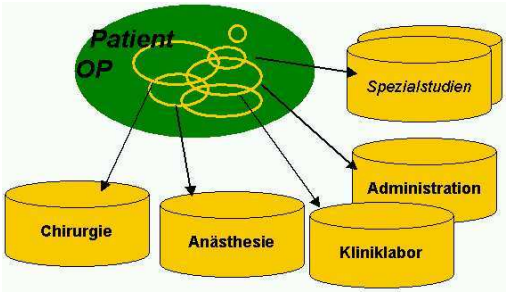


Abbildung 9.2: Ein häufiges Bild: Die Datenhaltung ist geprägt von autonomen Fachabteilungen. Hier unterhalten die Chirurgie, die Anästhesiologie, die Verwaltung und die klinische Chemie jeweils ihr eigenes Klinkinformationssystem (KIS) und speichern darin ihre jeweilige Sicht auf einen Patienten und dessen OP.

Die Übertragbarkeit des EuroSCORE auf andere Patientenkollektive wurde von Nashef et al. (2002) für die USA belegt.

9.2 Probleme und Herausforderungen

Das Herzzentrum Lahr betreibt seit seinem Bestehen 1995 mehrere Informationssysteme, s. Abb. 9.2. Sie unterstützen die Ärzte, das medizinische Personal, das Labor und die Verwaltung in ihrer Arbeit. Leider sind diese Systeme hochspezialisiert und nicht auf Interaktion oder Integration hin konzipiert. So war es in der Vergangenheit relativ mühevoll und zeitaufwändig, Informationen aus den Hauptsystemen und ihren proprietären Datenstrukturen zu extrahieren und zusammenzuführen. Da Zeit auf Seiten des medizinischen Personals eine rare Ressource ist, war dies ein ausgesprochenes Forschungshemmnis. Zudem entstanden häufiger auch Fehler: Ein harmloses Beispiel ist die Fehlausweisung von nur 3 statt der tatsächlich 25 LV-Aneurysmektomien, einer besonders anspruchsvolle Operation im Jahresbericht 1999.

Die Probleme und Herausforderungen sind nicht untypisch für eine Klinik, wie auch in Kuhn und Giuse (2001), Lenz und Kuhn (2001) und Haux (2002) diskutiert wird:

Koexistenz unverbundener Klinik-Informations-Systeme (KIS) unter der Leitung autonomer Fachabteilungen. Sie werden von vielen als schutzwürdige Investitionen betrachtet. Modifikationen sind sehr aufwän-

dig, da wartungsvertragliche und haftungsrechtliche Hemmnisse bestehen;

heterogene Datenquellen: Neben den KIS gibt es weitere, z.T. handaufbereitete, wertvolle Datenbestände aus Klinikstudien;

partielle Konsistenz: Aufgrund von Übertragungsfehlern u.ä. stimmen redundante Einträge nicht überein, widersprechen sich oder lassen sich nicht zuordnen;

komplizierter Zugang zu Daten und

langwierige Auswertungsprozeduren bei übergreifenden Fragestellungen.

Da eine „große“ vollintegrierte KIS-Lösung aus organisatorischen, ressourcentechnischen und prinzipiellen Schwierigkeiten keine Realisierungschancen hatte, wurde eine schlanke, ergebnisorientierte Lösung der Probleme entwickelt.

9.3 Aufbau eines Data-Marts

Ein *Data-Mart* ist ein kleines *Data-Warehouse*, das projektspezifisch die benötigten Daten führt und bewusst darauf verzichtet, entsprechend eines Warenlagers sämtliche Informationen bereitzuhalten (Fernandez und Schneider 1996; Inmon 1998; Imhoff 1999). Stattdessen wird der Datenumfang zielorientiert definiert und ggf. evolviert, um in kurzer Zeit zu praktischen Lösungen zu gelangen.

In der ersten Phase wurde in engagierter Zusammenarbeit mit dem Domänenexperten vor Ort, Dr. Alexander Albert, ein *Data-Mart*-System geplant. Hierzu gehört die Modellierung der Anforderungen, der Datenstruktur, der Quell- und Zieldaten und ihrer Dimensionen sowie eine erste Qualitätsbeurteilung. Der grobe Datenfluss ist in Abb. 9.3 illustriert. Um Störungsrisiken für den operativen Klinikbetrieb zu minimieren und KIS-Umbauarbeiten zu vermeiden, wurden Spiegelungsprozeduren für alle relevanten Quelldaten konzipiert. Diese sind unumgänglich, da die folgende Verarbeitung auf einem aus Sicherheitgründen isolierten Unix-Server stattgefunden hat.

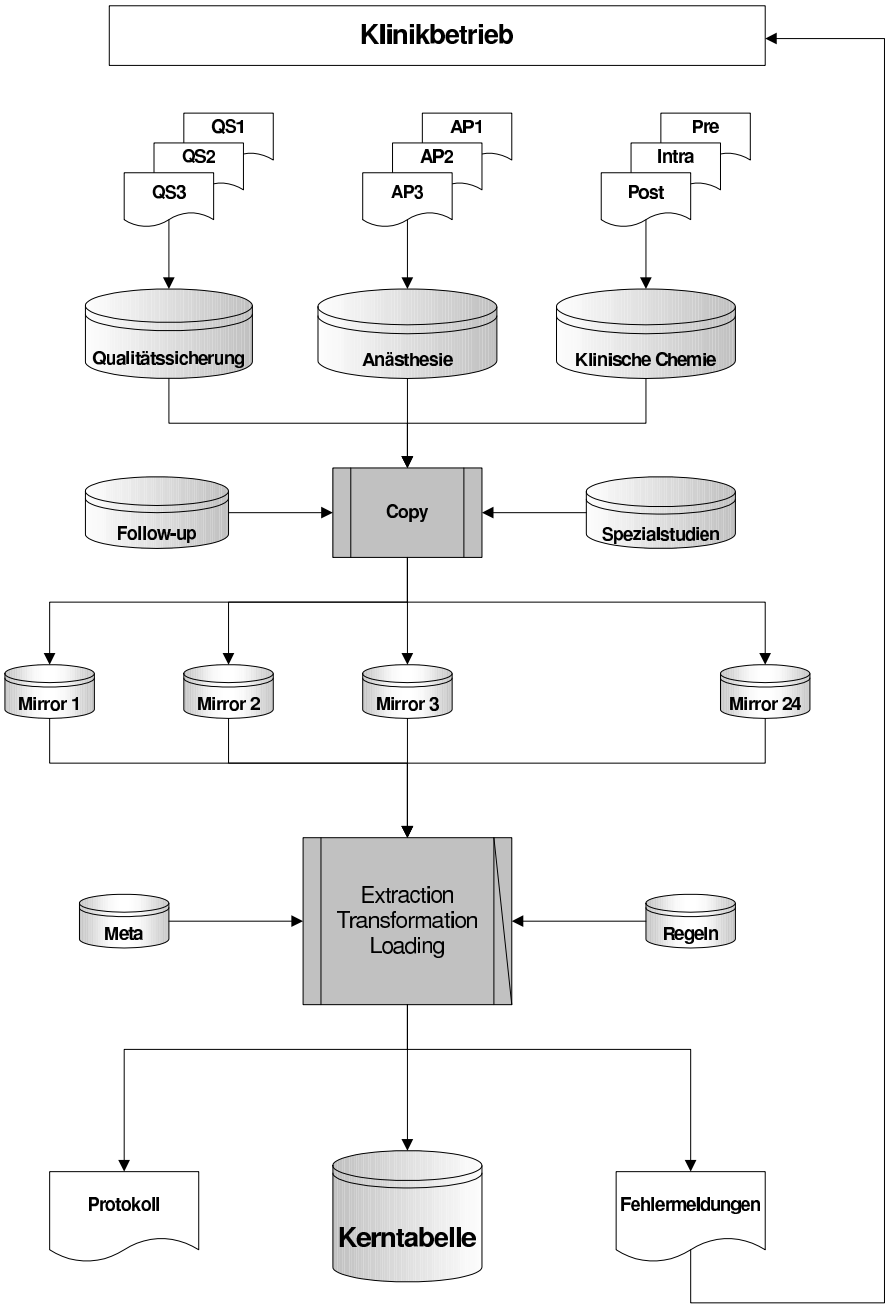


Abbildung 9.3: ETL-Struktur der realisierten Data-Mart-Lösung.

Nach diesem „***E**xtraction*“-Schritt aller relevanten Daten von diversen Quellen werden diese transformiert („***T**ransformation*“) und die Zieltabellen geladen („***L**oading*“). So einfach diese Schritte klingen mögen, in der Praxis ist mit Design, Implementierung und Tests ein hochkomplexer Arbeitsaufwand verbunden. Dass etwa 50-80 % des Gesamtaufwandes eines Datamining-Projektes in diesen Vorbereitungsschritten stecken, ist vielfach beschrieben worden und konnte in diesem Projekt bestätigt werden (z.B. TwoCrows 1999; Pyle 1999) .

