Angewandte Multivariate Statistik

Vorlesung 701-0102-00, mit Ergänzung 401-0102-99

und Weiterbildungslehrgang in Angewandter Statistik

Frühlingssemester 2010, ETH Zürich

Werner Stahel
Seminar für Statistik, ETH Zürich

Februar-Mai 2010

Inhaltsverzeichnis

1 Einleitung

1.1 Fragen der multivariaten Statistik

a Zur Charakterisierung von Personen, Objekten oder anderer Beobachtungseinheiten werden meistens einige bis viele Merkmale erfasst. Wenn bei Patienten der Blutdruck, das Alter, das Geschlecht, das Gewicht, die Behandlungsart und weitere Daten notiert werden, dann liegt das Interesse oft darin, die Zielgrösse Blutdruck als Funktion von erklärenden Grössen darzustellen, um kausale Zusammenhänge wie die Wirksamkeit der Behandlung zu erfassen. Dies führt zur Grundfragestellung der statistischen Regressionsmethodik. Oft sind aber **mehrere Merkmale von gleichrangigem Interesse**: Bei Insektenlarven werden die Längen mehrerer Gliedmassen ausgemessen, bei Patienten sind unterer und oberer Blutdruck, die Konzentration mehrerer Substanzen im Blut oder andere messbare Grössen wichtig, bei einer chemischen Reaktion sind die Konzentrationen mehrerer Agenzien beteiligt, usw.

b Die **multivariate Statistik** befasst sich mit Fragestellungen, in denen mehrere Variable gemeinsam und gleich bedeutend betrachtet werden sollen.

c Für die **grafische Darstellung** *einer einzigen* Variablen stehen Varianten von Histogrammen und Boxplots im Vordergrund. Für die Wiedergabe der gemeinsamen Verteilung von *zwei* Variablen bildet das Streudiagramm (Abb. 1.2.b(i)) die „Grundfigur“. Die **gemeinsame Verteilung von *mehreren* Variablen** ist schon schwieriger darzustellen, und so beruhen die **grafischen Methoden** in der multivariaten Statistik auf einer Vielfalt von Ideen.

d Die Grundlage der Statistik mit einer einzigen Variablen oder einer einzigen Zielgrösse in der Regression bilden **Wahrscheinlichkeits-Verteilungen**. Zum Verständnis von Zusammenhängen zwischen mehreren Variablen ist es noch wichtiger, **Wahrscheinlichkeitsmodelle** zu entwickeln, die die Zusammenhänge beschreiben können. Auf dieser Grundlage kann man die **Grundfragen der schliessenden Statistik** nach dem Zusammenhang zwischen beobachteten Daten und Parameterwerten eines Modells beantworten. Schliesslich wird es wichtig sein, Methoden zur Überprüfung von Modell-Voraussetzungen zu erörtern.

e Bei der Betrachtung einer einzelnen Zufallsvariablen sind die Fragestellungen der **Schätzung und Prüfung eines Mittelwertes** und des **Vergleichs von zwei oder mehreren Gruppen** die grundlegenden Problemstellungen, die am Anfang der schliessenden Statistik stehen. Sie werden in der multivariaten Statistik auf mehrere Variable ausgedehnt, ebenso die übliche **Varianzanalyse** und die **Regression** mit einer Zielgrösse auf die gemeinsame Betrachtung **mehrerer Zielgrössen**.

f Immer wieder geht man in der multivariaten Statistik von Fragen aus, die zunächst bei der Untersuchung einer einzelnen Zufallsgrösse nahe liegen, und verallgemeinert das Problem und die entsprechenden Methoden auf den Fall mehrerer Variablen. Dem einfacheren Fall geben wir deshalb den Namen **univariate Statistik**.

1.2 Beispiele

a ▷ **Iris-Arten.** Im Jahre 1935 hat der Biologe E. Anderson an Iris-Blumen verschiedener Artzugehörigkeit die Länge und Breite von Sepal-Blättern und Petal-Blättern gemessen; beide Blättertypen sind Teile der Blüte. R. A. Fisher (1936) hat die Daten von Anderson als grundlegendes Beispiel für ein multivariates Verfahren, die „Diskriminanzanalyse“, verwendet; dadurch wurde dieser Datensatz wohl zum berühmtesten der Statistik überhaupt. Von den 150 Pflanzen, deren Messungen Fisher benützte, gehörten je 50 den Arten Iris setosa, Iris virginica und Iris versicolor an.

Die Fragestellung der Studie lautete, ob sich rein anhand der gemessenen Blütenblätter die **Pflanzen den einzelnen Arten zuweisen** lassen. Es wurde vermutet, dass Iris versicolor eigentlich ein Hybrid der beiden anderen Arten sei. (Diese Informationen zum Datensatz sind Andrews and Herzberg (1985) entnommen.)

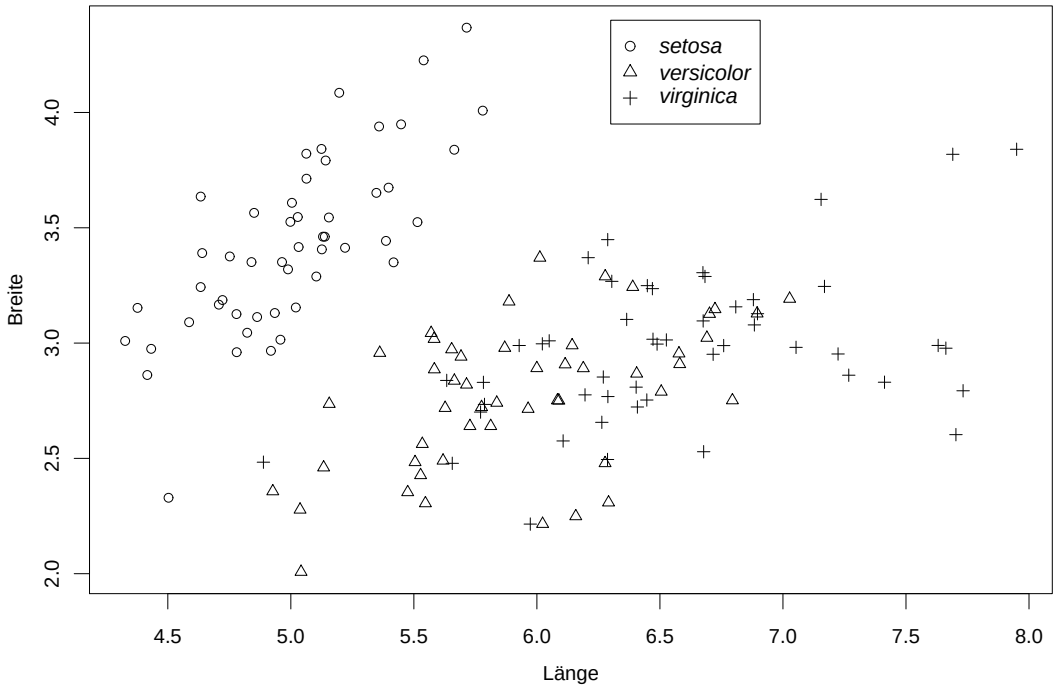

Abbildung 1.2.b (i): Streudiagramm der Länge und Breite der Sepalblätter im Beispiel der Irisblüten

b ▷ Betrachten wir zunächst nur zwei Variable, die Länge und Breite der Sepalblätter! Es ist naheliegend, die Daten in einem Streudiagramm aufzuzeichnen (Abbildung 1.2.b (i)). Die Punkte, die den einzelnen Pflanzen entsprechen, sind durch Symbole markiert, die die Art wiedergeben.

Man sieht, dass – wie zu erwarten ist – die beiden Variablen mit einander zusammenhängen; längere Blütenblätter sind meist auch breiter als kürzere. Man spricht von Korrelation der beiden Variablen. Ebenfalls ist ersichtlich, dass die gemeinsame Betrachtung dieser korrelierten Variablen weiter hilft als die separate Darstellung (Abbildung 1.2.b (ii)): In der gemeinsamen Darstellung kann die Art „setosa“ klar von den anderen beiden getrennt werden, und das gelingt weder für die Länge noch für die Breite alleine. ◁

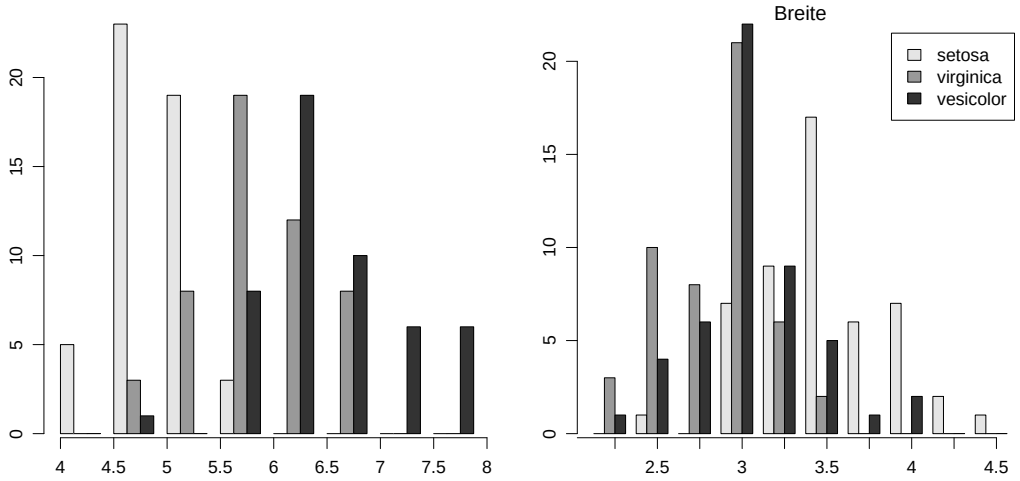

Abbildung 1.2.b (ii): Histogramme von Länge und Breite der Sepalblätter, aufgeteilt nach den drei Arten Iris setosa, virginica und versicolor

c ▷ **Ader-Verengung. Diagnostische Tests in der Medizin** dienen dazu, die Patienten im Hinblick auf eine spezifische Krankheit in Kranke und Gesunde einzuteilen. Oft wird dies auf Grund einer Messung einer einzelnen Variablen oder gar einer Ja-Nein-Antwort getan. Es kann aber eine wesentlich genauere Diagnose möglich sein, wenn mehrere Symptom-Merkmale gleichzeitig verwendet werden.

Ein einfaches Beispiel liefert die Diagnose einer Ader-Verengung auf Grund des Herzschlag-Volumens (Vol) und des Pulses (Rate). Abbildung 1.2.c zeigt die Daten nach logarithmischer Transformation. (Quelle: Finney, 1947, Biometrika 34 und auch Fahrmeir, Hamerle and Tutz, 1996). ◁

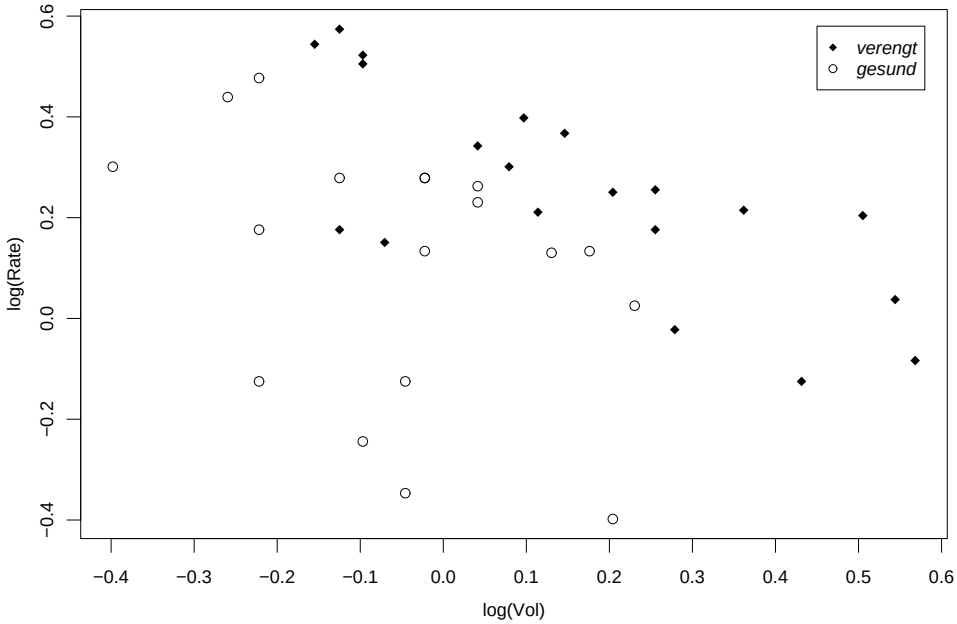

Abbildung 1.2.c: Daten im Beispiel der Ader-Verengung

d ▷ **Fossilien.** Aus Fossilien, die man in verschiedenen Schichten von Meeres-Ablagerungen findet, will man auf Umweltbedingungen (Temperatur, Nährstoffgehalt) der entsprechenden Zeitperioden schliessen.

Bollmann, Henderiks and Brabec (2002) sammelten deshalb Messungen verschiedener morphologischer Merkmale von Cocolithen der Art *Gephyrocapsa* an 110 Stellen der Weltmeere in der obersten Schicht (Holozän). Abbildung 1.2.d zeigt diese Stellen und ein schematisches Bild eines Cocolithen. Die Grösse liegt im Bereich von 1-5 μm. An den Probenahmestellen wurden auch die heutigen Umweltbedingungen aufgenommen.

Zusätzlich wurden Cocolithen von tieferen Schichten ausgemessen. Die Grundidee besagt, dass die morphologischen Merkmale von den Umweltbedingungen abhängen. Die Beziehungen kann man auf Grund der Proben von heute modellieren. Falls die Beziehungen gleich geblieben sind, kann man sie benützen, um **von den morphologischen Merkmalen der tieferen Schicht auf die damaligen Umweltbedingungen** zu schliessen.

Die Gruppe fand, dass sich bessere Prognosen ergaben, wenn die Individuen zuerst, mit Hilfe der erwähnten Messungen, in Unterarten eingeteilt und dann die Anteile der Unterarten in den Proben für die Vorhersage eingesetzt wurden. Es stellt sich also auch die Frage, wie man solche **Unterarten festlegen** soll, und wie man die einzelnen Individuen diesen Unterarten zuordnen soll. Eine solche Definition ist in Teilfigur C angedeutet. ◁

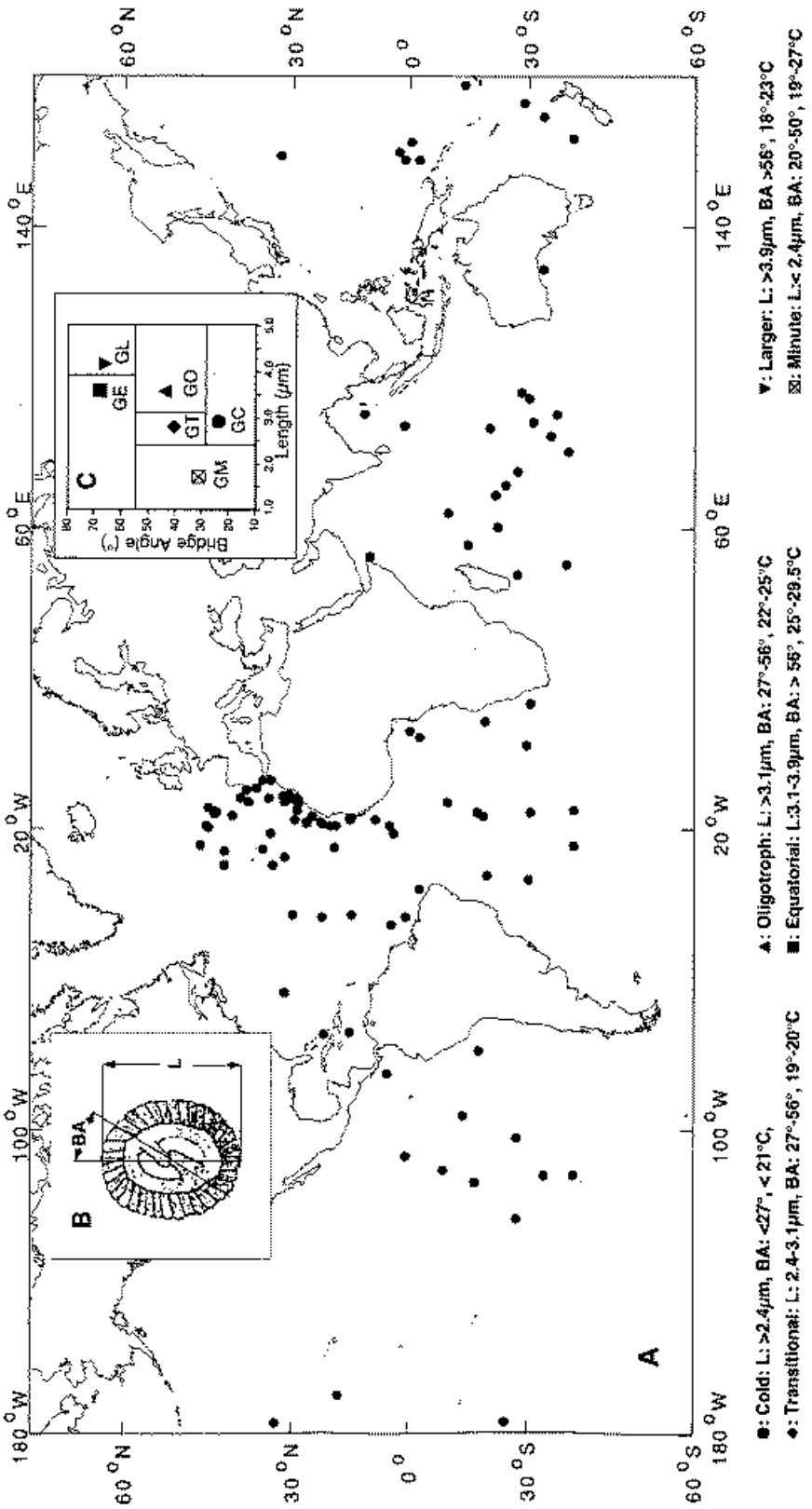

Abbildung 1.2.d: Probenahmestellen in Beispiel der Fossilien (A). (B) Für die Grundform der Cocolithen sind mit der Länge und dem „Brückenwinkel“ zwei wichtige Messungen angegeben. (C) zeigt eine Einteilung in Unterarten auf Grund der beiden Messgrössen.

e ▷ **Ökosystem.** In Studien über Ökosysteme wird die Beziehung zwischen verschiedenen Arten von Variablen untersucht. Jewell, Güsewell, Berry, Käuferle, Kreuzer and Edwards (2005) stellten die Frage, wie Beweidung die Vegetation beeinflusst. Sie bestimmten auf einer Alp im Kanton Tessin auf 82 Probeflächen erstens die Intensität der Beweidung via Beobachtung der Aufenthaltsorte und über die Exkremente, zweitens bestimmten sie bodenchemische Grössen – den pH, die Phosphat-, Nitrat- und Kohlenstoff-Konzentration – und Häufigkeiten von 64 Pflanzenarten. Der Ort der Probenahme legt überdies physikalisch Umweltgrössen wie Hangneigung und -exposition und Höhe über Meer fest.

Abbildung 1.2.e zeigt, wie die Abundanzen der 6 häufigsten Arten von der Beweidungsintensität abhängen. Wie kann man die Vegetation als ganzes charakterisieren, ohne die Abundanzen aller 64 Arten einzeln studieren zu müssen? Wie stark hängen die Variablengruppen mit einander zusammen? ◁

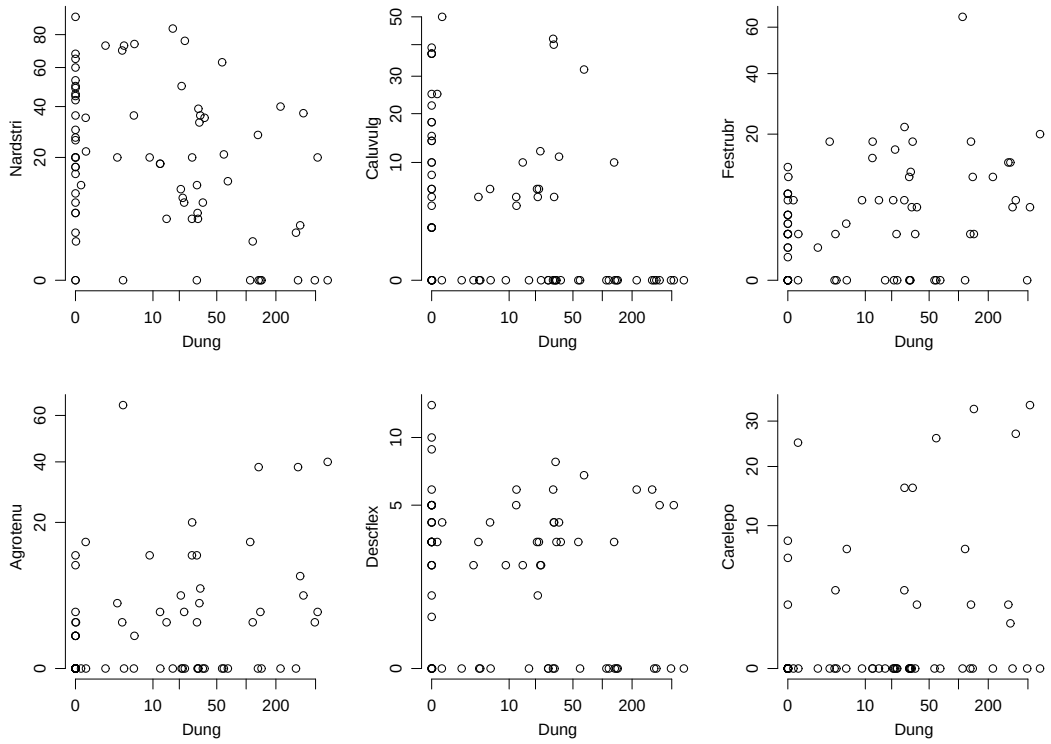

Abbildung 1.2.e: Abhängigkeit der 6 häufigsten Pflanzen-Arten von der Beweidungsintensität im Beispiel des Ökosystems

f ▷ **Abstimmungen.** Mit statistischen Analysen von Abstimmungs-Resultaten kann man es bis in die Tagespresse schaffen. Für 14 eidgenössische Vorlagen der Jahre 1995-96 wurden die Ja-Stimmen-Anteile der Kantone zusammengestellt. Zeigen sich darin interpretierbare Muster?

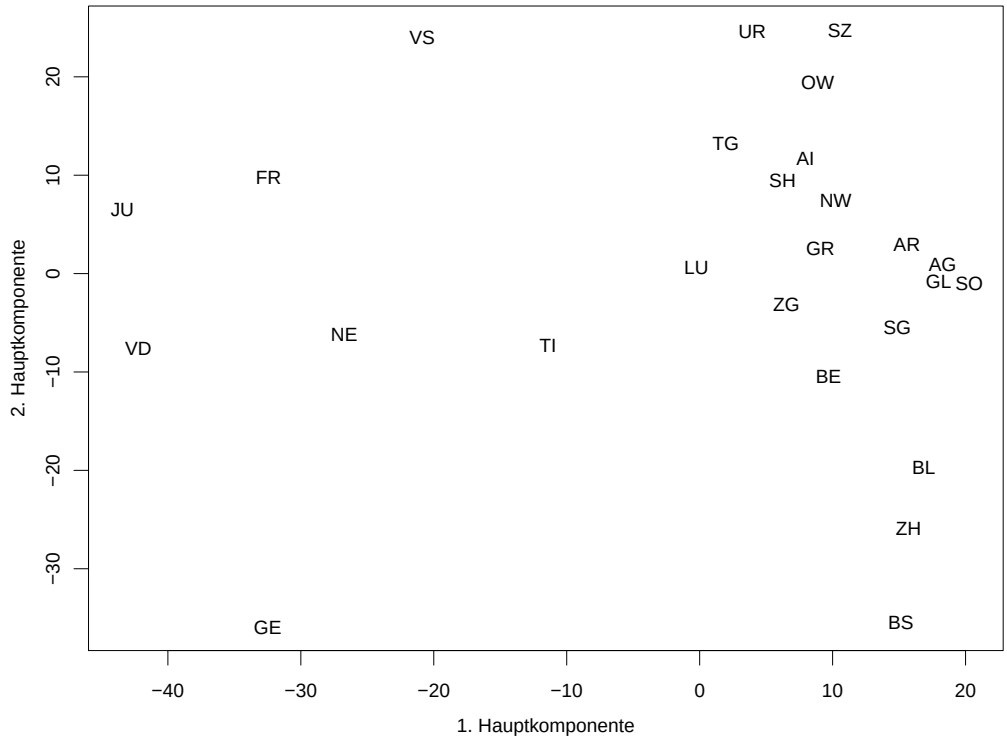

Abbildung 1.2.f: Erste zwei Hauptkomponenten im Beispiel der Abstimmungen

Abbildung 1.2.f zeigt eine **grafische Darstellung der Ergebnisse,** die sich aus einer multivariaten Analyse, einer so genannten Hauptkomponenten-Analyse der Anteile ergibt. Man sieht in der linken Hälfte Westschweizer Kantone, rechts die Ostschweiz und in der Mitte das südliche Tessin. Oben erscheinen die ländlichen Kantone, unten die städtischen. Diese „Geographie“ wurde zur Erzeugung der Darstellung nicht benützt. Offenbar genügt die Ähnlichkeit im Abstimmungsverhalten, um diese Gegensätze wiederzufinden. ◁

g ▷ **NIR-Spektren.** Spektren spielen in der Chemie eine wichtige Rolle. Sie ermöglichen es, die **Zusammensetzung von Gemischen** zu ermitteln ohne chemische Analyse, also ohne für jeden Stoff, der in Frage kommt, mit einer kleinen Probe eine ganz spezifische Nachweis-Reaktion durchzuführen. Es gibt viele Arten von Spektren. Die optischen Spektren messen die Absorption von Licht, das durch eine Probe hindurchgestrahlt wird (oder die Reflexion an einer festen Probe) in Abhängigkeit von der Wellenlänge des Lichts.

Im Idealfall zeigt sich für jede Substanz im Gemisch ein „Peak“ im Spektrum; für eine bestimmte Wellenlänge absorbiert diese Substanz viel Licht, für die anderen Wellenlängen vernachlässigbar wenig, und diese charakteristischen Wellenlängen sind für die verschiedenen Substanzen so unterschiedlich, dass sich die Peaks nicht überlappen. Die Analyse solcher Spektren ist relativ einfach: Die Grösse (Fläche) der Peaks ist direkt proportional zum Anteil der entsprechenden Substanz im Gemisch.

Für Wellenlängen im Bereich des nahen Infrarot (NIR) ist das leider nicht der Fall. Die einzelnen Substanzen absorbieren über grössere Bereiche eines NIR-Spektrums, die

„Peaks“ sind verschmiert und überlappen sich stark. Es ist besser, die Idee des Peaks zu ersetzen durch die Vorstellung eines spezifischen Spektrums allgemeinerer Form für jede Substanz.

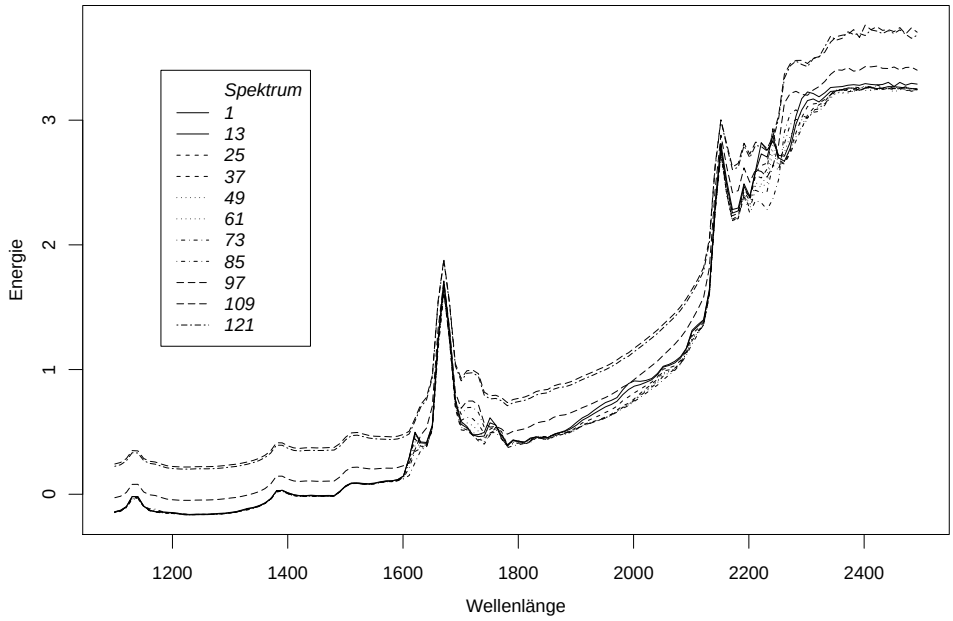

Abbildung 1.2.g (i): 11 NIR-Spektren, die den Verlauf einer chemischen Reaktion dokumentieren

Die Absorption eines Gemisches für eine bestimmte Wellenlänge können wir als Variable ansprechen. Für jede Wellenlänge, für die die Absorption gemessen wird, erhalten wir also eine Variable. Abbildung 1.2.g (i) zeigt 11 von 121 NIR-Spektren, die im Verlauf einer chemischen Reaktion aufgenommen wurden.