Eine Besonderheit dieses Projektes liegt in der sorgfältigen Integration von historischen Daten, die den hohen medizinwissenschaftlichen Wert des Datenbestandes begründen. Dies hatte zur Konsequenz, dass sämtliche Versionswechsel der drei Hauptinformationssysteme nachgebildet werden mussten. Die ohnehin schon eigentümlich komplexe Datenbankstrukturen wurden von den Herstellerfirmen wiederholt modifiziert. Leider muss man hier mit einem weitreichenden Mangel an Dokumentation und nachvollziehbarer Prozessbeschreibung umgehen.

Sind die Daten von der Chirurgie, der Anästhesiologie und der klinischen Chemie zusammengeführt worden, muss das **Problem der partiellen Konsistenz** gelöst werden, zum Beispiel

- OP-Identifikationsnummern passen nicht zusammen – verursacht z.B. durch Lesefehler eines Formularscanners. Indem Editdistanzen auf Name und Geburtsdaten gebildet werden, kann man Bindungshypothesen generieren (s. Abs. 2.4.1). Mit diesem Verfahren konnten über 1000 „Altfälle“ restauriert werden.

- Einträge fehlen, sind unplausibel oder widersprechen sich.

Hier wurden die einzelnen Quellenzuverlässigkeiten beurteilt und es wurde ein umfangreicher Regelsatz entwickelt. Er wird auf jeden neuen Fall angewendet und damit werden die Daten konsolidiert. Es wurden spezielle webbasierte Inspektionswerkzeuge entwickelt, die den Zusammenbau des Data-Marts integriert nachvollziehbar machen. Sie können auf einzelne Fälle oder definierbare Fallgruppen angewendet werden und sind damit von großem Wert, um die Regelbasis gezielt zu testen und zu warten. Wichtige Komponenten sind in der Diplomarbeit von Bert Arnrich (2001) entstanden und dort detaillierter beschrieben.

Kerninhalte und EuroSCORE-Integration

Die Data-Mart-Kerntabelle umfasst nun mehr als 250 Merkmale von jeweils über 13.000 Operationen und mehr als 5 Millionen Laboreinzelmessungen von Blutproben. Die Merkmale umfassen prä-, peri- und postoperative Parameter, wie sie im Klinikworkflow formulargestützt oder papierlos über jede individuelle Operation aufgezeichnet werden. Kurz nachdem die Daten elektronisch erfaßt sind, werden sie automatisch in den Data-Mart transferiert und verarbeitet. Über Webmasken können essentielle Informationen nachgetragen werden, z.B. Versterben nach Verlassen der Klinik. Dies ist für die korrekte Handhabung der „30-Tage"-Standardletalität wichtig, sofern das Ableben innerhalb dieser 30 Tage nach Operation (post-op) eintritt.

Eine wichtige Besonderheit ist die Integration von Risikomodellen, die auf dem EuroSCORE basieren. Neben der Anpassung und Präparation aller Einflussmerkmale werden einige Risikomodelle standardmäßig angewendet und sind fallbasiert verfügbar:

ES-num: ganzer Zahlwert des *simple-additive* EuroSCORE, wie in Abs. 9.1.2 erwähnt;

ES-log: Vergleichsletalitätswert entspricht dem Durchschnitt der 132 an der EuroSCORE-Studie beteiligten Kliniken, wie in Abs. 9.1.2 erwähnt;

ES-LahrCal: Klinikspezifische Letalitätsschätzung das auf dem ES-num Wert beruht;

ES-LahrLog: wie ES-log, aber auf das klinikspezifische Patentienkollektiv kalibriert.

Die mittlere Letalität aller Fälle beträgt etwa 2,0 %, wobei der Anteil von $\frac{1}{4}$ dringlicher und Notfall-Operationen den Schnitt von 1,5 % für die elektiven (geplanten) Operationen verschlechtert. Damit liegt das Herzzentrum deutlich unter den EuroSCORE-Zahlen (3,9 %) und den US-Zahlen (1,2 % vs. 2,2 % für Bypass-Operationen (CABG). Eine genauere Merkmalsanalyse findet sich z.B. in Walter et al. 2003).

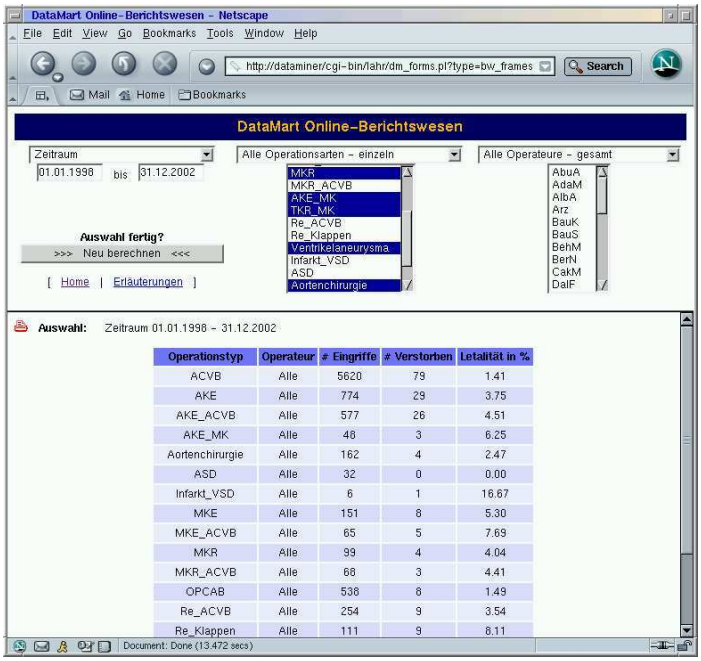
9.4 Integrativer Zugang für Auswertungen im Intranet-Portal

Um die gesammelten Daten einfach und effektiv zugänglich zu machen wurde ein Auswertungs-Portal zum Data-Mart geschaffen. Es steht allen registrierten MitarbeiterInnen im Kinikintranet zur Verfügung. Das Auswertungsportal wird u.a. genutzt für:

- die Auswahl von Datensätzen für weitere medizinische Analysen. Die Selektion erfolgt über Tabellen oder alternativ mittels hyperbolischem Treebrowser, s. Abb. 7.2, S. 180;
- der Export der Datensätze erfolgt aus Datenschutzgründen voll-anonymisiert. Dabei wird die Fall-ID mit einer geheimen one-way Hashfunktion verschlüsselt, die selbst Klinikpersonal keinerlei Rückschluss auf den Patienten erlaubt.
- Das Online-Berichtswesen über Operationszahlen in diversen Kategorien und frei wählbaren Zeiträumen (s. Abb. 9.4);
- risikoadjustierte temporale Performanzanalysen, sogennante VLADs, s. Abs. 9.4 unten;
- risiko-adjustierte Hypothesentests und Signifikanzberechnungen für gruppenspezifische Abweichungen, s. Abs. 9.5 unten.

Dabei können die Auswertungen in beliebig wählbaren Untergruppen durchgeführt werden. Besonders interessante Gruppen-Differenzierungen sind direkt in der Eingabemaske kombinierbar, i.e. Operationsart, Operateur und Zeiträume.

Personenbezogene, nicht-aggregierte Auswahlmöglichkeiten unterliegen Zugangsbeschränkungen. Zum Beispiel darf ein Operateur eine von ihm selbst durchgeführte Operation im Detail auswerten, nicht aber die Operationen eines Kollegen. Personenbezogene Patientendaten werden im Data-Mart grundsätzlich nicht verarbeitet. Die „Falldaten“ werden aber unter Pseudonym gespeichert, um eine Rückverfolgung mit den anderen KIS, etwa zu Prüfzwecken, zu ermöglichen. Im Datenexport erfolgt eine vollständige Anonymisierung.



Operationstyp	Operateur	# Eingriffe	# Verstorben	Letalität in %
ACVB	Alle	5620	79	1.41
AKE	Alle	774	29	3.75
AKE_ACVB	Alle	577	26	4.51
AKE_MK	Alle	48	3	6.25
Aortenchirurgie	Alle	162	4	2.47
ASD	Alle	32	0	0.00
Infarkt_VSD	Alle	6	1	16.67
MKE	Alle	151	8	5.30
MKE_ACVB	Alle	65	5	7.69
MKR	Alle	99	4	4.04
MKR_ACVB	Alle	68	3	4.41
OPCAB	Alle	536	8	1.49
Re_ACVB	Alle	254	9	3.54
Re_Klappen	Alle	111	9	8.11

Abbildung 9.4: Bildschirmansicht des Online-Berichtswesens. Oben Auswahlmaske, unten Resultat.

Abb. 9.4 zeigt das Online-Berichtwesen und Abb. 9.5 den Blick auf ein Auswahlmenü für die VLADs. Neben den numerischen Fakten in Zahlenkolonnen lassen sich die OLAP-Funktionalitäten des Excel-Programmes nutzen, um schnell und effizient Standardgraphiken in Untergruppen zu analysieren.