Die multivariate Statistik ist dazu da, aus der gemeinsamen Information, die in diesen (beliebig) vielen Variablen steckt, das Wesentliche herauszuholen – wenn möglich die Konzentration der Substanzen in jedem Gemisch. Dabei kommt ihr ein Naturgesetz zu gute: Das Spektrum eines Gemisches ist die gewichtete Summe der Spektren der einzelnen Substanzen; die Gewichte sind die Anteile der Substanzen im Gemisch. Natürlich gilt dieses Gesetz nicht exakt, sondern nur als Näherung und nur bis auf Messfehler. Es wird oft als Gesetz von Lambert und Beer bezeichnet (obwohl das eigentliche Gesetz von Lambert und Beer nur den Zusammenhang der Intensität mit den charakteristischen Grössen der Messapparatur beschreibt).

Mit diesem Grundgesetz und entsprechenden multivariaten Methoden können in der Chemie Reaktionen erforscht oder Prozesse überwacht werden. Im Beispiel können mit ihrer Hilfe 4 Phasen des Prozesses identifiziert werden, die nacheinander ablaufen; Abbildung 1.2.g (ii) stellt die Abfolge dieser Phasen im zeitlichen Verlauf dar.

Das Gesetz werden wir durch ein Wahrscheinlichkeitsmodell beschreiben, das als **Modell der linearen Mischung** bezeichnet wird. Es findet auch in der Analyse von Umweltschadstoffdaten und anderen Gebieten Anwendung. ◁

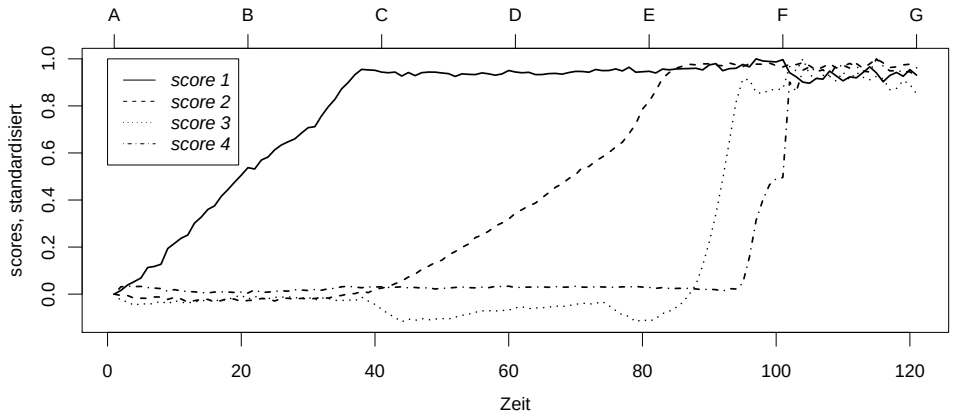

Abbildung 1.2.g (ii): Zeitliche Abfolge von vier Phasen des Prozesses im Beispiel der NIR-Spektren

h **Kunden-Segmentierung.** Ein wichtiges Anwendungsfeld für multivariate Daten stellen die immer grösser werdenden Dateien über Kunden von grösseren Firmen dar. Für Marketing-Zwecke sollen Kunden mit speziellen Interessen oder speziellem Verhalten gefunden werden, damit sie gezielt mit Werbung eingedeckt oder sonst speziell behandelt werden können.

In gewissen Anwendungen benützt man Daten aus der Vergangenheit, um Regeln für Vorhersagen abzuleiten. Beispielsweise werden die Daten von Schuldnern einer Bank benützt, um vorherzusagen, welche von ihnen mit erhöhter Wahrscheinlichkeit zahlungsunfähig werden. Um die Regel herzuleiten, benützt man Daten aus der Vergangenheit und vergleicht die Gruppe jener, die zahlungsunfähig wurden, mit jenen, die keine Probleme machten. Das ist wieder eine Frage der Unterscheidung zwischen Gruppen. Ebenso kann man vorhersagen wollen, welcher Ertrag von den einzelnen Kunden erwartet werden kann. Das wäre eine Frage der Regression.

In anderen Anwendungen gibt es keine bestimmte Zielgrösse oder Klassierung, die man vorhersagen will, sondern man möchte gerne „sinnvolle Gruppen“ in den Daten finden.

1.3 Fragestellungen

a Die vorgestellten Beispiele illustrieren die folgenden Fragestellungen, die mit den Methoden der multivariaten Statistik beantwortet werden können.

- Zunächst will man sich anhand von **grafischen Darstellungen** ein Bild von den Daten machen, sei es zur **Bereinigung** (screening) der Daten, zur **explorativen Analyse**, also zur Suche nach Unerwartetem, oder später zur Veranschaulichung von Zusammenhängen, die sich aus Analysen mit anderen Methoden ergeben haben.

b
- Betrachtet man die gemeinsame Verteilung von zwei Variablen, dann kann man fragen: **Sind die Variablen unabhängig?** Wenn nicht: **Welche Art von Abhängigkeit ist vorhanden? Wie stark ist sie?** Die bekannteste Art der Charakterisierung eines Zusammenhangs ist der Korrelationskoeffizient (von Pearson), der die Stärke eines linearen Zusammenhangs misst. Man kann die Fragen und die Begriffe auch auf Zusammenhänge zwischen Gruppen von Variablen ausdehnen.

  Vollständig sind die Zusammenhänge erst bestimmt, wenn man ein Modell für die gemeinsame Verteilung aller Variablen festlegt. Die Verallgemeinerung der **Normalverteilung** auf den Fall von mehreren Variablen bildet das Modell, das in der multivariaten Statistik noch eine viel zentralere Rolle spielt als in der univariaten.

c
- Im Datensatz über die Iris-Pflanzen sind mit den drei Arten drei Gruppen vorgegeben. Eine erste Frage lautet: **Gibt es Unterschiede zwischen Gruppen von Beobachtungen?** Wie in der univariaten Statistik ist für diese Frage ein statistischer **Test** gesucht, der die Nullhypothese **gleicher Erwartungswerte** (oder gleicher Werte eines anderen Lageparameters) oder gar gleicher Verteilung für die Gruppen prüft.

d
- Die Unterschiede zwischen Gruppen sind in den Beispielen der Iris-Pflanzenund der Fossilien vor allem nützlich, wenn sie es ermöglichen, neue Individuen, für die die Gruppenzugehörigkeit nicht bekannt ist, einer Gruppe zuzuordnen. Wir fragen also: **Welcher Gruppe soll man ein neues Individuum** mit unbekannter Gruppenzugehörigkeit auf Grund seiner Merkmale **zuordnen?** Daran schliesst sich die Frage an: **Wie sicher kann man bei einer solchen Zuordnung sein?**

  Zusammen mit der vorhergehenden bilden diese Fragen den Ausgangspunkt der **Diskriminanz-Analyse**. Die Zuordnung wird manchmal auch genauer als **Identifikations-Analyse** bezeichnet.

e
- Für die Bankkunden sind keine Gruppen vorgegeben, sondern man fragt: **Lassen sich die Beobachtungseinheiten sinnvoll in Gruppen einteilen?** Das ist ein recht vage gestelltes Problem, und dem entsprechend bietet das Gebiet der **Cluster-Analyse** eine Vielfalt von Methoden an, die zu Gruppen mit verschiedenen Eigenschaften führen.

f
- Eine sinnvolle Gruppenbildung setzt voraus, dass man die Ähnlichkeit oder die Unähnlichkeit zwischen Beobachtungseinheiten sinnvoll quantifizieren kann. Man fragt also: **Wie sollen die unterschiedlichen Werte der Variablen für zwei Beobachtungseinheiten zu einem Ähnlichkeits- oder Unähnlichkeitsmass zusammengefasst werden?** Es kann auch sinnvoll sein, nach der Ähnlichkeit oder Unähnlichkeit zwischen Variablen zu fragen.

Hat man Unähnlichkeiten zwischen Beobachtungen oder Variablen einmal ermittelt, dann können sie nicht nur zur Bildung von Gruppen benützt werden, sondern sie bilden auch die Grundlage für weitere **distanz-basierte Methoden**, die meistens zur grafischen Darstellung dienen.

g • Zweidimensionale Daten lassen sich in einem Streudiagramm sehr einfach darstellen. Man kann deshalb auch fragen: **Kann man die multivariaten Daten so auf zwei Dimensionen reduzieren, dass die wesentliche Information erhalten bleibt?** Oder weniger restriktiv: **Kann man die Dimension der Daten reduzieren, ohne wesentliche Information zu verlieren?** Im Beispiel der Spektren hat diese Frage eine klare Grundlage: Da sich im Verlaufe der Reaktion lediglich die Konzentrationen der wenigen beteiligten chemischen Substanzen ändern, müssen sich die „wesentlichen Unterschiede" der Spektren mit den Veränderungen der Konzentrationen erfassen lassen. Die gesuchte Dimension wird also höchstens gleich der Anzahl chemischer Substanzen sein.

Eine Methode, solche Dimensionen im Sinne einer beschreibenden Methode zu finden, bildet die **Hauptkomponenten-Analyse**. Modelle, die die Idee formalisieren, laufen unter den Stichworten **Faktor-Analyse** und **lineare Mischungs-Modelle**.

h • Die Situation mit mehreren Gruppen bildet in der univariaten Statistik als einfache Varianzanalyse den Ausgangspunkt für die allgemeinen Modelle der Varianzanalyse und der linearen Regression. In Analogie zu diesen Methodiken fragt man: **Wie hängt die gemeinsame Verteilung mehrerer Zielgrössen von (mehreren) erklärenden Grössen ab?** Die erklärenden Grössen können dabei wieder kontinuierliche Variable oder Faktoren (nominale, kategorielle Variable) sein. Wichtige Ziele solcher Analysen sind wieder Hinweise auf Kausalität oder die Vorhersage der Zielgrössen aus den erklärenden Variablen.

Die **multivariate Regression** und die **multivariate Varianzanalyse** (MANOVA) verallgemeinern die Begriffe und Methoden der „gewöhnlichen", univariaten Regression und Varianzanalyse. Ab und zu wird der Begriff „multivariate Regression" auch für die Regression mit einer einzigen Zielgrösse und mehreren erklärenden Variablen, also für die *multiple* Regression, verwendet.

i Die meisten Fragestellungen der multivariaten Statistik sind also denen der univariaten Statistik gleich, betreffen aber **mehrere Zielgrössen** gleichzeitig. Neu sind die Problemstellungen der Cluster-Analyse und der Dimensionsreduktion. Die Betrachtung der gemeinsamen Verteilung mehrerer, zusammenhängender Zielgrössen bringt aber nicht nur da neue Begriffe und Vorstellungen ins Spiel.

j **Data mining.** In der Geschäftswelt hat sich ein Schlagwort etabliert: „Data mining". Es geht um die Analyse von Daten, die im Geschäftsverkehr in Computer-lesbarer Form anfallen, vor allem um Daten über Kunden, Lager, Bestellungen oder andere Transaktionen. Im Vergleich zu Umfragen oder Versuchen entstehen so in aller Regel grosse bis riesige Datensätze. Was in ihnen enthalten ist, ist aber vorgegeben und kann meistens nicht durch statistische Fragestellungen mitbestimmt werden. Oft müssen die Daten auch aus verschiedenen Datenbanken zuerst aufbereitet und in einem „**data warehouse**" zusammengetragen werden, bevor sie für eine Analyse taugen.

Das Schlagwort „data mining“ will sagen, dass mit geeigneten Werkzeugen in einem solchen „Datenberg“ Bodenschätze zu finden sind. Diese geeigneten Werkzeuge umfassen zunächst statistische Verfahren – neben den verschiedenen Regressionsmethoden auch die Verfahren der multivariaten Statistik. Dazu kommen Algorithmen, die von Informatikern und Ingenieuren erfunden worden sind, und die von Statistikern oft kritisch beurteilt werden.

Die Fragestellungen, die in den Anwendungsgebieten des Data mining im Vordergrund stehen, sind

- die Bestimmung aller Kunden mit bestimmten Merkmalen (Datenbank-Abfrage und -Management),
- die übersichtliche Darstellung der Kundendaten (Beschreibung),
- die Zuordnung der Kunden zu bestimmten Gruppen (Diskriminanz-Analyse),
- die Suche nach einer möglichen Gruppenenteilung (Cluster-Analyse),
- die Vorhersage von Zielgrössen wie der künftig mit den einzelnen Kunden erzielte Umsatz aus bekannten erklärenden Variablen (Regression).

k **Kategorielle und kontinuierliche Daten.** Kategorielle Variable sind solche, die nur endlich viele mögliche Werte haben können, die also die Zugehörigkeit zu einer „Kategorie“ angeben. Kontinuierliche Variable können dagegen prinzipiell alle Zahlen als Werte haben – allenfalls eingeschränkt auf alle nicht-negativen Zahlen oder auf ein Intervall. Sie sind quantitativ zu interpretieren. Für eine dritte Sorte von Daten, die diskreten geordneten Variablen, sind nur „diskrete“ Werte möglich, beispielsweise die ganzen Zahlen. Diese Werte weisen eine sinnvolle Ordnung auf und sind eventuell auch quantitativ zu interpretieren.

In den Beispielen wurden kontinuierliche Daten betrachtet – bis auf die Gruppierungsvariablen, die natürlich kategoriell sind. Wenn von multivariater Statistik die Rede ist, denkt man meistens an kontinuierliche Variable, und das wird auch in diesem Skript so sein.

Eigentlich können alle angeführten Fragen auch für kategorielle Merkmale formuliert werden. Man findet in den Büchern über multivariate Statistik dazu wenig. Eine Ausnahme bilden Fahrmeir et al. (1996). Ein Stichwort in dieser Richtung heisst „loglineare Modelle“, und dies wird oft in Büchern über kategorielle Daten und über verallgemeinerte lineare Modelle behandelt.

Geordnete, diskrete Variable können oft wie kontinuierliche behandelt werden. Spezifische Methoden für solche Variable sind wenig(e) bekannt.

l **Zusammenhang mit anderen Gebieten der Statistik.** Wir werden für die schliessende Statistik Wahrscheinlichkeitsmodelle brauchen. Das grundlegende Modell bildet – wie zu erwarten ist – die Normalverteilung, hier die multivariate Normalverteilung (3.2). Sie wird auch in anderen Gebieten der Statistik gebraucht:

- In den Gebieten der **Zeitreihen** und der **räumlichen Statistik** wird jeweils zunächst nur eine Grösse betrachtet, die zeitlich oder räumlich variiert. Für die gemeinsame Verteilung der Werte für verschiedene Zeitpunkte oder verschiedene Orte sind Normalverteilungen mit spezieller Struktur geeignet. Wenn mehrere

Variable im Zeitverlauf oder in Abhängigkeit vom Ort betrachtet werden, kommt man zusätzlich zum Zusammenhang mit der multivariaten Normalverteilung ins eigentliche Gebiet der multivariaten Statistik.

- **Varianzanalyse-Modelle** mit Zufallseffekten für eine einzige Zielgrösse können ebenfalls durch jeweils ein Modell für die gemeinsame Verteilung der Werte der Zielgrösse für alle Versuchsbedingungen ersetzt werden(siehe 9.2).
- Die multivariate Normalverteilung tritt als **Verteilung von Schätzungen** in allen Gebieten der Statistik auf. Oft werden aus den Beobachtungen Schätzungen für mehrere Parameter ausgerechnet. Fassen wir sie zu einer „mehrdimensionalen Statistik“ zusammen! Bei Annahme einer eindimensionalen Normalverteilung für die Beobachtungen sind lineare mehrdimensionale Statistiken, die aus ihnen ausgerechnet werden, gemeinsam multivariat normalverteilt.

Die multivariate Variante des Zentralen Grenzwertsatzes sagt, dass viele mehrdimensionale Statistiken (auch nichtlineare) näherungsweise multivariat normalverteilt sind bei wachsender Zahl der Beobachtungen – und das gilt nicht nur bei normalverteilten Beobachtungen.

m **Geometrie.** Der Begriff der gemeinsamen Verteilung von zwei Variablen ruft sofort nach grafischen Darstellungen wie dem Streudiagramm (1.2.b). Man identifiziert Beobachtungen mit Punkten in der Ebene. Diese Vorstellung lässt sich auch für drei Variable durchhalten; ab vier wird's schwieriger, und man hofft, dass die wesentlichen Einsichten, die sich aus der Geometrie des zwei- und dreidimensionalen Raumes ergeben, auch für höhere Dimensionen Gültigkeit haben. Formal lassen sich die Begriffe der Geometrie ohne Schwierigkeiten für den m-dimensionalen Raum formulieren. Die Analogie zu den Begriffen der multivariaten Statistik geht weiter und wird das Verständnis unterstützen. Es zeigt sich aber, dass in höheren Dimensionen unangenehme Eigenschaften auftreten, die der zwei- und dreidimensionalen Anschauung widersprechen. Man spricht vom „**Fluch der Dimension**“.

1.4 Software

a **Statistikpakete.** Die meisten Methoden, die in diesem Skript besprochen werden, sind in den umfassenderen Statistik-Software-Paketen verfügbar und über die üblichen „Graphical User Interfaces“ benützbar.

Die klassischen grossen Statistikprogramme heissen SAS und SPSS. Sie umfassen die meisten gängigen Methoden in bewährter Form. Das Programm S-Plus und die freesoftware-Version R sind teilweise weniger ausgereift, beruhen aber auf der Sprache S, auf die wir gleich zurückkommen. Weitere Programme, die generell etwas weniger umfassend, aber auch billiger sind, heissen Systat, Stata, ...

In diesem Skript werden die Funktionen der Sprache S, die die vorgestellten Methoden implementieren, jeweils in einem Abschnitt am Schluss eines Kapitels zusammengestellt.

b **Data mining.** In Anbetracht des finanzkräftigen Marktes bieten die grossen Softwarehäuser auch Module an, die speziell auf die Anwendungen des data mining ausgerichtet sind. Für solche Module ist charakteristisch, dass

- sie gut mit riesigen Datenbanken umgehen können,
- einfach sind in der Benützung,
- neben den wichtigsten klassischen Verfahren einige ad-hoc-Verfahren (besser Algorithmen) mit gutem Marketing enthalten.

Die bekanntesten Pakete sind „Clementine“, das mit SPSS in Verbindung steht, und „SAS Data miner“.

c Da Analysen grösserer empirischer Studien recht komplex werden können, bewährt sich die Benützung einer Kommando-orientierten Statistik-Software oder, anders gesagt, einer **benützerfreundlichen Programmiersprache** mit einer umfassenden Sammlung von Statistik-Funktionen.

Wesentlich an einer solchen Sprache ist, dass sie

- vektor- und matrixorientiert ist und
- die Ergebnisse aller Funktionen in einer Form abspeichert, die ihre weitere Verwendung leicht macht.

d **Statistik-Sprache S.** Eine solche Sprache heisst „S“. Es gibt eine kommerzielle Implementation mit Namen **S-Plus** und eine Free-Software-Version namens „**R**“. Die beiden Varianten sind weitgehend identisch in der Syntax und den zur Verfügung stehenden Grund-Funktionen.

Die Sprache S ist zum Kommunikationsmittel der Statistikerinnen und Statistiker der ganzen Welt geworden. Neue Verfahren werden als „libraries“ von den Forschenden (kostenlos) zur Verfügung gestellt und können ohne Weiteres von allen ohne grosse Einarbeitung benützt werden.

Diese grosse Flexibilität und Reichhaltigkeit zieht es nach sich, dass nicht alle Teile des Systems gleich zuverlässig sind und Fehlermeldungen oft schwierig zu interpretieren sind.

e Neben S erfüllen die Software-Pakete Matlab und Mathematica die vorher erwähnten Voraussetzungen der Matrix-Orientierung und des Speicherns von Resultaten. Die Statistik-Funktionen sind aber viel weniger umfassend und teilweise schlecht konzipiert.

f Als Einführungen in die Sprache S gibt es Bücher von Becker, Chambers and Wilks (1988), Chambers (1998), Chambers and Hastie (1992), Venables and Ripley (1999), Venables and Ripley (2000). Anleitungen in unterschiedlicher Länge findet man auch auf dem Web unter `www.R-project.org`.

1.5 Zu diesem Skript

a **Zielsetzung.** Dieses Skript ist primär für den Nachdiplomkurs in angewandter Statistik gedacht. Im Vordergrund stehen in diesem Skript die Problemstellungen und die sinnvolle Anwendung der ihnen entsprechenden Methoden. Formale mathematische Theorie fehlt weitgehend. Immerhin wird die lineare Algebra eingeführt und benützt, da sie für das Verständnis der Begriffe und Modelle sehr nützlich ist.

Unterlagen zur Vorlesung Multivariate Statistik. Für die Vorlesung Multivariate Statistik stehen die Folien und einzelne vertiefende Abschnitte zur Verfügung. Dieses Skript dient zur ergänzenden Lektüre.

b **Voraussetzungen.** Dieses Skript setzt voraus, dass die Grundbegriffe der univariaten Statistik (der frequentistischen Art) bekannt sind. Grundkenntnisse der Matrix-Rechnung (oder der linearen Algebra) sind von Vorteil; sie werden in Kapitel 2 kurz repetiert.

c **Lehrziele**

- Sie sollen zu angewandten Problemstellungen die geeigneten Methoden identifizieren können.
- Sie sollen die Grundideen der klassischen Methoden, die auf der multivariaten Normalverteilung beruhen, kennen.
- Sie sollen das Potenzial einiger Methoden der explorativen multivariaten Analyse abschätzen können.
- Sie sollen die Modelle, auf denen die Analyse-Verfahren beruhen, verstehen und dadurch die Ziele, Voraussetzungen und Einschränkungen der Methoden richtig beurteilen können.

d **Zusätzliche Lehrziele** für die erweiterte Version der Vorlesung:

- Sie sollen die Wahrscheinlichkeitsmodelle, auf denen die Verfahren beruhen, auch mathematisch begreifen und einfache Herleitungen wiedergeben können.
- Sie sollen die Querverbindungen zur linearen Algebra verstehen und benützen können.

e **Reihenfolge.** Erfahrungsgemäss bildet das Kapitel über die Modelle für angewandt ausgerichtete Lernende eine Hürde, die nach der Repetition und Anwendung der Linearen Algebra demotivierend wirken kann. Deshalb schalte ich in der Vorlesung die Abschnitte über Hauptkomponenten und Biplot (7.1 - 7.2) dazwischen.

f **Notation** In diesem Skript werden Klammern unorthodox, aber überlegt verwendet:

- $\{..\}$ Die geschweiften Klammern werden ausschliesslich für Mengen verwendet.
- $\langle..\rangle$ Diese eckigen Klammern umschliessen Argumente von Funktionen.
- $(..)$ Die gewöhnlichen Klammern zeigen die Priorität der Rechenoperationen (wie in $(a+b)c$).
- $[..]$ Die üblichen eckigen Klammern werden für Vektoren und Matrizen gebraucht.

L Literatur

a Es gibt viele Bücher über multivariate Statistik. Sie sind umfassender und meistens anspruchsvoller als dieses Skript. Die folgenden Bücher werden speziell empfohlen:
Fahrmeir et al. (1996): Deutsch, sehr gut, leider sehr teuer
Mardia, Kent and Bibby (1979): Umfassend. Theorie und Beispiele. Sehr gut.
Rencher (1995): Entspricht im Niveau etwa diesem Skript. Gut verständlich.
Rencher (1998): Nachschlagewerk mit vielen Hinweisen auf Literatur (ohne Cluster-Analyse und Distanzmethoden).
Krzanowski (2000): Angewandt.

Das Buch von Flury (1997) ist sehr anschaulich, angewandt und didaktisch, also sehr gut lesbar. Es umfasst nur einen Teil der Vorlesung und ist dennoch recht dick(... Seiten).

Das Buch von Gnanadesikan (1997) hat die explorative Datenanalyse als Schwerpunkt und hat beim ersten Erscheinen 1977 Marksteine gesetzt.

b **Weitere Lehrbücher** über multivariate Statistik: Anderson (1984), Morrison (1967, 1976), Muirhead (1982), Bilodeau and Brenner (1999), Seber (1984), Srivastava and Carter (1983), Chatfield and Collins (1980), Cooley and Lohnes (1971), Green and Caroll (1976), Harris (1975), Johnson and Wichern (1982, 1988, 1992), Karson (1982), Kendall (1957, 1961), Manly (1986, 1990), Timm (2002)

c **Spezialitäten.** Für spezielle Anwendungsgebiete schreiben Maxwell (1977) (Verhaltenswissenschaften); Tatsuoka (1971) (Psychologie).

Spezielle Themen: Verteilungen: Johnson and Kotz (1972); Modelle für Messfehler: Fuller (1987), Brown (1993); Fehlende Daten: Schafer (1997), Little and Rubin (1987); Grafik: Everitt (1978).

2 Beschreibende Statistik

2.1 Grafische Darstellungen

a **Streudiagramm.** Was wir für unsere Augen festhalten, findet fast immer in zwei Dimensionen seinen Niederschlag, typischerweise auf Papier oder auf dem Bildschirm. Deshalb sind auch grafische Darstellungen normalerweise zweidimensional. Die gemeinsame Verteilung von zwei Variablen lässt sich da auf natürliche Weise im Streudiagramm darstellen. Abbildung 2.1.a zeigt die logarithmierten Längen und Breiten der Sepalblätter der 50 Iris-setosa-Pflanzen im Beispiel der Iris-Blüten.

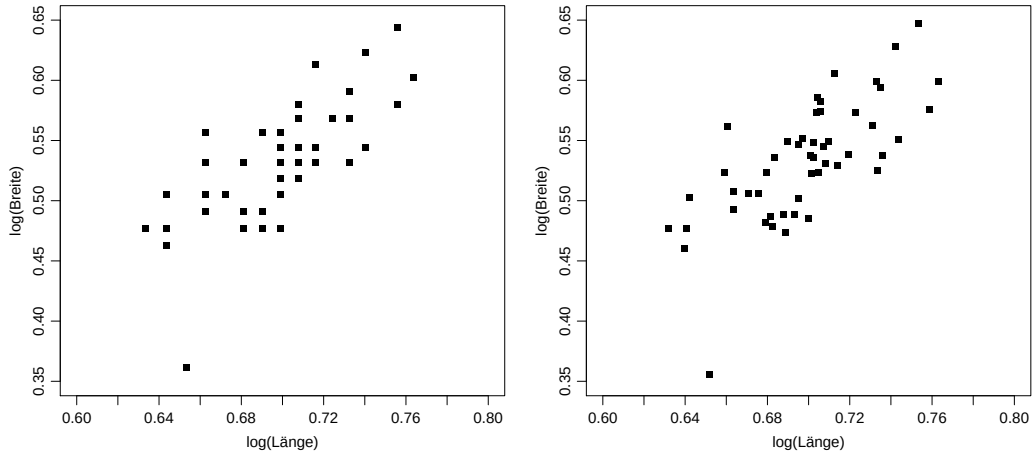

Abbildung 2.1.a: Logarithmierte Längen und Breiten der Sepalblätter von Iris setosa, ohne und mit „Verrütteln“

Benützt man eine übliche Funktion für ein Streudiagramm (linkes Bild in der Abbildung), so wird nicht sichtbar, dass einige Punkte als Folge gerundeter Beobachtungen aufeinander zu liegen kommen. Solche **multiplen Punkte** kann man unter anderem sichtbar machen, indem man die Beobachtungen „verrüttelt“ (englisch *to jitter)*, das heisst, indem man eine kleine, zufällige Zahl addiert. Das ist für das rechte Bild geschehen.

b Mit einigem Aufwand kann man **drei Dimensionen** ausnützen, indem für das zweidimensionale Bild eines dreidimensionalen Objekts der Blickpunkt stetig verändert wird, oder durch andere Formen der Erzeugung eines dreidimensionalen Eindrucks. Auf diese Weise können drei Variable gemeinsam dargestellt werden. Ab vier Variablen muss man sich etwas Anderes einfallen lassen.

c Um mehrere Variable in zwei Dimensionen (ohne Animation, also statisch) darzustellen, gibt es viele Ideen. Etliche sind in Stahel (2002), Kap. 3.6, beschrieben. Es gibt ganze Bücher voller ausgearbeiteter Arten von solchen Darstellungen.

- Drei Bücher, die das Nützliche mit dem Künstlerischen verbinden, stammen von E. Tufte (1983, 1990, 1997).
- W. Cleveland hat die Effektivität von statistischen grafischen Darstellungen untersucht und daraus einen neuen Stil von grafischen Darstellungen entwickelt, der den Namen „trellis“ trägt. Er stellt in seinen Büchern Cleveland (1993) und Cleveland (1994) viele nützliche Anwendungen dieser Darstellungen für die explorative Datenanalyse und die Präsentation von statistischen Resultaten vor.

d **Streudiagramm-Matrix.** Eine grundlegende, einfache Art der Darstellung von mehr als zwei Variablen zeigt die Streudiagramme von allen möglichen Paaren von Variablen, zu einer Matrix von Diagrammen zusammengestellt (Abbildung 2.1.d).