Temporale Performanz – VLADs Auswertungen

Eine temporale Performanzdarstellung ist auch unter dem Namen *variable life-adjusted display* (VLAD) bekannt (Lovegrove et al. 1997; Poloniecki et al. 1998). Es handelt sich um eine kumulative risikoadjustierte Letalitätsdarstellung, um Schwankungen in der Versterbensrate zu visualisieren.

Ausgangsbasis für die Berechnung sind die risikoadjustierten erwarteten Letalitätszahlen (*expected mortality*) $l_i^{expected}$ jeder Operation i. Sie schwanken in Lahr je nach konkreten Merkmalsausprägungen zwischen etwa 0,002 und 0,1. Diese Schätzung wird akkumuliert – abzüglich der tatsächlich beobachteten Letalität (*observed mortality*), die 1 ist, wenn der

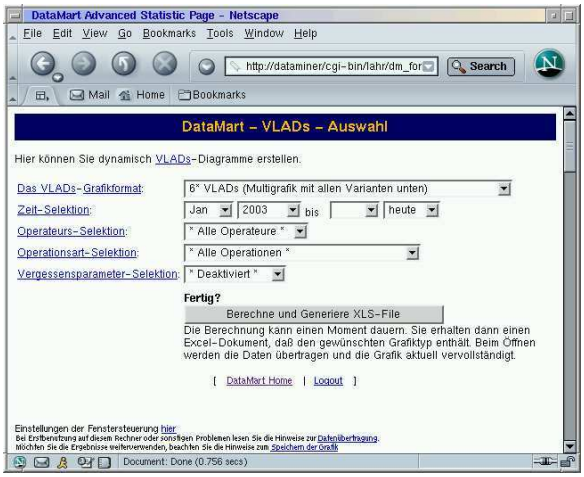


Abbildung 9.5: Bildschirmansicht einer Auswahlmaske für VLADs im Auswertungsportal des Data-Marts.

Patient innerhalb von 30 Tagen post-op verstirbt, sonst 0. Die Performanz

$$NLS(i) = \sum_{j=1}^{i} (l_i^{expected} - l_i^{observed}) \qquad \text{mit} \quad l_i^{observed} \in \{0, 1\} \qquad (9.1)$$

wird auch *net life saved* genannt, da sie die Nettozahl quasi geretteter Leben ausdrückt. Das Bezugsniveau liegt bei der gewählten Referenzbewertung, siehe Abs. 9.3.

Im ersten Beispiel werden alle Operationen ausgewählt, die ein bestimmter Operateur im Zeitraum 08.2000–02.2003 leitend ausgeführt hat (Albert, Walter, Rosendahl, Arnrich, Beller, und Ennker 2003). In Abb. 9.6 sind die OPs chronologisch auf der Abszisse aufgetragen und (in Abständen) mit dem Datum markiert. In allen geraden Kalendermonaten dienen zusätzlich Achsenmarkierungen als Indikator für die Zeitskalierung. Die sägezahnförmige Kurve steigt flach für einfache und stärker für besonders riskante, aber geglückte Operationen. Anfänglich versterben drei Patienten, was an den „Einbrüchen“ erkennbar ist. Langfristig lassen sich durchschnittlich gute Ergebnisse darin erkennen, dass die NLS-Werte um die Horizontalachse schwanken. Ist die Operationsperformanz überdurchschnittlich gut, steigt die Sägezahnkurve und umgekehrt. Gibt es eine zeitliche Häufung der Ereignisse, ist dies ein Indiz für ein Problem, das der Aufmerksamkeit der ärztlichen Leitung bedarf. Umgekehrt sind längere steigende Kurven ein positives Zeichen und eine Motivation für auch künftig anhaltenden Einsatz. Am Ende des Zeitraums wurden hier 1,7 „Leben

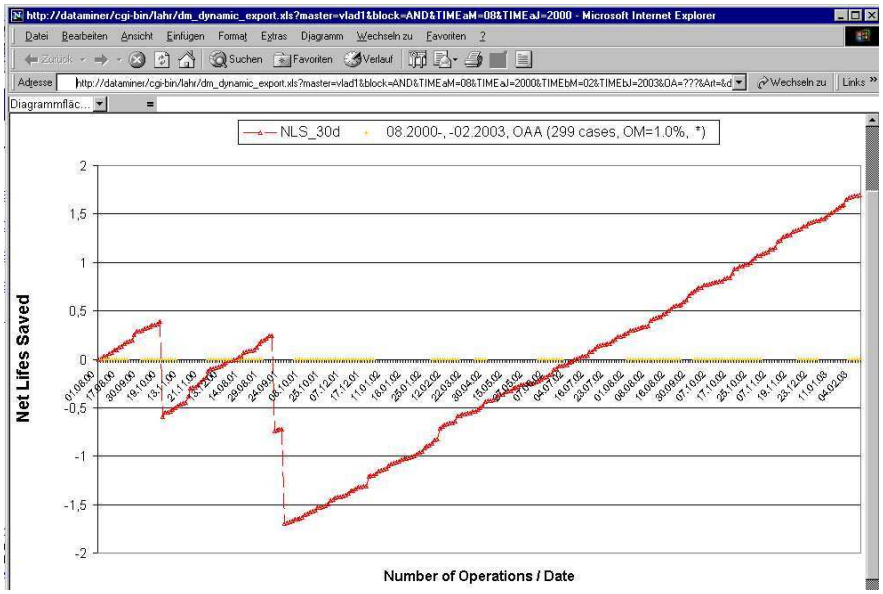


Abbildung 9.6: Operationsperformanz als VLAD-Kurve für einen Operateur. Aufgetragen sind die Anzahl der „netto geretteten Leben“ über die Operationszahl für einen einzelnen Operateur.

netto gerettet“ .

Steht einem ein solches Instrument zur Performanzanalyse zur Verfügung, lassen sich beispielsweise auch Fragen zum Spektrum angewandter Therapietechniken stellen. Das nächste Beispiel beleuchtet die Einführung der minimal-invasiven Bypassoperationen, dem so genannten *Off-Pump-Coronary-Artery-Bypass*-Verfahren (OPCAB). Diese Operationstechnik verlangt eine besonders geübte Operateurshand, denn anstatt mit Hilfe einer Herz-Lungen-Maschine den Herzmuskel stilllegen zu können, werden die Nähte an den Herzkranzgefäßen ausgeführt, während das Herz schlägt. OPCAB Daten dienen auch in Abs. 5.9 als Beispiel für visuelles Datamining (s. a. Abb. 5.20 für die Detektion einer ternären Merkmalsbeziehung).

Abb. 9.7 zeigt die VLADs für OPCABs seit Klinikbestehen. Bei den frühen Operationen traten mehrfach Komplikationen auf, aus deren Aufarbeitung Erfahrungen und Erkenntnisse gesammelt wurden. Somit gelangt man zu einer Verbesserung der Operationstechniken: Die Kurve steigt kontinuierlich – es stellt sich nach einiger Zeit langsam Routine ein. Dann, nach einer erfolgreichen (über 2 Jahre dauernden) Phase, schleichen sich mutmaßlich Nachlässigkeiten ein, es wird seltener operiert (an den Abzissenmarkern und Datumsangaben sichtbar) und Todesfälle häufen sich

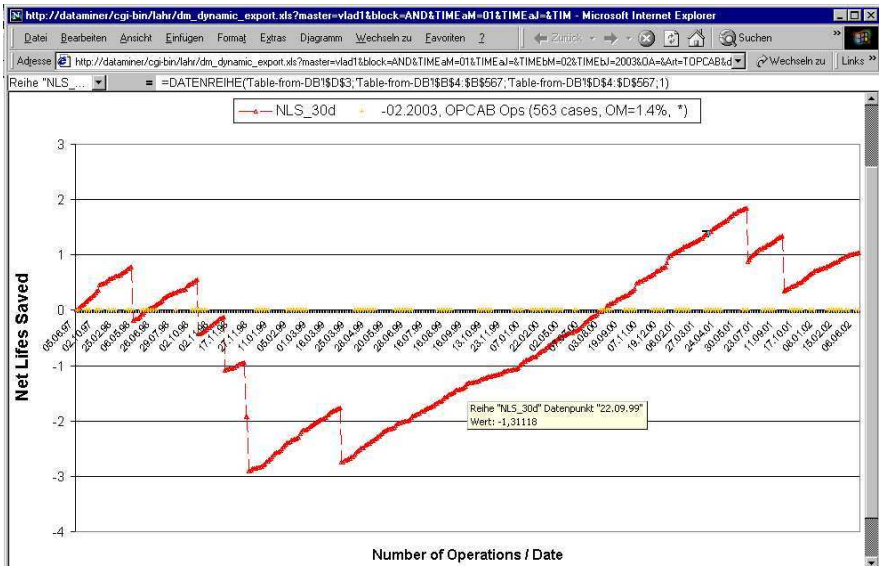


Abbildung 9.7: Operationsperformanz als VLAD-Kurve für einen Operationstyp: hier das minimal invasive Bypassoperationsverfahren *Off Pump Coronary Artery Bypass* (OPCAB). Technisch ist hier eine dynamisch generierte Excel-Datei mit Datenbankkopplung im Browser-Plugin zu sehen. Durch solche Integrationsformen lassen sich hochentwickelte Benutzungsmöglichkeiten von Standardsoftwarekomponenten nutzen: Hier z.B. erlaubt die sichtbare Tooltip-Funktion die detaillierte Dateninspektion via Maus (s. Datum).

wieder. Diese Form der „Lernkurve“ ist typisch und wurde schon mehrfach in der Literatur beschrieben. Mit der Auswertungsform der VLADs lässt sich die Operationsperformanz zeitlich analysieren und Fehlentwicklungen können zeitnah aufgespüren werden, außerdem gibt sie den Verantwortlichen die Möglichkeit, frühzeitig Gegenmaßnahmen zu treffen.

9.5 Risikoadjustierte Hypothesentests und Konfidenzintervalle

Eine zweite wichtige Größe in der Performanzbeurteilung für eine Operationsauswahl Φ ist neben der NLS (Gl. 9.1) der *Risk Adjusted Mortality Quotient* (RAMQ)

$$RAMQ(\Phi) = \frac{OM}{EM} = \frac{\sum_{i\in\Phi} l_i^{observed}}{\sum_{i\in\Phi} l_i^{expected}}, \tag{9.2}$$

der die beobachtete und die erwartete Letalität ins Verhältnis setzt. Ein Wert von 1 ist normal, >1 beziffert zu viele Verstorbene und umgekehrt, weist <1 auf eine überdurchschnittliche Operationsperformanz hin.

Ist die betrachtete Anzahl von Fällen $|\Phi|$ klein, sind die statistischen Fluktuationen relativ groß und das Problem der richtigen Ergebnisinterpretation wird besonders deutlich. Zum Beispiel ist bei einer Fallzahl $|\Phi| = 50$ und einer Eintrittswahrscheinlichkeit von $p = 2\,\%$ ein Verstorbener zu erwarten. Findet man null, zwei oder drei Ereignisse, dann sind die Schwankungen von NLS und $RAMQ$ sehr groß und es stellt sich die Frage: Wann ist eine Abweichung signifikant?

Die übliche Methode ist die Bildung von Kontingenztabellen und anschließendem χ^2-Test (s. Abs. 4.12.1). Um nicht zu kleine Gruppengrößen zu bekommen, bietet sich die Aggregation von EuroSCORE-Gruppen an, z.B. in drei ungefähr gleichstarke Gruppen mit niedrigem, mittlerem und hohem Risiko (Nashef et al. 1999 schlugen die EuroSCORE-Intervalle 0-2, 3-5, 6++ vor).

Ist die Grundgesamtheit Φ klein, verwirft man allzu viel Information. Fisher's exakter Test liefert dann zwar noch einen p-Wert, aber keine Konfidenzintervalle.

Zur Lösung wurde hier eine Methode entwickelt, integrierte Hypothesentests durchzuführen (Walter et al. 2003). Die Kernidee ist die näherungsfreie Modellierung des genauen Risikomixes, wobei jede Einzeloperation als unabhängige Bernoulli-Zufallsvariable gesehen wird. In beliebigen Gruppen wird mit der Methode der Monte-Carlo-Simulation die Verteilung der erwarteten Gesamtereigniszahl ermittelt.

Im Intranet-Portal des Data-Marts werden hierzu eine Grundgesamtheit und das Risikomodell gewählt. Im zweiten Schritt werden zu vergleichende Untergruppen definiert. In Pseudocode formuliert, laufen folgende Schritte ab:

```
foreach group {
    for(j=0; j<#topLoop; j++) {
        for(i=0; i<201; i++) {
            foreach case i in group {
                wirf Münze mit individueller Letalität l_i^expected;
                akkumuliere gefundene Letalität;
            }
        }
        berechne Verteilungsfunktion;
        akkumuliere Verteilungsfunktion;
    }
    verorte die OM, ermittle Quantile in Verteilungsfunktion;
    speichere relevante Verteilungsfunktionsparameter;
}
```

Hierzu sei bemerkt, dass die zweitinnerste Schleife bis genau 201 läuft, um bequem u.a. die beiden CI-95 %-Konfidenzintervallgrenzen extrahieren zu können (Abs. 4.4 f). Sie liegen dann genau an Position 5 und 195 (im Indexbereich {0, . . . , 200}) der sortierten Ereignissummen. Diese Schleife wird #topLoop (etwa 5–50-mal) mal wiederholt, um die Restriktion auf ganzzahlige Ergebnisse zu lösen und eine Präzisierung der mittleren Verteilungsfunktion zu erhalten. Bei kleinen Gruppengrößen sollte jeder Fall mindestens 10.000-mal „gewürfelt“ werden. Anhand der Verteilungsfunktion wird der p-Wert ermittelt, wobei auf korrekte Behandlung von Bindungen in der Verteilungsfunktion zu achten ist.

Abb. 9.8 zeigt das Ergebnis eines konkreten Beispiels. Die interaktiv generierten Ergebnisse werden im Auswertungsportal präsentiert (gleichzeitig werden sie auch persistent in einer Datenbank gespeichert). Die gewählte Grundgesamtheit umfasst alle Patienten, deren *Creatinin-Clearance*-Blutwert (CC) präoperativ bestimmt wurde. Dies ist ein hochgerechneter Nierenleistungsparameter für die Kreatininausscheidung und berücksichtigt geschlechtsspezifisch das Körpervolumen nach Cockcroft-Gault (für weitere Details s. Walter et al. 2003). Zwar ist ein dichotomer Nierenwert, der Serumkreatinin-Wert, im EuroSCORE vertreten, doch, so die Hypothese, ist dies eine zu grobe Berücksichtigung des Risikofaktors Nierenschwäche.