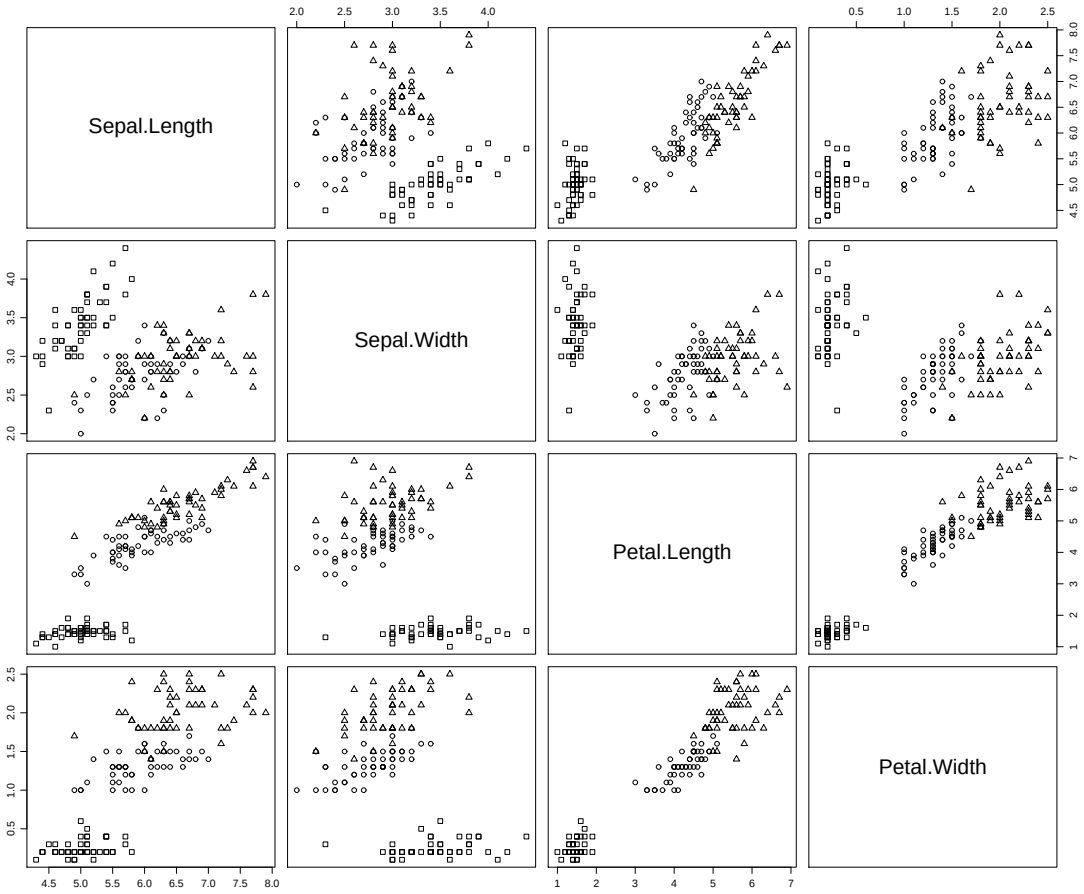

Abbildung 2.1.d: Streudiagramm-Matrix der Iris-Daten. Die verschiedenen Symbole stehen für die drei verschiedenen Iris-Arten.

In jeder Zeile dieser Matrix wird eine Variable in vertikaler Richtung aufgezeichnet, während alle anderen der Reihe nach als horizontale Achse verwendet werden. In jeder Spalte der Matrix erscheint jeweils eine Variable als horizontale Achse und alle anderen als vertikale. Da in der „Diagonalen“ der Anordnung die Variablen gegen sich selber aufgetragen werden müssten, sind diese Feldchen andersartig zu gebrauchen. In Abbildung 2.1.d wird hier der Name der Variablen geschrieben. Oft werden dort auch Histogramme der Variablen gezeigt.

Weil durch das Vertauschen der Achsen nichts wesentlich Neues herauskommt, kann man sich pro Variablenpaar auf eines der beiden möglichen Streudiagramme beschränken. Häufig wird deshalb nur die untere Hälfte der Matrix gezeigt.

e **Coplot.** Die Streudiagramme stellen nur zweidimensionale Verteilungen der Beobachtungen dar. Ein so genannter **Coplot** kann kompliziertere Zusammenhänge zwischen vier Variablen zeigen. Er beruht darauf, dass zunächst zwei Variable klassiert werden, beispielsweise in je 6 überlappende Bereiche mit gleicher Anzahl Beobachtungen, und die Bildfläche dementsprechend in 6×6 Teilflächen aufgespalten wird. In der Teilfläche $[h, \ell]$ werden nur die Daten verwendet, die der entsprechenden Kombination der Klassen der beiden Variablen entspricht. Für diese Daten wird ein Streudiagramm der beiden weiteren Variablen gezeigt.

In Abbildung 2.1.e bilden die Länge und Breite der Sepalblätter der Irisblüten die Achsen der Streudiagramme, während für die Aufteilung die Länge der Petalblätter einerseits und die Pflanzenart andererseits verwendet wurden. Die letztgenannte Variable ist ja bereits eine kategorielle Variable und musste nicht klassiert werden.

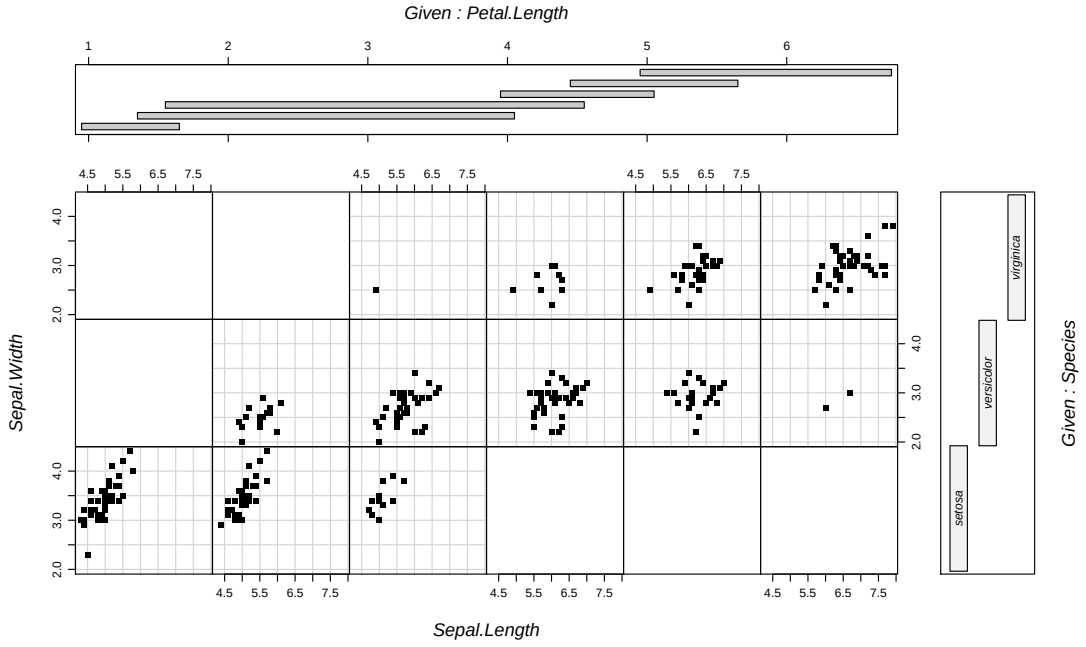

Abbildung 2.1.e: Coplot der Iris-Daten

2.2 Symbole

a **Streudiagramme mit Symbolen.** In einem Streudiagramm von zwei Variablen können weitere Variable mit Symbolen verschiedener Art dargestellt werden.

▷ In Abbildung 2.2.a bilden zwei Variable, die den Boden charakterisieren, die Koordinaten des Streudiagramms, während die Symbole die Anzahl der gefundenen Exemplare von fünf wichtigen Pflanzenarten wiedergeben.

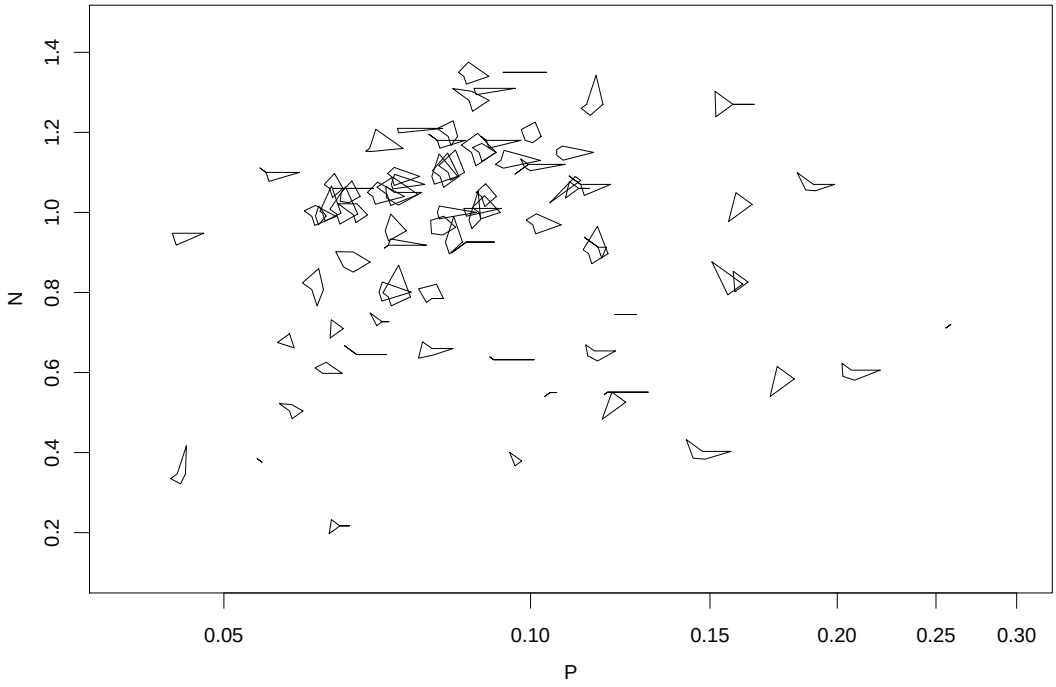

Abbildung 2.2.a: Streudiagramm von zwei Bodenvariablen mit Symbolen, die die Anzahl Exemplare von fünf wichtigen Pflanzenarten zeigen.

◁

b **Symbole.** Die Symbole, die in Abbildung 2.2.a verwendet wurden, sind „Sterne“ (stars). Die Grösse der ersten zusätzlichen Variablen wird vom „Ankerpunkt“ der Beobachtung aus gegen rechts aufgetragen, die zweite Variable in Richtung 72^o (bei insgesamt 5 zusätzlichen Variablen), die dritte in 144^o-Richtung, also nach links oben, etc. Die so erhaltenen Punkte werden verbunden zu einem Fünfeck.

Der Kreativität, solche Symbole zu finden, die die Werte von zusätzlichen Variablen ausdrücken, sind keine Grenzen gesetzt. Wenn nur eine zusätzliche Variable dargestellt werden soll, sind „Bubbles“, also Kreise mit verschiedenen Radien, naheliegend, bei zwei Variablen Rechtecke, und „stars“ können verschieden gestaltet werden.

Eine extravagante Idee bilden die „faces“, künstliche Gesichter, bei denen die „Gesichtszüge“ und Kopfformen den Variablenwerten entsprechen. Die Hoffnung besteht, dass unser Auge Gesichtszüge besonders gut unterscheiden kann.

c **Grafische Elemente.** Daneben gibt es noch weitere grafische Elemente, die Daten kodieren können.

- **Blinken lassen** ist ein äusserst wirksames Mittel für eine zweiwertige Variable, das aber für einen Bericht nicht verwendbar ist.

Weitere Aspekte, die teilweise quantitative Variable und teilweise nur nominale (kategorielle) oder zweiwertige Variable wiedergeben können, sind, in abnehmender Wirksamkeit entsprechend der Beurteilung des Autors,

- Grösse, Farbe, Orientierung (einer Linie, eines Rechtecks, ...), Form (Stern, Kreis, ...),
- Intensität oder Schwarz-Betrag, Farbton, Farbsättigung, Text (Identifikationsnummer, Name einer Gruppe, ...).

2.3 Dynamische Grafik

a **Dynamische grafische Elemente.** In einem Skript oder einem Buch kann man Daten nur auf den zwei Dimensionen des Papiers, statisch grafisch darstellen, und man leistet sich oft nicht einmal Farben. Wenn Daten am Rechner analysiert und auf einem Bildschirm dargestellt werden, hat man einige Möglichkeiten mehr:

- Man kann Darstellungen „bewegen“ und damit beispielsweise einen dreidimensionalen Eindruck erzeugen, indem man Drehungen einer dreidimensionalen „Punktewolke“ veranschaulicht.
- Man kann Interaktionen des Benützers, der Benützerin erlauben: Wenn auf einen Punkt in der Darstellung „geklickt“ wird, kann der Rechner diesen beispielsweise beschriften. Ausserdem kann eine Darstellung über die übliche Art einer „Menü-Steuerung“ beeinflusst werden.

b **Linked Views.** Auf dem Bildschirm können mehrere Darstellungen gleichzeitig gezeigt werden. Die einfachste Variante besteht in der bereits besprochenen Streudiagramm-Matrix. Wenn dann in einem dieser „Panels“ Punkte interaktiv markiert werden, ist es sehr nützlich, wenn sie gleichzeitig in allen Darstellungen hervorgehoben werden. Da die Hervorhebung meist mit Farbe geschieht, nennt man diese Interaktion auch „Anmalen“ *(brushing)*.

Es ist schwierig, auf dem Papier über interaktive und dynamische Grafik etwas Schlaues zu schreiben oder zu zeigen. Probieren Sie es aus!

2.4 Kennzahlen

a Die Verteilung einer einzigen Variablen X wird oft durch Mittelwert $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ und Standardabweichung s beschrieben. Letztere ist die Wurzel aus der (empirischen) Varianz

$$\widehat{\mathrm{var}}\langle X\rangle = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(Die Abkürzung var für die Varianz trägt einen Hut, weil wir bei den Modellen die „theoretische“ Varianz var (ohne Hut) einführen, für die $\widehat{\text{var}}$ eine Schätzung ist.)

b In der multivariaten Statistik betrachten wir mehrere Variable $X^{(j)}$, $j = 1, 2, ..., m$. Grundlegend für die Beschreibung der gemeinsamen Verteilung ist ein Mass für den Zusammenhang zwischen zwei Variablen $X^{(j)}$ und $X^{(k)}$. In Analogie zur Varianz bildet man die **Kovarianz**

$$\widehat{\text{cov}}\langle X^{(j)}, X^{(k)}\rangle = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^{(j)} - \bar{x}^{(j)})(x_i^{(k)} - \bar{x}^{(k)}) \; .$$

Damit das Mass für den Zusammenhang von den Messeinheiten der Variablen unabhängig wird, standardisiert man diese Grösse mittels Division durch die Standardabweichungen und erhält die (Produkt-Momenten oder Pearson-) **Korrelation**

$$\widehat{\rho}\langle X^{(j)}, X^{(k)}\rangle = \frac{\widehat{\text{cov}}\langle X^{(j)}, X^{(k)}\rangle}{\sqrt{\widehat{\text{var}}\langle X^{(j)}\rangle\, \widehat{\text{var}}\langle X^{(k)}\rangle}} \; .$$

Ihr Wert liegt zwischen -1 und 1. Grosse Werte beschreiben einen starken „positiven“, linearen Zusammenhang; grosse Werte von $X^{(1)}$ kommen gehäuft mit grossen Werten von $X^{(2)}$ zusammen, und kleine ebenso. Stark negative Korrelation bedeutet, dass grosse Werte von $X^{(1)}$ häufig mit kleinen Werten von $X^{(2)}$ zusammengehen. Variable, die keinen (linearen) Zusammenhang zeigen, führen zu Korrelations-Werten nahe 0.

c Es gibt natürlich noch andere Arten, Lage, Streuung und Zusammenhang zwischen zwei Variablen zu messen. Mittelwert, Varianz und Kovarianz respektive Pearson-Korrelation sind diejenigen, die zu mathematisch einfachen Resultaten führen.