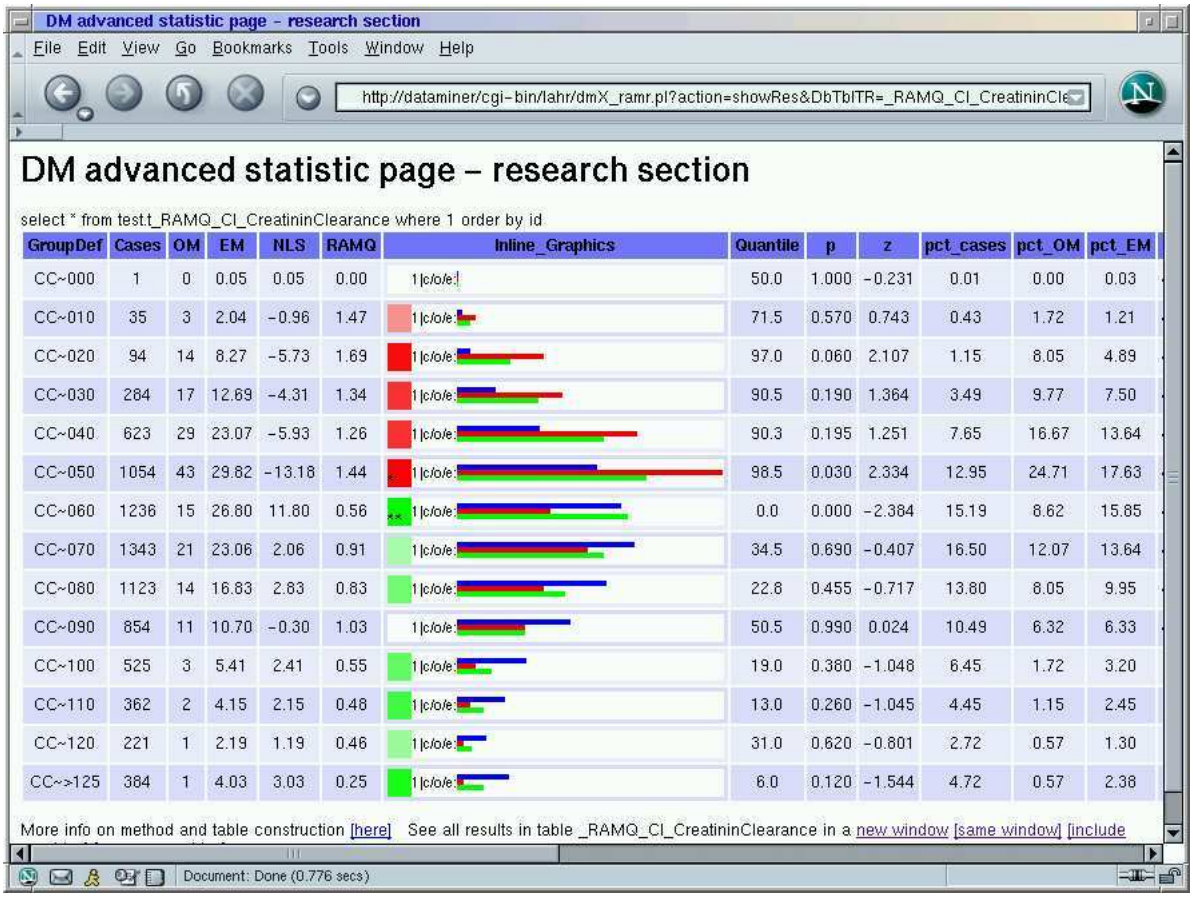


Abbildung 9.8: Beispiel für Risiko-adjustierte Hypothesentests hier integriert im Auswertungsportal des Data-Marts. Über eine Webmaske werden die Grundgesamtheit (hier alle Patienten, für die die *Creatinin-Clearance* (CC) präoperativ bekannt ist) und eine Untergruppenaufteilung ausgewählt (hier CC, gerundet auf 10 ml/min). Der jeweils oberste, blaue Balken der Inline-Graphiken ist proportional zur relativen Fallzahl (*cases, pct_cases*) der Untergruppe (*GroupDef*). Der zweite, rote Balken ist proportional zur Gesamtzahl der beobachteten (*OM=Observed Mortality, pct_OM*), der untere, grüne Balken ist proportional zur Gesamtzahl der beobachteten Letalitäten (*EM=Expected Mortality, pct_EM*). Sichtbar sind ferner die p-Werte und die Quantilenniveaus der *OM*. Interessante Untergruppen werden auf einen Blick durch das linke Farbfeld in den Inline-Graphiken erkennbar, angezeigt auf einer rot-grün Skala, die unter- bis überdurchschnittliche Performanz signalisiert. Je kleiner der p-Wert, desto gestättigter die Farbe. Sinkt er unter die konventionelle 5 %-Marke, wird das Feld mit einem * gekennzeichnet und bekommt einen weiteren Stern unter der 1 %-Marke.

Teilt man die Patienten nun nach ihren gerundeten CC-Werten in Untergruppen, so erkennt man an der gedachten Linie der blauen Balkenenden die Verteilung der Merkmalsausprägungen, s. a. Legende von Abb. 9.8. Unter der Nullhypothese der Unabhängigkeit der Operationen vom gewählten Merkmal CC sollten die Balkentripel untereinander etwa gleiche Länge aufweisen. Dies trifft nicht zu, denn schon der EuroSCORE berücksichtigt den negativ korrelierten Kreatinin-Blutserumwert, der eine Erhöhung des vorhergesagten Risikos mit längeren dritten, grünen Balken bei kleinen CC-Werten verursacht.

Unter der Annahme der Nullhypothese der optimalen Risikomodellierung sollten keine systematischen Unterschiede (Residuen) zwischen erwarteter und beobachteter Letalität (grüne und rote Balken) bestehen. Jedoch findet sich sehr wohl ein durchgängiger Trend, d.h. eine klare „Rotverschiebung“ bzw. „Grünverschiebung“ vor und nach der 55 ml/min-CC-Grenze. Diese rot-grüne Längendifferenz korrespondiert mit der Farbe des kompakten Signifikanzindikators links von den Balken in den Inline-Graphiken. Rot signalisiert die unter- und grün die überdurchschnittliche Performanz, wobei die Farbsättigung vom p-Wert abhängt. Sinkt der p-Wert unter die allgemein anerkannte 5 %- (1 %-) Signifikanzmarke, wird das Feld mit einem Stern „*“ (bzw. Doppelstern „**“) gekennzeichnet (siehe Zeile CC=50, 60).

In diesem Feld verdichtet sich die gewonnene Information über ungewöhnliche Abweichung und wird so auf einen Blick erfassbar. Die tabellierten NLS- und $RAMQ$-Werte untermauern diese Information – was sich an der blau-roten Balkendifferenz bzw. im Längenverhältnis widerspiegelt. Sind die weiteren Kennwerte der Monte-Carlo-Verteilungsfunktion hier verdeckt, so ist die Essenz der Konfidenzanalyse sichtbar an den p-Werten und den %-Quantilenniveaus, d.h. der Position der *OM* auf der simulierten *EM*-Verteilung.

9.6 Interaktive Präsentation von Merkmalensähnlichkeiten

In Kap. 7 wurden einige Beispiele zur Visualisierung von Objekten, insbesondere von Dokumenten und ihrer Ähnlichkeitsstruktur, vorgestellt. Man kann nun den Standpunkt der Betrachtung wechseln und danach fragen, wie ähnlich oder unähnlich die im Data-Mart verfügbaren Merkmale sind.

Eine Antwort kann mittels der in Abs. 4.13 erörterten Assoziationsmaße gegeben werden. Im Folgenden wird der Entropie-basierte, symmetrische Unsicherheitskoeffizient (*normalized mutual entropy*) $U(x_i, x_j)$ in Gl. 4.82 verwendet, um die Distanzwerte $\delta_{ij} = \log(U(x_i, x_j) + 0.0001)$ zweier Patientenmerkmale i und j zu berechnen.

Abb. 9.9 zeigt einen Schnappschuss der resultierenden HMDS-Einbettung von 62 Merkmalen in den $\mathbb{H}^2$ (s. Abs. 7.3). Die Beschriftung sind die Datenbankmerkmalsbezeichner, die nach bestimmten Prä- und Suffixregeln gewählt wurden. Für den Herzchirurgen erschließen sich aus dieser Art der Darstellung interessante Zusammenhänge, für die sich z.B. fachspezifische oder allgemeine physiologische Bezüge finden – oder die entsprechende Überlegungen auslösen, also zur Hypothesenbildung sehr nützlich sind.

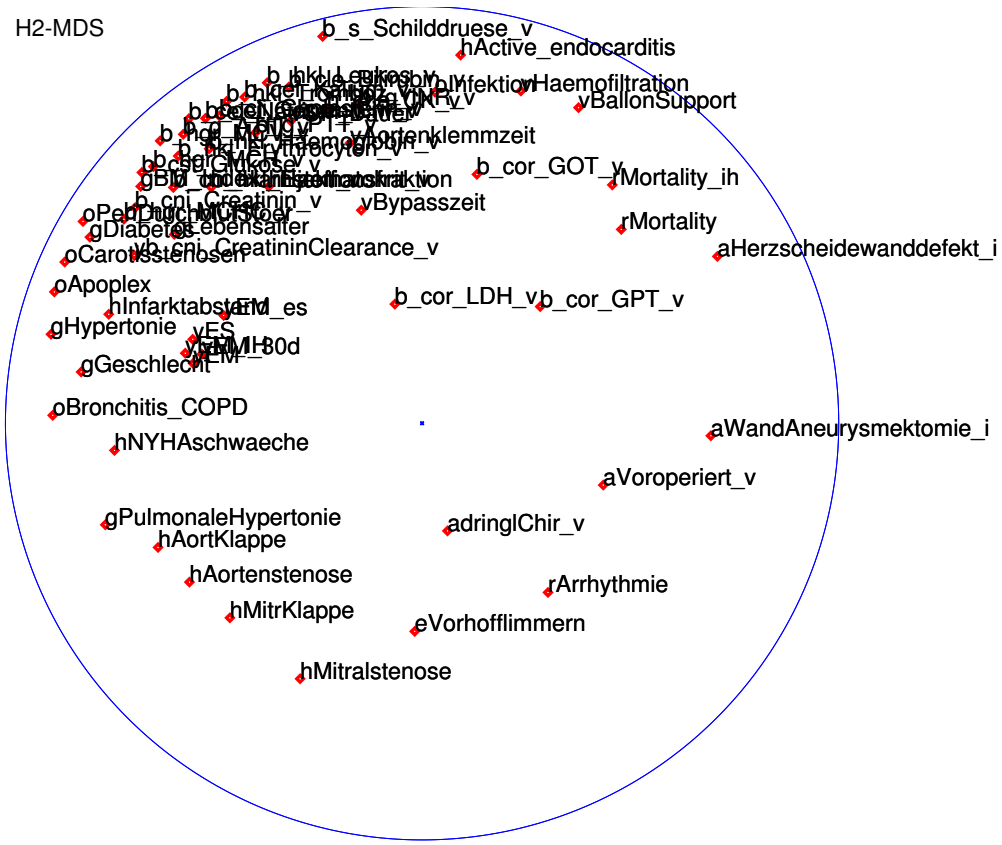


Abbildung 9.9: Ein Schnappschuss der $\mathbb{H}^2$ HMDS-Einbettung für 62 Merkmale des Data-Marts mittels des Entropie-basierten Unsicherheitskoeffizient $U(x_i, x_j)$ (Gl. 4.82). Erläuterungen siehe Text.

Zum Beispiel findet sich in Abb. 9.9 rechts oben die Information über das Verstorbensein des Patienten (*rMortality* und *rMortality_ih*) unmittelbar neben den Merkmalen, die den schlechten Zustand des Patienten nach der Operation und die deswegen notwendigen lebensrettenden Maßnahmen beschreiben (*vBallonSupport*, *vHaemofiltration*, *(v..*)=Verlauf), die Operation mit dem bekanntermaßen höchsten Sterberisiko (*aHerzscheidewanddefekt_i*) und bestimmte Blutwerte (*b..*), die einen sehr schlechten Zustand schon vor (*v..*) der Operation ausdrückt (*b_cor_LDH, b_cor_GPT_v, b_cor_GOT_v* (Hinweis auf Herzinfarkt kurz vor der Operation).

Gut erkennbar ist die Kohäsion der Merkmalsgruppe *hMitralklappe*, *hMitralstenose*, *hAortenstenose*, *hAortenklappe*. Die Begriffe bezeichnen die häufigsten Herzklappenerkrankungen. Nebenan liegen die Merkmale bzw. Folgeerkrankungen, welche bei den Patienten mit solchen Herzklappenerkrankungen typischerweise anzutreffen sind: *gPulmonale_Hypertonie* (durch Stauung des Lungenkreislaufs), *hNYHAschwaeche* (*New York Heart Ass.* Code) und *eVorhofflimmern*.

Grundsätzlich kann eine solche Darstellung der Merkmalszusammenhänge für einen erfahrenen Spezialisten ein Mittel zur ausführlichen Diskussion der mitunter sehr vielschichtigen Zusammenhänge seiner Arbeit sein. Die räumliche Präsentation der Merkmale gibt Anlass, anders über die Zusammenhänge nachzudenken und deren komplexes Wechselspiel neu zu sehen. Sich darin bewegen zu können, ist Anreiz, interaktiv neue Ansichten zu erzeugen und mental zu verstehen, abzubilden und damit, in mehrfacher Hinsicht, zu visualisieren.

Literatur

Agrawal, R. und R. Srikant (1994). Fast algorithms for mining association rules. In *Proc Int Conf on Very Large Databases (VLDB)*, pp. 487–499.

Agresti, A. (1990). *Categorial Data Analysis*. John Wiley and Sons.

Akaike, H. (1973). Information Theory and an Extention of the maximum Likelihood Principle. In *Proc 2nd Int Symp Information Theeory*, pp. 267–281.

Albert, A., B. Arnrich, U. Rosendahl, C. Beller, J. Walter, A. Mortasawi, F. Dalladaku und J. Ennker (2002, 22.-25.09 Monaco). Comparison of cold hyperkalaemic blood vs. crystalloid (Kirsch) cardioplegia in isolated coronary artery bypass grafting. In *16th Annual meeting of the European Association for cardio-thoracic surgery*, pp. 0990,276.

Albert, A., C. Beller, B. Arnrich, J. Walter, U. Rosendahl, H. Priss und J. Ennker (2nd Asia Pacific Scientific Forum, Honolulu, 2003, June). Blood cell alterations before the onset of stroke: high granulocytes count as an independent risk factor for stroke after cardiac surgery.

Albert, A., J. Walter, U. Rosendahl, B. Arnrich, C. Beller und J. Ennker (2003). Variable life adjusted displays: risikoadjustierte, zeitnahe Analysen der Performance. In *69. Jahrstagung der Deutschen Gesellschaft für Kardiologie, Herz- und Kreislaufforschung, Zeitschrift für Kardiologie*, Volume 92, Mannheim, pp. 1/1188.

Albert, A., J. Walter, U. Rosendahl, T. Schröder und J. Ennker (1999). *Dokumentationsverfahren in der Herzchirurgie V*, Chapter Wissensgewinnung aus Datenbanken mittels interaktivem Data Mining, pp. 69–73. Steinkopff Darmstadt.

Andrews, D. (1972). Plots of high dimensional data. *Biometrics 28*, 125–136.

Arnrich, B. (2001, Juni). Datamining in der Herzchirurgie. Diplomarbeit Universität Bielefeld.

Arnrich, B. und J. Walter (2000, März). Energieverbrauchsprognose Gas. Workshop ökoBudget, Bielefeld, Deutsche Bundesstiftung Umwelt.

Bay, S. D. und M. J. Pazzani (1999). Detecting Change in Categorical Data:

Mining Contrast Sets. In *Knowledge Discovery and Data Mining*, pp. 302–306.

Bertin, J. (1982). *Graphische Darstellungen*. De Gruyter, Berlin, New York.

Bi, Z., C. Faloutsos und F. Korn (2001, San Francisco, CA, August). The DGX Distribution for Mining Massive, Skewed Data. In *Proc Int Conf KDD 2001*.

Bishop, C. M., M. Svensen und C. K. I. Williams (1998). GTM: The Generative Topographic Mapping. *Neural Computation 10*(1), 215–234.

Blanusa, D. (1955). Über die Einbettung hyperbolischer Räume in euklidische Räume. *Monatshefte Mathematik 59*, 217–229.

Borg, I. und P. Groenen (1997). *Modern Multidimensional Scaling*. Springer-Verlag.

Breimann, L., J. Friedman, R. Olshen und C. Stone (1984). *Classification and regression trees*. Wadsworth Inc.

Broomhead, D. und G. King (1986). Extracting Qualitative Dynamics from Experimental Data. *Physica 20D*, 217–236.

Bulirsch, R. (1965). *Numerische Mathematik*.

Buntime, W. L. (1994). Operations for learning with graphical models. *J of AI Research*, 159–225.

Card, S. K., J. D. MacKinlay und B. Schneiderman (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann.

Chang, K.-y. und J. Ghosh (2001). A Unified Model for Probabilistic Principal Surfaces. *IEEE transactions on pattern analysis and machine inteligence 23*(1), 22–41.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer und R. Wirth (2000). CRISP-DM 1.0: step-by-step data mining guide. http://www.crisp-dm.org/.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association 68*, 361–368.

CIE (1931). *Commission Internationale de l'Eclairage Proceedings*. Cambridge University Press.

Cochran, W. (1954). Some methods for strengthening the χ^2 tests. *Biometrics 10*, 417–51.

Cox, T. F. und M. A. Cox (1994). *Multidimensional Scaling*. Monographs on Statistics and Appied Probability. Chapman and Hall.

Coxeter, H. S. M. (1969). *Introduction to Geometry*, Chapter 13.7 Barycentric Coordinates, pp. 216–221. Wiley New York.

CSRS, A. W. (2002, September). Coronary Artery Bypass Surgery in New York State 1997-1999. Technical report, New York State Department of Health.

Daugman, J. und C. Downing (1995). Gabor Wavelets for Statistical Pattern Recognition. In M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*, pp. 414–419. Cambridge, Massachusetts: MIT Press.

deLeeuw, J. und I. Stoop (1986). An upper bound for SSTRESS. *Psychometrika 51*, 149–153.

Dempster, A., N. Laird und D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B 39*, 1–38.

Duchamp, T. und W. Stuetzle (1996). Extremal Properties of Principal Curves in the Plane. *Annals of Statistics 24*, 1511–1520.

Ekman, G. (1954). Dimensions of Color Vision. *Journal of Psychology 68*, 530–536.

Ennker, J. (1998). Transparenz und Qualitätssicherung in der Herzchirurgie. *Management und Krankenhaus 11*.

Ester, M. ., H. Kriegel, J. Sander und X. Xu (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc 2nd Int Conf on Knowldege Discovery and Data Mining (KDD'96)*, pp. 226–231. AAAI Press.

Ester, M. und J. Sander (2000). *Knowlegde Discovery in Databases*. Springer.

Faloutsos, C. und K.-I. Lin (1995). FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. In M. J. Carey and D. A. Schneider (Eds.), *Proc ACM SIGMOD Int Conf on Management of Data*, San Jose, California, pp. 163–174.

Fayyad, U., G. Grinstein und A. Wierse (2002). *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann.

Fayyad, U., D. Haussler und P. Stolorz (1996, November). Mining Scientific Data. *Communications of the ACM 39*(11).

Fayyad, U., G. Piatetsky-Shapiro und P. Smyth (1996). *From Data Mining to Knowledge Discovery: An Overview*, Chapter 1, pp. 1–36. In Fayyad et al. Fayyad, Piatetsky-Shapiro, Smyth und Uthurusamy (1996).

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth und R. Uthurusamy (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Fayyad, U. und R. Uthurusamy (2002). Storage Law. *Communications of ACM Special Issue on Data Mining August*.

Fernandez, P. M. und D. Schneider (1996). The Ins and Outs (and everything in between) of Data Warehousing. In *ACM SIGMOD Tutorial Notes*.

Fienberg, S. E. (1979). Graphical methods in statistics. *The American Statistician* (33), 165–178.

Fischler, M. und O. Frischein (1987). *Readings in Computer Vision – Issues, Problems, Principles, and Paradigms*. Morgan Kaufmann.

Fogel, I. und D. Sagi (1989). Gabor Filters as Texture Discriminators. *Biological Cybernetics 61*, 103–113.

Fraser, A. und H. Swinney (1986). Independent coordinates for strage attractors from mutual information. *Physical Review 33A*, 1134–1140.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics 19*(1), 1–141. (with discussion).

Gehrke, J., R. Ramakrishnan und V. Ganti (2000). RainForest - A Framework for Fast Decision Tree Construction of Large Datasets. *Data Mining and Knowledge Discovery 4*(2/3), 127–162.

Giegerich, R., S. Kurtz und J. Stoye (1999). Efficient Implementation of Lazy Suffix Trees. In S. Verlag (Ed.), *Proc 3rd Workshop on Algorithmic Engineering (WAE99)*, Lecture Notes in Computer Science 1668, pp. 30–42.