Ein wichtiger Grund, auch andere Kennzahlen zu betrachten, ist die mangelnde **Robustheit** der besprochenen Masse: Wenn ein Ausreisser auftaucht, also eine Beobachtung, die „schlecht zum Grossteil der Beobachtungen passt“, dann hat er einen grossen Einfluss auf die Kennzahlen, vor allem auf die Varianzen und Kovarianzen. Das ist von der Interpretation her oft unerwünscht und von der schliessenden Statistik her ineffizient. Robuste Methoden sind also wichtig, aber wir wollen sie nicht in diesem Kapitel behandeln.

d **Rangkorrelation.** Eine einfache und bekannte Alternative sei immerhin erwähnt: Wie in der univariaten Statistik kann man einen allzu grossen Einfluss von Ausreissern auf die Korrelation vermeiden, indem man zu Rängen übergeht. Man wendet also auf jede Variable $X^{(j)}$ die **Rangtransformation** an und rechnet dann die Pearson-Korrelation der transformierten Daten aus. Dieses Mass heisst dann **Spearman's Rangkorrelation**. Näheres siehe Stahel (2002), Kap. 3.3.

Allerdings löst diese Idee das Problem der Robustheit nur teilweise. Wenn zwei Variable eng (positiv) korreliert sind, dann kann man eine Beobachtung mit einem hohen Wert der ersten und einem tiefen Wert der zweiten Variablen hinzufügen, und dieser Ausreisser wird einen grossen Einfluss auf die Korrelation haben – auch auf die Rangkorrelation.

2.5 Matrix-Notation

a Multivariate Statistik braucht die Matrizen-Rechnung und Ergebnisse der Linearen Algebra. Ohne diese Werkzeuge zu arbeiten würde der Aushebung einer grossen Baugrube mit einer Schaufel gleichen. Hier werden diese Hilfsmittel schrittweise eingeführt anhand der Anwendung auf die multivariate Statistik. Die Begriffe und Sätze sind in der nötigen Allgemeinheit im Anhang zusammengestellt.

b Die Daten, die in der multivariaten Statistik den Ausgangspunkt bilden, lassen sich in natürlicher Weise als Matrix schreiben,

$$\boldsymbol{x} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \end{bmatrix} .$$

Das Element $x_i^{(j)}$ enthält den Wert der $j$ten Variablen für die $i$te Beobachtung. Tabelle 2.5.b zeigt einen Auszug aus dieser **Datenmatrix** für die Art Iris setosa im Beispiel der Iris-Blüten.

Wir benützen die ersten vier Beobachtungen der ersten zwei Variablen als konkretes **Zahlenbeispiel** für die Veranschaulichung der folgenden Rechnungen.

	Sepal-Blätter		Petal-Blätter	
Nr.	Länge	Breite	Länge	Breite
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
		...		
49	5.3	3.7	1.5	0.2
50	5.0	3.3	1.4	0.2

Tabelle 2.5.b: Daten des Beispiels der Iris-Blüten

c Die einzelnen Variablen entsprechen einer Spalte dieser Matrix, also einem **Vektor**

$$\underline{x}^{(j)} = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix} , \qquad \underline{x}^{(2)} = \begin{bmatrix} 3.5 \\ 3.0 \\ 3.2 \\ 3.1 \end{bmatrix}$$

Die Zeilen fassen die Werte der Variablen für eine Beobachtung i zusammen. Da es

üblich ist, dass Vektoren vertikal geschrieben werden, ist

$$\underline{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \dots \\ x_i^{(m)} \end{bmatrix}, \qquad \underline{x}_3 = \begin{bmatrix} 4.7 \\ 3.2 \end{bmatrix}$$

Wenn wir einen solchen Vektor als eine Zeile geschrieben brauchen, dann bilden wir den „Zeilenvektor“ $\underline{x}_i^T = [x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}]$, $\underline{x}_3^T = [4.7,\ 3.2]$ (T für „transponiert“).

d **Mittelwert.** Mit Hilfe des Vektors

$$\underline{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

lassen sich Summen als „Skalarprodukt“ schreiben: $\sum_i x_i^{(j)} = \underline{1}^T \underline{x}^{(j)}$. Deshalb gilt

$$\bar{x}^{(j)} = \tfrac{1}{n} \underline{1}^T \underline{x}^{(j)} \ .$$

Man kann mit dem Matrixprodukt sogar den Vektor aller Mittelwerte sehr einfach schreiben:

$$\underline{\bar{x}}^T = \tfrac{1}{n} \underline{1}^T \boldsymbol{x} \qquad = \tfrac{1}{4} [1,1,1,1] \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.7 & 3.2 \\ 4.6 & 3.1 \end{bmatrix} = [4.825,\ 3.2] \ .$$

(Der Vektor $\underline{\bar{x}}$ soll wieder ein Spaltenvektor sein. Die rechte Seite der Gleichung gibt aber eine Zeile, ergibt also $\underline{\bar{x}}^T$.)

e **Zentrierte Daten.** Für die Varianzen und Kovarianzen brauchen wir die „zentrierten Variablen“ $x_i^{(j)} - \bar{x}^{(j)}$. Auch diese zentrierten Daten lassen sich sehr kurz schreiben:

$$\boldsymbol{x}_c = \boldsymbol{x} - \underline{1}\, \underline{\bar{x}}^T \ .$$

Dabei muss man beachten, dass der letzte Term ein unübliches Matrix-Produkt ist: Während üblicherweise ein Zeilenvektor mit einem gleich langen Spaltenvektor (zum Skalarprodukt) multipliziert wird, ist hier die Reihenfolge umgekehrt, und es entsteht eine Matrix. Das sieht man am ausgeschriebenen Beispiel:

$$\begin{aligned} \boldsymbol{x}_c &= \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.7 & 3.2 \\ 4.6 & 3.1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [4.825,\ 3.2] = \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.7 & 3.2 \\ 4.6 & 3.1 \end{bmatrix} - \begin{bmatrix} 4.825 & 3.2 \\ 4.825 & 3.2 \\ 4.825 & 3.2 \\ 4.825 & 3.2 \end{bmatrix} \\ &= \begin{bmatrix} 0.275 & 0.3 \\ 0.075 & -0.2 \\ -0.125 & 0 \\ -0.225 & -0.1 \end{bmatrix} . \end{aligned}$$

f **Varianz-Kovarianzmatrix.** Der Ausdruck für die Kovarianz in 2.4.b besteht im Wesentlichen aus einer Summe von Produkten, also einem Skalarprodukt:

$$\begin{aligned}\widehat{\mathrm{cov}}\langle X^{(j)}, X^{(k)}\rangle &= \tfrac{1}{n-1}\, \underline{x}_c^{(j)T}\, \underline{x}_c^{(k)} \\ \widehat{\mathrm{cov}}\langle X^{(1)}, X^{(2)}\rangle &= \tfrac{1}{n-1}\, [0.275, 0.075, -0.125, -0.225] \begin{bmatrix} 0.3 \\ -0.2 \\ 0 \\ -0.1 \end{bmatrix} = 0.03\ .\end{aligned}$$

Die Varianzen erhält man, indem man im allgemeinen Ausdruck $j = k$ setzt. Nun kann man eine Matrix-Gleichung schreiben, die gleich alle Varianzen und Kovarianzen wiedergibt:

$$\begin{aligned}\tfrac{1}{n-1}\, \boldsymbol{x}_c^T \boldsymbol{x}_c &= \begin{bmatrix} \widehat{\mathrm{var}}\langle X^{(1)}\rangle & \widehat{\mathrm{cov}}\langle X^{(1)}, X^{(2)}\rangle & \ldots & \widehat{\mathrm{cov}}\langle X^{(1)}, X^{(m)}\rangle \\ \widehat{\mathrm{cov}}\langle X^{(2)}, X^{(1)}\rangle & \widehat{\mathrm{var}}\langle X^{(2)}\rangle & \ldots & \widehat{\mathrm{cov}}\langle X^{(2)}, X^{(m)}\rangle \\ \vdots & \vdots & \ldots & \vdots \\ \widehat{\mathrm{cov}}\langle X^{(m)}, X^{(1)}\rangle & \widehat{\mathrm{cov}}\langle X^{(m)}, X^{(2)}\rangle & \ldots & \widehat{\mathrm{var}}\langle X^{(m)}\rangle \end{bmatrix} \\ &= \widehat{\mathrm{var}}\langle \underline{X}\rangle = \widehat{\boldsymbol{\Sigma}}\ .\end{aligned}$$

Die Ausdrücke $\widehat{\mathrm{var}}\langle \underline{X}\rangle$ und $\widehat{\boldsymbol{\Sigma}}$ sind Bezeichnungen für diese Matrix, die alle Varianzen und Kovarianzen zusammenfasst und die (empirische) **Varianz-Kovarianz-Matrix** oder kurz **Kovarianz-Matrix** heisst. Sie hat für die ganze multivariate Statistik eine grundlegende Bedeutung. Die zunächst etwas überraschende Konvention, die Varianzen in die „Diagonale“ dieser Matrix zu schreiben, wird sich für die mathematischen Zusammenhänge sehr bewähren. Das zeigt sich schon an der Überlegung, mit der sie hier eingeführt wurde.

Im Beispiel wird

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}} &= \tfrac{1}{n-1} \begin{bmatrix} 0.275 & 0.075 & -0.125 & -0.225 \\ 0.3 & -0.2 & 0 & -0.1 \end{bmatrix} \begin{bmatrix} 0.275 & 0.3 \\ 0.075 & -0.2 \\ -0.125 & 0 \\ -0.225 & -0.1 \end{bmatrix} \\ &= \begin{bmatrix} 0.0492 & 0.0300 \\ 0.0300 & 0.0467 \end{bmatrix}\ .\end{aligned}$$

Es ist nützlich, festzuhalten, dass die Kovarianzmatrix **symmetrisch** ist; es ist ja $\widehat{\mathrm{cov}}\langle X^{(j)}, X^{(k)}\rangle = \widehat{\mathrm{cov}}\langle X^{(k)}, X^{(j)}\rangle$.

> `var`

g **Korrelationsmatrix.** Die Korrelationen kann man nun schreiben als

$$\widehat{\rho}\langle X^{(j)}, X^{(k)}\rangle = \widehat{\rho}_{jk} = \frac{\widehat{\boldsymbol{\Sigma}}_{jk}}{\sqrt{\widehat{\boldsymbol{\Sigma}}_{jj}\, \widehat{\boldsymbol{\Sigma}}_{kk}}}$$

und sie ebenfalls in eine Matrix, die **Korrelationsmatrix**, zusammenfassen.

▷ Tabelle 2.5.g zeigt die untere Hälfte der Korrelationsmatrix. In der „Diagonalen“ werden oft Einsen eingesetzt; die Korrelation jeder Variablen mit sich selber ist =1.

Eine grafische Veranschaulichung dieser Zahlen ist durch die Streudiagramm-Matrix in 2.1.d gegeben. ◁

> `cor`

Sepal.Length	1			
Sepal.Width	0.743	1		
Petal.Length	0.267	0.178	1	
Petal.Width	0.278	0.233	0.332	1
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width

Tabelle 2.5.g: Korrelationen für Iris setosa im Beispiel der Iris-Blüten

h Die einfachste Kovarianzmatrix ist die **Einheitsmatrix**

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

(die Matrix mit Einsen in der Diagonalen und sonst lauter Nullen). Sie besagt, dass die Komponenten $X^{(j)}$ von $\underline{X}$ Varianz 1 haben (und, falls $\bar{x}^{(j)} = 0$ ist, deshalb standardisierte Variable sind), und dass sie unkorreliert sind.

2.6 Projektionen und lineare Transformationen

a **Linearkombinationen von Variablen.** In der multivariaten Statistik spielen Linearkombinationen (übersetzt: gewichtete Summen plus Konstante) der Variablen $X^{(j)}$ eine zentrale Rolle.

▷ Im **Beispiel der Irisblüten** kann es sinnvoll sein, die Fläche und die Form eines Blattes zu erfassen statt der Länge und der Breite. Wenn wir die gemessenen Grössen Länge und Breite zuerst logarithmieren, dann ist die Summe der logarithmierten Werte eine gute Näherung für die logarithmierte Fläche, bis auf eine Konstante a, die für elliptische Blätter $= \log_{10}\langle \pi/4 \rangle = -0.105 \approx -0.1$ beträgt. Die Differenz der beiden logarithmierten Grössen erfasst die Form. ◁

Betrachten wir den allgemeineren Fall einer Linearkombination $Y = a + b_1 X^{(1)} + b_2 X^{(2)}$. Wir bilden also die Werte $y_i = a + b_1 x_i^{(1)} + b_2 x_i^{(2)}$ und schreiben dies mit Vektoren als

$$y_i = a + \underline{b}^T \underline{x}_i \; .$$

Für das Zahlenbeispiel tun wir so, als ob die gegebenen Zahlen bereits logarithmierte Längen und Breiten wären, damit wir die bisherigen Ergebnisse verwenden können. Es wird dann beispielsweise

$$y_3 = -0.1 + [1 \;\; 1] \begin{bmatrix} 4.7 \\ 3.2 \end{bmatrix} = 7.8 \; .$$

b Der **Mittelwert** der y_i ist

$$\begin{aligned}\bar{y} &= a + \tfrac{1}{n}\sum_i \left(b_1 x_i^{(1)} + b_2 x_i^{(2)}\right) = a + \tfrac{1}{n}\left(b_1 \sum_i x_i^{(1)} + b_2 \sum_i x_i^{(2)}\right) \\ &= a + \left(b_1 \tfrac{1}{n}\sum_i x_i^{(1)} + b_2 \tfrac{1}{n}\sum_i x_i^{(2)}\right) = a + b_1\bar{x}^{(1)} + b_2\bar{x}^{(2)} \\ &= a + \underline{b}^T\underline{\bar{x}}\end{aligned}$$

Die Formel $\bar{y} = a + \underline{b}^T\underline{\bar{x}}$ gilt auch noch, wenn $\underline{x}_i$ mehr als zwei Variable umfasst – und der Gewichtungsvektor $\underline{b}$ dementsprechend mehr Komponenten hat.

c Die (empirische) **Varianz** der y_i kann in einer längeren Rechnung mit gleicher Zielrichtung umgewandelt werden:

$$\begin{aligned}\widehat{\mathrm{var}}\langle Y\rangle &= \tfrac{1}{n-1}\sum_i (y_i - \bar{y})^2 = \tfrac{1}{n-1}\sum_i \left(a + b_1 x_i^{(1)} + b_2 x_i^{(2)} - (a + b_1\bar{x}^{(1)} + b_2\bar{x}^{(2)})\right)^2 \\ &= \tfrac{1}{n-1}\sum_i \left(b_1(x_i^{(1)} - \bar{x}^{(1)}) + b_2(x_i^{(2)} - \bar{x}^{(2)})\right)^2 \\ &= \tfrac{1}{n-1}\Big(b_1^2\sum_i (x_i^{(1)} - \bar{x}^{(1)})^2 + 2b_1b_2\sum_i (x_i^{(1)} - \bar{x}^{(1)})(x_i^{(2)} - \bar{x}^{(2)}) \\ &\qquad + b_2^2\sum_i (x_i^{(2)} - \bar{x}^{(2)})^2\Big) \\ &= b_1^2\widehat{\mathrm{var}}\langle X^{(1)}\rangle + 2b_1b_2\widehat{\mathrm{cov}}\langle X^{(1)}, X^{(2)}\rangle + b_2^2\widehat{\mathrm{var}}\langle X^{(2)}\rangle\ .\end{aligned}$$