Göppert, J. (1997). *Die topologische interpolierende selbstorganisierende Karte in der Funktionsapproximation*. Shaker Verlag.

Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*, 325–328.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrica*.

Grassberger, P. und I. Procaccia (1983). Characterization of Stange Attractors. *Physical Review Letters 50*, 346–349.

Han, J. und M. Kamber (2001). *Data Mining Concepts and Techniques*. Series in Datamanagement Systems. Morgan Kaufmann.

Hand, D., H. Mannila und P. Smyth (2001). *Principles of Data Mining*. MIT Press.

Handels, H. (2000). *Medizinische Bildverarbeitung*. B.G.Teubner Stuttgart - Leibzig.

Hanley, J. und B. McNeil (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology 143*, 29–36.

Hartigan, J. (1967). Representation of similarity matrices by trees. *J. Am. Statist, Ass. 62*, 1140–1158.

Hastie, T., R. Tibshirani und J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

Haux, R. (2002). Health Care in the Information Society: What Should Be the Role of Medical Informatics?1. *Methods of Information in Medicine 41*(1), 31–35.

Hearst, M. (1995, May). TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proc ACM SIGCHI Conf Human Factors in Computing Systems (CHI)*, Denver, CO, pp. 59–66.

Henderson, H. V. und P. F. Velleman (1981). Building regression models interactively. *Biometrics 37*, 391–411.

Hermann, T. (2002). *Sonification for Exploratory Data Analysis*. Ph. D. thesis, Technische Fakultät, Universität Bielefeld.

Heston, T. F., D. J. Norman, J. M. Barry, W. M. Bennett und R. A. Wilson (1997). Cardiac Risk Stratification in Renal Transplantation Using a Form of Artificial Intelligence. *The American journal of cardiology 79*, 415–417.

Hoerl, A. und R. Kennard (1970). Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hornik, K., M. Stinchcombe und H. White (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks 2*, 359–366.

Hunt, R. W. G. (1995). *The reproduction of colour* (5th ed.). Kingston-upon-Thames, Fountain Press.

Imhoff, C. (1999, March). Intelligent Solutions: Will the Real Data Mart Please Stand Up? *DM Review*.

Inmon, B. (1998, May). Data Mart Does Not Equal Data Warehouse. *DM Review*.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer 1*, 69–91.

Jähne, R. (2001). *Digitale Bildverarbeitung*. Springer-Verlag.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proc. 10th European Conference on Machine Learning (ECML)*, Number 1398, Chemnitz, DE, pp. 137–142.

Jones, J. und L. Palmer (1987). An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *Journal of Neurophysiology 58*, 1233–1258.

Jordan, M. I. (Ed.) (1999). MIT Press.

Kaufman, L. und P. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

Keim, D. A. (2001). *Datenvisualisierung und Data Mining*, Volume BTW Tutorial. Oldenburg: Deutsche Informatik Akademie.

Keim, D. A. und H.-P. Kriegel (1994). VisDB: Database exploration using multidimensional visualization. In *IEEE Computer Graphics and Applications*, Volume 14, pp. 40–49.

Kleinbaum, D. G. (1994). *Logistic Regression*. Springer Verlag.

Klock, H. und J. Buhmann (1997). Multidimensional Scaling by Deterministic Annealing. In *Proc EMMCVPR Venice*.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biol. Cyb. 43*(1), 59–69.

Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.), Volume 30 of *Springer Series in Information Sciences*. Springer.

Kohonen, T., S. Kaski, K. Lagus, J. Salojärvi, V. Paatero und A. Saarela (2000, May). Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery 11*(3), 574–585.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1–27.

Kuhn, K. A. und D. A. Giuse (2001). From Hospital Information Systems to Health Information Systems. *Methods of Information in Medicine 40*(4), 275–287.

Lamping, J. und R. Rao (1994). Laying Out and Visualizing Large Trees Using a Hyperbolic Space. In *ACM Symposium on User Interface Software and Technology*, pp. 13–14.

Lamping, J., R. Rao und P. Pirolli (1995). A focus+context technique based on hyperbolic geometry for viewing large hierarchies. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 401–408.

Lenz, R. und K. A. Kuhn (2001). Intranet Meets Hospital Information Systems: The Solution to the Integration Problem? *Methods of Information in Medicine 40*(2), 99–105.

Lewis, D. (1997). Reuters-21578 Dataset Distribution 1.0. http://www-.daviddlewis.com/resources/testcollections/reuters21578/.

Littmann, E., A. Meyering, J. Walter, T. Wengerek und H. Ritter (1992). Neural Networks for Robot Hand Control. In M. van der Meer (Ed.), *Statusseminar des BMFT Neuroinformatik*, pp. 253–262. Deutsche Forschungsanstalt für Luft- und Raumfahrt e.V.

Lodhi, H., J. Shawe-Taylor, N. Cristianini und C. Wat (2001). Text Classification using String Kernels. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*, pp. 563–569. MIT Press.

Loh, W.-Y. und Y.-S. Shih (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*.

Lovegrove, J., O. Valencia, T. Treasure, C. Sherlaw-Johnson und S. Gallivan (1997). Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet 350*, 1128–1130.

MacAdam, D. (1942). Visual sensitivities to color differences in daylight. *J Opt Soc America*, 247–274.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *5th Merkley Symp Math Stat Prob*, Volume 1, pp. 281–297.

Mandelbrot, B. (1983). *The Fractal Geometry of Nature*. Freeman and Co.

Mann, H. und D. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist. 18*, 52–60.

Marquardt, D. W. (1963). *J. Soc Appl. Math. 11*, 431–441.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Martinetz, T., S. Berkovich und K. Schulten (1993). "Neural-Gas" Network for Vector Quantization and its Application to Time-Series Prediction. *IEEE TNN 4*(4), 558–569.

McCulloch, W. und W. Pitts (1943). A Logical Calculus of Ideas Immanent in the Nervous System. *Bulletin of mathematical Biophysics 5*, 115–133.

Michie, D., D. Spiegelhalter und C. Taylor (1994). *Machine Learning, Neural and Statistical Classication*. Ellis Horwood.

Miettinen und Nurminen (1985). Comparative analysis of two rates. *Statistics in Medicine 4*, 213–226.

Minsky, M. und S. Papert (1969). *Perceptrons : an introduction to computational geometry*. MIT Press, Cambridge.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Moise, E. (1974). *Elementary Geometry from an Advanced Standpoint*. Addison-Wesley, Reading, MA.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics 38*(8).

Morgan, F. (1993). *Riemannian Geometry: A Beginner's Guide*. Jones and Bartlett Publishers.

Munzner, T. (1997). H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space. In *Proceedings of the 1997 IEEE Symposium on Information Visualization, Phoenix, AZ*, pp. 2–10.

Munzner, T. (1998). Drawing Large Graphs with H3Viewer and Site Manager. In *Proceedings of Graph Drawing '98, Montreal, Canada, Springer-Verlag*, Lecture Notes in Computer Science 1547, pp. 384–393.

Nashef, S., F. Roques, B. Hammill, E. Peterson, P. Michel, F. Grover, R. Wyse, und T. Ferguson (2002). Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. *European Journal of Cardio-thoracic Surgery 22*(1), 101–105.

Nashef, S., F. Roques, P. Michel, E. Gauducheau, S. Lemeshow und R. Salamon (1999). European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-thoracic Surgery 16*, 9–13.

Obermayer, K., H. Ritter und K. Schulten (1990, nov). A principle for the formation of the spatial structure of cortical feature maps. In *Proc. Natl. Acad. Sci., USA Neurobiology*, Volume 87, pp. 8345–8349.

Ontrup, J. und H. Ritter (1998). Perceptual Grouping in a Neural Model: Reproducing Human Texture Perception. Technical report, Technical Report SFB360-TR-98/6.

Ontrup, J. und H. Ritter (2001). Text Categorization and Semantic Browsing with Self-Organizing Maps on non-euclidean Spaces. In *Proc PKDD-2001*, pp. 338–349. Springer LNAI 2168.

Orr, G. B. und K.-R. Müller (Eds.) (1998). *Neural Networks: Tricks of the Trade*, Volume 1524 of *Lecture Notes in Computer Science*. Springer.

Parsonnet, V., A. Bernstein und M. Gera (1996). Clinical usefulness of risk-stratified outcome analysis in cardiac surgery in New Jersey. *The annals of thoracic surgery 61*(2), 8–11.

Poggio, T. und F. Girosi (1990). Networks for Best Approximation and Learning. *Proc. of IEEE 78*, 1481–1497.

Poloniecki, J., O. Valencia und P. Littlejohns (1998, June). Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal 316*, 1697–1700.

Powell, M. (1987). Radial basis functions for multivariable interpolation: A review. In *Algorithms for Approximation*, pp. 143–167. Oxford: Clarendon Press.

Press, W., B. Flannery, S. Teukolsky und W. Vetterling (1988). *Numerical Recipes in C – the Art of Scientific Computing*. Cambridge Univ. Press.