Dieses Resultat lässt sich nun mit Hilfe der Kovarianzmatrix ebenfalls elegant schreiben:

$$\widehat{\mathrm{var}}\langle Y\rangle = [b_1,\, b_2]\begin{bmatrix} \widehat{\mathrm{var}}\langle X^{(1)}\rangle & \widehat{\mathrm{cov}}\langle X^{(1)}X^{(2)}\rangle \\ \widehat{\mathrm{cov}}\langle X^{(1)}X^{(2)}\rangle & \widehat{\mathrm{var}}\langle X^{(2)}\rangle \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \underline{b}^T\widehat{\mathbf{\Sigma}}\underline{b}\ .$$

Wieder gilt die letzte Formel auch für mehr als zwei Variable $X^{(j)}$.

d **Projektion.** Der Cosinus-Satz aus der Vektor-Geometrie beschreibt den engen Zusammenhang des Skalarproduktes $\underline{b}^T\underline{x}_i$ mit dem Dreieck, das aus den („Orts“-) Vektoren $\underline{b}$ und $\underline{x}_i$ und der Verbindung ihrer Spitzen gebildet wird (Abbildung 2.6.d). Es gilt

$$\underline{b}^T\underline{x}_i = \|\underline{b}\|\,\|\underline{x}_i\|\,\cos\langle\underline{b},\underline{x}_i\rangle\ ,$$

wobei $\cos\langle\underline{b},\underline{x}_i\rangle$ den Cosinus des Winkels zwischen $\underline{b}$ und $\underline{x}_i$ und $\|\underline{c}\|$ die Länge eines Vektors $\underline{c}$ bedeuten. (Diese Länge kann man auch als Wurzel aus dem Skalarprodukt von $\underline{c}$ mit sich selber schreiben, $\|\underline{c}\| = \sqrt{\underline{c}^T\underline{c}}$.)

Man sieht in der Figur auch, dass der Vektor $\widetilde{\underline{x}}_i$, die so genannte „Projektion“ des Vektors $\underline{x}_i$ auf die Richtung von $\underline{b}$, die Länge $\|\underline{x}_i\|\,\cos\langle\underline{b},\underline{x}_i\rangle = \underline{b}^T\underline{x}_i/\|\underline{b}\|$ hat. Wenn der Vektor $\underline{b}$ so gewählt wird, dass er die Länge $\|\underline{b}\| = 1$ hat, dann bilden also die $y_i = \underline{b}^T\underline{x}_i$ die Längen der Projektionen der Beobachtungsvektoren $\underline{x}_i$ auf die Richtung $\underline{b}$. Wenn nur zwei Variable vorhanden sind ($m = 2$), dann können solche **Einheitsvektoren** oder **Richtungsvektoren** $\underline{b}$ durch den Winkel β charakterisiert werden, den sie mit der horizontalen Achse einschliessen; sie haben die Form

$$\underline{b} = \begin{bmatrix} \cos\langle\beta\rangle \\ \sin\langle\beta\rangle \end{bmatrix}\ .$$

Projektionen spielen in der multivariaten Statistik eine wichtige Rolle; wir kommen im nächsten Abschnitt darauf zurück.

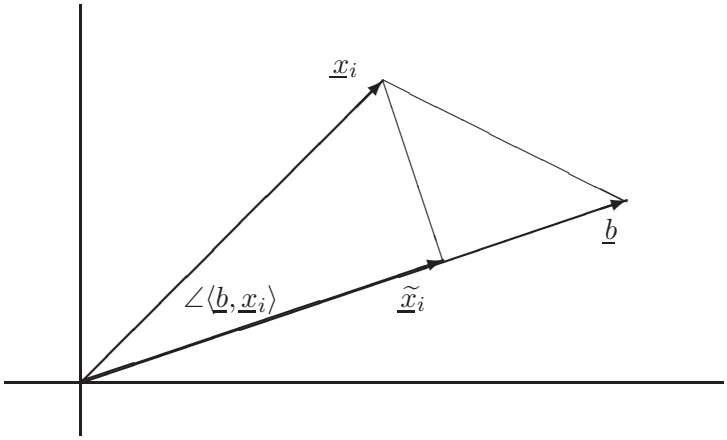

Abbildung 2.6.d: Veranschaulichung der Bedeutung von Skalarprodukt, Zwischenwinkel und Projektion

e **Lineare Transformation.** Wie oben erwähnt, könnten die Blütenblätter der Irispflanzen statt durch die (logarithmierte) Länge $X^{(1)}$ und Breite $X^{(2)}$ auch durch die (logarithmierte) Fläche $Y^{(1)} = a + X^{(1)} + X^{(2)}$ und die Form-Variable $Y^{(2)} = X^{(2)} - X^{(1)}$ erfasst werden. Die neuen Grössen $\underline{Y} = [Y^{(1)}, Y^{(2)}]^T$ entstehen aus den alten durch eine **lineare Transformation**. Wir sprechen allgemein von einer linearen Transformation, wenn

$$\underline{y} = \underline{a} + \boldsymbol{B}\underline{x}$$

gilt. Im Beispiel ist

$$\underline{y} = \begin{bmatrix} -0.1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \underline{x} \,.$$

f **Wichtige Beispiele linearer Transformationen.** Transformationen lassen sich **geometrisch als Abbildungen** interpretieren und veranschaulichen – besonders gut in zwei Dimensionen. Punkte $\underline{x}_i$ in der Ebene gehen in Bildpunkte $\underline{y}_i$ über.

Eine besonders einfache Abbildung ist die **Streckung** vom Nullpunkt aus. Man erhält den Bildpunkt $\underline{y}_i$, indem man $\underline{x}_i$ komponentenweise mit dem Streckungsfaktor b multipliziert. Das ist eine lineare Transformation mit $\underline{a} = \underline{0}$ und

$$\boldsymbol{B} = \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix} \,.$$

Matrizen, die nur auf der Diagonalen von 0 verschiedene Zahlen enthalten, nennt man übrigens **Diagonal-Matrizen** und schreibt diese Matrix auch als $\mathrm{diag}\langle b, b \rangle$.

Eine **Spiegelung** an der X-Achse erhält man mit der Diagonalmatrix $\boldsymbol{B} = \mathrm{diag}\langle 1, -1 \rangle$, eine Punktspiegelung am Nullpunkt mit $\boldsymbol{B} = \mathrm{diag}\langle -1, -1 \rangle$. Mit allgemeinen Diagonalmatrizen kann man Streckungen und Spiegelungen erhalten, die in die beiden Achsenrichtungen verschieden wirken; $\boldsymbol{B} = \mathrm{diag}\langle b, 1 \rangle$ streckt die Daten nur in x-Richtung.

Eine etwas unüblichere Abbildung ist die Scherung. Eine Scherung in horizontaler Richtung entsteht mit der Transformationsmatrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} .$$

Abbildung 2.6.f veranschaulicht diese Abbildungen.

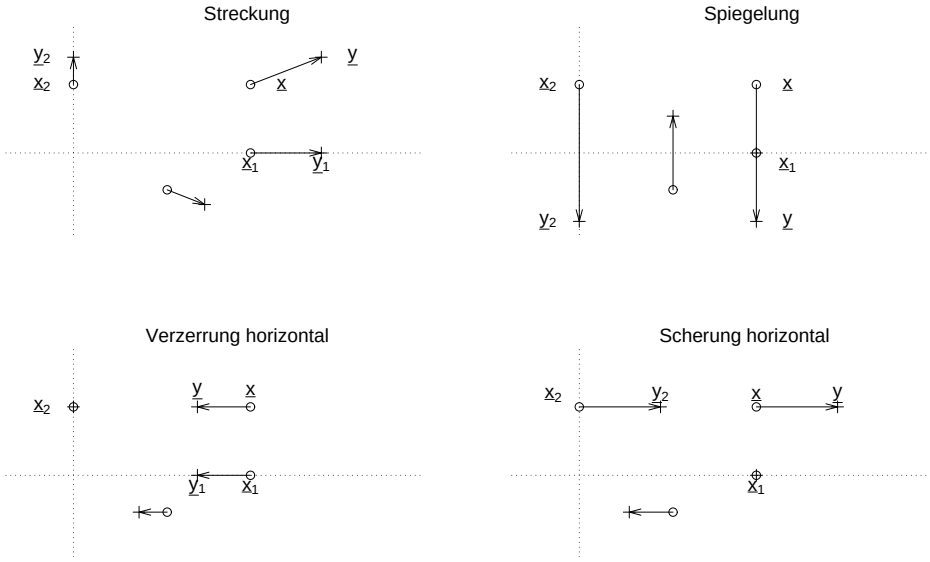

Abbildung 2.6.f: Vier lineare Abbildungen: Streckung mit $b = 1.4$, Spiegelung an der horizontalen Achse, Verzerrung in horizontaler Richtung mit $b = 0.7$ und Scherung in horizontaler Richtung mit $a = 0.3$

g **Drehung.** Eine wichtige Art von Transformationen bilden die Drehungen. Die entsprechenden Matrizen haben die Form

$$\boldsymbol{B} = \begin{bmatrix} \cos\langle\beta\rangle & -\sin\langle\beta\rangle \\ \sin\langle\beta\rangle & \cos\langle\beta\rangle \end{bmatrix} \quad \text{resp.} \quad = \begin{bmatrix} -\cos\langle\beta\rangle & \sin\langle\beta\rangle \\ \sin\langle\beta\rangle & \cos\langle\beta\rangle \end{bmatrix}$$

Wie kann man das sehen? Abbildung 2.6.g zeigt, dass für einen Punkt auf der horizontalen Achse, mit der Form $\underline{x}_1 = [x^{(1)} \;\; 0]^T$ der Bildpunkt

$$\underline{y}_1 = \begin{bmatrix} x^{(1)} \cdot \cos\langle\beta\rangle \\ x^{(1)} \cdot \sin\langle\beta\rangle \end{bmatrix}$$

herauskommt, und das ist wirklich gleich $\boldsymbol{B}\underline{x}_1$. Ebenso erhält man für $\underline{x}_2 = [0, \; x^{(2)}]^T$ den Bildpunkt $\underline{y}_2 = [-x^{(2)} \cdot \sin\langle\beta\rangle \; , \; x^{(2)} \cdot \cos\langle\beta\rangle]^T = \boldsymbol{B}\underline{x}_2$. Den allgemeinen Punkt mit dem Vektor $[x^{(1)}, \; x^{(2)}]^T$ erhält man, indem man $\underline{x}_1$ und $\underline{x}_2$ zusammenzählt. Den entsprechenden Bildpunkt erhält man also, indem man $\underline{y}_1$ und $\underline{y}_2$ zusammenzählt (wie

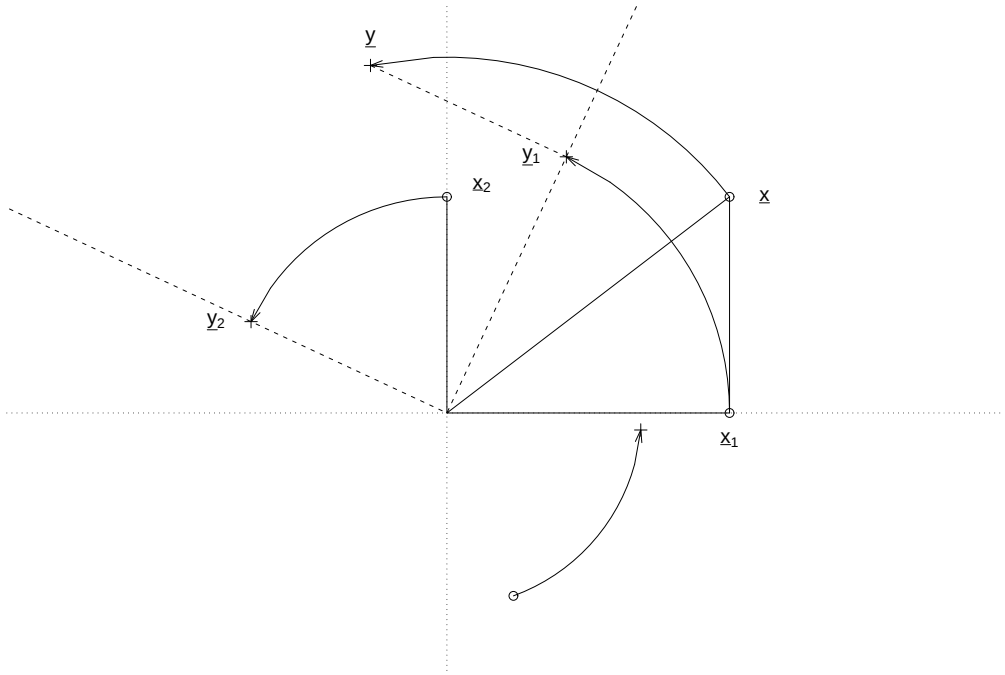

Abbildung 2.6.g: Drehung um den Winkel $\beta = 65^o$

der Name „linear“ der Abbildung ausdrückt und man leicht nachrechnet). Das führt zum Ergebnis $\underline{y} = \boldsymbol{B}\underline{x}$ mit der oben aufgeschriebenen Transformtionsmatrix $\boldsymbol{B}$.

In Abbildung 2.6.g (ii) werden die Beobachtungen des Zahlenbeispiels (2.5.b) so gedreht, dass der Mittelwertsvektor auf die horizontale Achse zu liegen kommt. Der Drehwinkel wird $-33.6°$.

h **Zwei Transformationen.** Wenn man zwei lineare Transformationen hinter einander ausführt, dann kann man die zusammengesetzte Abbildung auch als eine einzige lineare Transformation schreiben, da

$$\begin{aligned} \underline{y} &= \underline{a} + \boldsymbol{B}\underline{x}\,, \quad \widetilde{\underline{y}} = \widetilde{\underline{a}} + \widetilde{\boldsymbol{B}}\underline{y} \\ \widetilde{\underline{y}} &= \widetilde{\underline{a}} + \widetilde{\boldsymbol{B}} \cdot \big(\underline{a} + \boldsymbol{B}\underline{x}\big) = \widetilde{\underline{a}} + \widetilde{\boldsymbol{B}} \cdot \underline{a} + \widetilde{\boldsymbol{B}}\boldsymbol{B}\underline{x} \end{aligned}$$

gilt und das Ergebnis wieder die Form „Vektor plus Matrix mal $\underline{x}$ hat.

i **Mittelwert und Varianz der transformierten Daten.** Zurück zur Statistik! Der Mittelwert der transformierten Werte $\underline{y}_i = \underline{a} + \boldsymbol{B}\underline{x}_i$ ist ja die „Zusammenstellung“ der Mittelwerte der einzelnen Komponenten, für die das Resultat aus 2.6.b gilt. Daraus wird $\overline{\underline{y}} = \underline{a} + \boldsymbol{B}\overline{\underline{x}}$.