Puzicha, J., Y. Rubner, C. Tomasi und J. Buhmann (1999). Empirical evaluation of dissimilarity measures for color and texture. In *Proc IEEE Int Conf on Computer Vision (ICCV)*, pp. 1165–1173.

Pyle, D. (1999). *Data Preparation For Data Mining*. Morgan Kaufmann.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE 77*(2), 257–286.

Rao, R. und S. Card (1994, April). The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information. In *Proc. ACM Conf. on Human Factors in Computing Systems SIGCHI*.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Ritter, H. (1993). Parametrized Self-Organizing Maps. In S. Gielen and B. Kappen (Eds.), *Proc. Int. Conf. on Artificial Neural Networks (ICANN-93), Amsterdam*, pp. 568–575. Springer Verlag, Berlin.

Ritter, H. (1999). Self-Organizing Maps on non-euclidean Spaces. In S. Oja, E. & Kaski (Ed.), *Kohonen Maps*, pp. 97–110. Amsterdam: Elsevier.

Ritter, H., T. Martinetz und K. Schulten (1991). *Neuronale Netze* (2 erw. ed.). Addison Wesley.

Ritter, H. J., T. M. Martinetz und K. J. Schulten (1989). Topology-Conserving Maps for Learning Visuo-Motor-Coordination. *Neural Networks 2*, 159–168.

Roques, F., S. Nashef, P. Michel, E. Gauducheau, C. de Vincentiis, E. Baudet, J. Cortina, M. David, A. Faichney, F. Gabrielle, E. Gams, A. Harjula, M. Jones, P. Pintor, R. Salamon und L. Thulin (1999). Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *Eur J Cardiothorac Surg. 15*(6), 816–822++.

Rubner, Y. und C. Tomasi (2000). *Perceptual Metrics for Image Database Navigation*. Kluwer Academic Publishers, Boston.

Rubner, Y., C. Tomasi und L. J. Guibas (1998, January). A Metric for Distributions with Applications to Image Databases. In *Proc IEEE Int Conf on Computer Vision (ICCV)*, pp. 59–66. Bombay, India.

Salton, G. und C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 5*(24), 513–523.

Sammon, Jr., J. W. (1969). A non-linear mapping for data structure analysis. *IEEE Transactions on Computers 18*, 401–409.

Schölkopf, B. und A. J. Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Schwartz, E. (1977). Spatial Mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biol. Cybernetics 25*, 181–194.

Shannon, C. E. (1948, July und Oct.). A Mathematical Theory of Communication. *Bell System Technical J. 27*, 379–423 and 623–656.

Shaw, L. J., R. Hachamovitch, G. V. Heller, T. H. Marwick, M. I. Travin, A. E. Iskandrian, K. Kesler, M. S. Lauer, R. Hendel, S. Borges-Neto, H. C. Lewin, D. S. Berman und D. Miller (2000). Noninvasive Strategies for the Estimation of Cardiac Risk in Stable Chest Pain Patients. *The American Journal of Cardiology 86*.

Shepard, D. (1968). A two-dimensional function for irregularly spaced data. In *In 23rd ACM National Conference*, pp. 517–524.

Skupin, A. (2002). A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications*, 50–58.

Smeraldi, F., O. Carmona und J. Bigun (2000). Saccadic search with Gabor features applied to eye detection and real-time head tracking. *Image and Vision Computing*.

Sonka, M., V. Hlavac und R. Boyle (1998). *Image Processing, Analysis and Machine Vision*. Pacific Grove, CA.

Srikant, R. und R. Agrawal (1995). Mining Generalized Association Rules. In *Proc Int Conf on Very Large Databases (VLDB)*, pp. 407–419.

Srikant, R. und R. Agrawal (1996). Mining Quantitative Association Rules in Large Relational Tables. In *ACM SIGMOD Int. Conf. on Management of Data*, pp. 1–12.

Strubecker, K. (1969). *Differentialgeometrie III: Theorie der Flachenkrümmung*. Walter de Gruyter, Berlin.

Takens, F. (1981). Detecing strange attractors in turbulence. In D. Rand and L. S. Young (Eds.), *Dynamical Systems and Turbulences*, Volume 898 of *Lecture Notes in Mathematics*, pp. 366–381. Springer, Berlin.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing 2*, 183–190.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. 58*, 267–288.

Tuckey, J. W. (1977). *Exploratory Data Analysis*. Reading MA: Addison- Wesley.

Tufte, E. (1983). *The Visual Display of Quantitative Information*. Graphics Press.

TwoCrows (1999). *Introduction to Data Mining and Knowledge Discovery* (3rd ed.). Pontomac, MD: Two-Crows-Corporation.

Walter, J. (1991). Visuo-motorische Koordination eines Industrieroboters und Vorhersage chaotischer Zeitserien: Zwei Anwendungen selbstlernender neuronaler Algorithmen. Diplomarbeit, Physik Department, Technische Universität München.

Walter, J. (1998). PSOM Network: Learning with Few Examples. In *Proc. Int. Conf. on Robotics and Automation (ICRA-98)*, pp. 2054–2059.

Walter, J. (2004). H-MDS: a New Approach for Interactive Visualization with Multidimensional Scaling in the Hyperbolic Space. *Information Systems, Elsevier 29*(4), 273–292.

Walter, J. und B. Arnrich (2000). Gabor Filters for Object Localization and Robot Grasping. In *Proc. Int. Conf. Pattern Recognition (ICPR, Lisbon)*, Volume 4, pp. 124–127.

Walter, J., B. Arnrich, U. Rosendahl und J. Ennker (2001). *Medizinischer Jahresbericht 2000*, Chapter Satistiken. Herzzentrum Lahr/Baden.

Walter, J., B. Arnrich und C. Schering (2000). Learning Fine Positioning of a Robot Manipulator based on Gabor Wavelets. In *Proc. Int. Joint Conf Neural Networks (IJCNN Como, Italy)*, Volume 5, pp. 137–142.

Walter, J., A. Mortasawi, B. Arnrich, A. Albert, I. Frerichs, U. Rosendahl und J. Ennker (2003). Creatinine clearance versus serum creatinine as a risk factor in cardiac surgery. *BioMed Central Surgery 3*(4).

Walter, J., C. Nölker und H. Ritter (2000). The PSOM Algorithm and Applications. In *Proc. Symp. Neural Computation*, pp. 758–764.

Walter, J., J. Ontrup, D. Wessling und H. Ritter (2003, November). Interactive Visualization and Navigation in Large Data Collections using the Hyperbolic Space. In *IEEE Int Conf Data Mining*, Melbourne, Florida, USA, pp. 355–362.

Walter, J. und H. Ritter (1995). Local PSOMs and Chebyshev PSOMs – Improving the Parametrised Self-Organizing Maps. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN-95), Paris*, Volume 1, pp. 95–102.

Walter, J. und H. Ritter (1996). Rapid Learning with Parametrized Self-Organizing Maps. *Neurocomputing 12*, 131–153.

Walter, J. und H. Ritter (2002, August, Edmonton, Canada). On Interactive Visualization of high-dimensional Data using the Hyperbolic Plane. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 123–131. SigKDD.

Walter, J., H. Ritter und K. Schulten (1990, June). Non-linear Prediction with Self-organizing Maps. In *Int. Joint Conf. on Neural Networks (IJCNN), San Diego, CA*, pp. 587–592.

Walter, J. und K. Schulten (1993). Implementation of Self-Organizing Neural Networks for Visuo-Motor Control of an Industrial Robot. *IEEE Transactions in Neural Networks 4*(1), 86–95.

Walter, J. A. (1996). *Rapid Learning in Robotics*. Dissertation, Technische Fakultät, Universität Bielefeld. http://www.techfak.uni-bielefeld.de/∼walter/pub/.

Wang, X., J. T. L. Wang, K.-I. Lin, D. Shasha, B. A. Shapiro und K. Zhang (2000, May). An Index Structure for Data Mining and Clustering. *Knowledge and Information Systems 2*, 161–184.

Weigend, A. S. und N. A. Gershenfeld (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley.

Wersing, H., J. J. Steil und H. Ritter (2001). A Competitive Layer Model for Feature Binding and Sensory Segmentation. *Neural Computation 13*(2), 357–387.

Wolf, P., R. D'Agostino, A. Belanger und W. Kannel (1991). Probability of stroke: a risk profile from the Framingham study. *Stroke 22*, 312–318.

Wyszecki, G. und W. Styles (1982). *Color Science: Concepts and Methods, Quantitative Data and Fomulae*. John Wiles and Sons, NY.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval 1*(1/2), 69–90.

Young, G. und A. Hauseholder (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika 3*, 19–22.

Yu, K., Z. Wen, X. Xu und M. Ester (2001). Feature Weighting and Instance Selecion for Callaborative Filtering. In *Proc Management of Information on the Web - Web Data and Text Mining (MIWI)*.

Zar, J. (1996). *Biostatistical analysis* (3ed ed.). Prentice Hall.

Zhang, T., R. Ramakrishnan und M. Livny (1996). BIRCH: an efficient data clustering method for very large databases. pp. 103–114.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *J. Gen. Psych. 33*, 251–256.