Die Varianzen der $Y^{(k)}$ erhält man ebenfalls aus dem vorhergehenden Fall (2.6.c). Und die Kovarianz zwischen der Fläche $Y^{(1)}$ und der Form $Y^{(2)}$? Eine analoge Rechnung ergibt, wenn $\underline{b}_1^T$ die erste Zeile der Transformationsmatrix $\boldsymbol{B}$ und $\underline{b}_2^T$ die zweite ist, $\widehat{\mathrm{cov}}\langle Y^{(1)}, Y^{(2)}\rangle = \underline{b}_1^T \boldsymbol{\Sigma} \underline{b}_2$. Stellt man die Varianzen und die Kovarianz zusammen, so

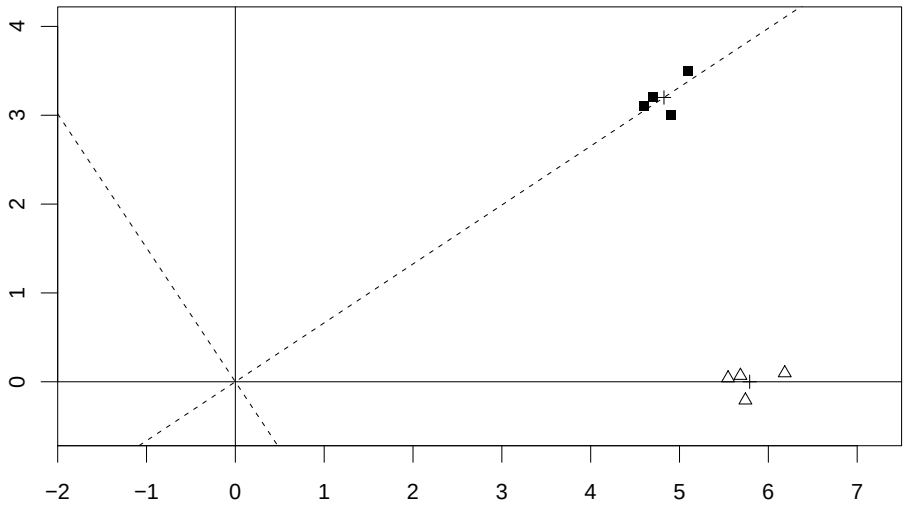

Abbildung 2.6.g (ii): Rotation der vier Beobachtungen des Zahlenbeispiels (2.5.b) um einen Winkel von $-33.6°$. Dem entspricht die Drehung des Koordinatenkreuzes um $+33.6°$. Die gedrehten Achsen sind gestrichelt eingezeichnet.

erhält man

$$\begin{aligned}\widehat{\mathrm{var}}\langle \underline{Y}\rangle &= \boldsymbol{B}\widehat{\mathrm{var}}\langle \underline{X}\rangle \boldsymbol{B}^T \\ &= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 0.0492 & 0.0300 \\ 0.0300 & 0.0467 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.0208 & 0.0128 \\ 0.0128 & 0.0751 \end{bmatrix}.\end{aligned}$$

Dabei ist $\boldsymbol{B}^T$ die transponierte Matrix $\boldsymbol{B}$.

j Die Herleitung dieser Resultate gelingt noch eleganter, wenn wir auf die Datenmatrix $\boldsymbol{x}$ zurückgreifen. Die Matrix $\boldsymbol{y}$ enthält die transformierten Beobachtungsvektoren $\underline{y}_i$ als Zeilen $\underline{y}_i^T$. Es gilt $\underline{y}_i^T = \underline{a}^T + \underline{x}_i^T \boldsymbol{B}^T$ (da sich beim Transponieren die Reihenfolge der Faktoren ändert, siehe 6).Deshalb wird

$$\boldsymbol{y} = \underline{1}\,\underline{a}^T + \boldsymbol{x}\,\boldsymbol{B}^T = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}[a,\ 0] + \begin{bmatrix} 5.1 & 3.5 \\ 4.9 & 3.0 \\ 4.7 & 3.2 \\ 4.6 & 3.1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Der Mittelwertsvektor ist gemäss 2.5.d gleich

$$\begin{aligned}\underline{\overline{y}}^T &= \tfrac{1}{n}\underline{1}^T\boldsymbol{y} = \tfrac{1}{n}\underline{1}^T\underline{1}\,\underline{a}^T + \tfrac{1}{n}\underline{1}^T\boldsymbol{x}\boldsymbol{B}^T = \tfrac{1}{n}\,n\,\underline{a}^T + \underline{\overline{x}}^T\boldsymbol{B}^T \\ \underline{\overline{y}} &= \underline{a} + \boldsymbol{B}\,\underline{\overline{x}} = \begin{bmatrix} -0.1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 4.825 \\ 3.2 \end{bmatrix} = \begin{bmatrix} 7.925 \\ -1.625 \end{bmatrix}.\end{aligned}$$

Die zentrierten, transformierten Daten werden

$$\boldsymbol{y}_c = \boldsymbol{y} - \underline{1}\,\underline{\overline{y}}^T = \underline{1}\,\underline{a}^T + \boldsymbol{x}\,\boldsymbol{B}^T - \underline{1}\,(\underline{a}^T + \underline{\overline{x}}^T\boldsymbol{B}^T) = (\boldsymbol{x} - \underline{1}\,\underline{\overline{x}}^T)\,\boldsymbol{B}^T = \boldsymbol{x}_c\,\boldsymbol{B}^T.$$

Für die Kovarianzmatrix erhalten wir schliesslich nach 2.5.f

$$\widehat{\mathrm{var}}\langle \underline{Y} \rangle = \tfrac{1}{n-1}\, \boldsymbol{y}_c^T \boldsymbol{y}_c = \tfrac{1}{n-1}\, \boldsymbol{B}\, \boldsymbol{x}_c^T \boldsymbol{x}_c \boldsymbol{B}^T = \boldsymbol{B}\, \widehat{\mathrm{var}}\langle \underline{X} \rangle\, \boldsymbol{B}^T \ .$$

Diese Formeln gelten auch dann, wenn $\boldsymbol{B}$ nicht quadratisch ist, also wenn aus den m X-Variablen nicht gleich viele Y-Variable erzeugt werden. Setzt man für $\boldsymbol{B}$ eine einzeilige Matrix $\underline{b}^T$ ein, dann erhält man die Formeln für eine Linearkombination (2.6.b und 2.6.c) als Spezialfall.

k **Die Identität.** Eine Frage, die zunächst eher überflüssig erscheint: Wie sieht die Transformation aus, die gar nichts verändert, für die also $\underline{y}_i = \underline{x}_i$ ist für alle möglichen $\underline{x}_i$? Blöde Frage? Nun, Sie wissen, dass für ein Verständnis der Addition von Zahlen die Null wichtig ist, die jede Zahl unverändert lässt, ebenso die Eins für die Multiplikation.

Bei den linearen Transformationen kombinieren wir eine Addition von Vektoren und eine Multiplikation eines Vektors mit einer Matrix (wobei die Matrix links vom Vektor steht). Für die Addition der Vektoren lässt der Nullvektor $\underline{0}$, der aus lauter Nullen besteht, den anderen Vektor unverändert; für die Multiplikation brauchen wir die **Einheitsmatrix** $\boldsymbol{I}$:

$$\underline{X} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(m)} \end{bmatrix} = \underline{0} + \boldsymbol{I}\,\underline{X} \ .$$

l **Rücktransformation, inverse Matrix.** Im Fall der Transformation von (logarithmierter) Länge und Breite zu (logarithmierter) Fläche und Form kann man aus den neuen Variablen die alten wieder zurückgewinnen mit der Umkehr-Transformation

$$\underline{X} = \boldsymbol{B}^{-1}(\underline{Y} - \underline{a}) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \left(\begin{bmatrix} Y^{(1)} \\ Y^{(2)} \end{bmatrix} - \begin{bmatrix} a \\ 0 \end{bmatrix} \right) \ ,$$

für das Beispiel, wie man leicht nachrechnen kann. Die Matrix $\boldsymbol{B}^{-1}$ ist die **Inverse** zur Matrix $\boldsymbol{B}$. Es gilt $\boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}$, da die Transformation mit der Matrix $\boldsymbol{B}$, gefolgt von der Umkehr-Transformation, die Identität liefert. Eine solche Matrix kann es nur zu quadratischen Matrizen geben, aber nicht alle quadratischen Matrizen haben eine Inverse. Diejenigen, für die es eine Inverse gibt, nennt man **reguläre** oder **invertierbare** Matrizen, die andern **singulär** (vgl. 2.A.0.j im Anhang).

m **Standardisierung.** In der univariaten Statistik war es oft nützlich, eine Stichprobe X zu standardisieren, indem man den Mittelwert abzog und durch die Standardabweichung dividierte; man bildete $z_i = (x_i - \bar{x})/\widehat{\sigma}$. Aus den Daten $\boldsymbol{x}$ mit Mittelwertsvektor $\underline{\overline{x}}$ und empirischer Kovarianzmatrix $\widehat{\boldsymbol{\Sigma}}$ wollen wir jetzt durch lineare Transformation zu einer Datenmatrix $\boldsymbol{z}$ mit $\underline{\overline{z}} = \underline{0}$ und $\widehat{\mathrm{var}}\langle \underline{Z} \rangle = \boldsymbol{I}$ gelangen.

Es ist leicht, $\underline{\overline{z}} = \underline{0}$ zu erreichen: Man braucht nur den Mittelwertsvektor $\underline{\overline{x}}$ von jedem $\underline{x}_i$ abzuziehen.

Für die zweite Forderung brauchen wir ein Resultat der linearen Algebra: Die Kovarianz-Matrix $\widehat{\boldsymbol{\Sigma}}$ ist symmetrisch und sogenannt **positiv semidefinit**, das heisst, es gilt für

jeden beliebigen Vektor $\underline{b}$ (der nicht $= \underline{0}$ ist),

$$\underline{b}^T \widehat{\boldsymbol{\Sigma}} \underline{b} \geq 0 \; .$$

Für jede solche Matrix lässt sich gemäss einem Resultat der linearen Algebra eine Matrix $\boldsymbol{B}$ finden, so dass $\boldsymbol{B}\,\boldsymbol{B}^T = \widehat{\boldsymbol{\Sigma}}$ – es gibt sogar unendlich viele. Die so genannte **Cholesky-Zerlegung** liefert eine davon in Form einer Dreiecksmatrix – im Beispiel

$$\boldsymbol{B} = \begin{bmatrix} 0.222 & 0 \\ 0.135 & 0.168 \end{bmatrix} , \quad \begin{bmatrix} 0.222 & 0 \\ 0.135 & 0.168 \end{bmatrix} \begin{bmatrix} 0.222 & 0.135 \\ 0 & 0.168 \end{bmatrix} = \begin{bmatrix} 0.0492 & 0.0300 \\ 0.0300 & 0.0467 \end{bmatrix} .$$

Jetzt setzen wir als Transformation $\underline{z} = \boldsymbol{C}(\underline{x} - \underline{\widehat{\mu}})$ an, mit $\boldsymbol{C} = \boldsymbol{B}^{-1}$, und prüfen nach! Es wird

$$\begin{aligned} \underline{z}_i &= \boldsymbol{C}(\underline{x}_i - \overline{\underline{x}}) = -\boldsymbol{C}\overline{\underline{x}} + \boldsymbol{C}\underline{x}_i \\ \overline{\underline{z}} &= \boldsymbol{C}(\overline{\underline{x}} - \overline{\underline{x}}) = \underline{0} \\ \widehat{\mathrm{var}}\langle \underline{Z} \rangle &= \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^T = \boldsymbol{C}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{C}^T = \boldsymbol{C}\boldsymbol{C}^{-1}(\boldsymbol{C}^{-1})^T\boldsymbol{C}^T = \boldsymbol{I} \; . \end{aligned}$$

Das Ziel ist also erreicht! (Allerdings wurde stillschweigend vorausgesetzt, dass $\boldsymbol{C}$ invertierbar, also nicht singulär sei, was gleichbedeutend ist mit der Forderung, $\widehat{\boldsymbol{\Sigma}}$ dürfe nicht singulär sein.)

Im Beispiel ist

$$\boldsymbol{C} = \begin{bmatrix} 4.51 & 0 \\ -3.62 & 5.94 \end{bmatrix} , \qquad \boldsymbol{z} = \boldsymbol{x}_c \boldsymbol{C}^T = \begin{bmatrix} 1.240 & 0.785 \\ 0.338 & -1.458 \\ -0.564 & 0.452 \\ -1.014 & 0.221 \end{bmatrix}$$

* Wie erwähnt liefert die Cholesky-Zerlegung nicht die einzige Matrix $\boldsymbol{B}$, für die $\boldsymbol{B}\,\boldsymbol{B}^T = \widehat{\boldsymbol{\Sigma}}$ gilt. Wir kommen auf andere Lösungen gleich anschliessend zurück.

n **Rücktransformation einer Drehung, orthogonale Matrizen.** Um eine Drehung um den Winkel β rückgängig zu machen, muss man nochmals drehen, diesmal um $-\beta$. Das führt zur Matrix

$$\begin{bmatrix} \cos\langle -\beta \rangle & -\sin\langle -\beta \rangle \\ \sin\langle -\beta \rangle & \cos\langle -\beta \rangle \end{bmatrix} = \begin{bmatrix} \cos\langle \beta \rangle & \sin\langle \beta \rangle \\ -\sin\langle \beta \rangle & \cos\langle \beta \rangle \end{bmatrix} = \boldsymbol{B}^T \; .$$

Also ist die transponierte Matrix $\boldsymbol{B}^T$ zugleich die Inverse, $\boldsymbol{B}^{-1}$, und deshalb

$$\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I} \; .$$

Das gilt auch für Spiegelungen. Matrizen mit dieser Eigenschaft heissen **orthogonale** Matrizen, und die zugehörigen Transformationen werden ebenfalls „orthogonal“ genannt.

Die zentrale Eigenschaft von orthogonalen Transformationen besteht darin, dass sie die Abstände zwischen Punkten nicht verändern und damit alle „Formen“ oder Muster von Punkten erhalten bleiben. Es lässt sich nämlich leicht nachrechnen, dass die

(quadrierten) Längen von Vektoren bei orthogonalen Transformationen unverändert bleiben:

$$\|\underline{y}_i\|^2 = \underline{y}_i^T \underline{y}_i = \underline{x}_i^T \boldsymbol{B}^T \boldsymbol{B} \underline{x}_i = \underline{x}_i^T \boldsymbol{I} \underline{x}_i = \underline{x}_i^T \underline{x}_i = \|\underline{x}_i\|^2 \ .$$

Anmerkung. Wir haben hier stillschweigend angenommen, dass keine Verschiebung erfolgt, also $\underline{a} = \underline{0}$ ist, und die Drehung erfolgte um den Nullpunkt. Man kann auch allgemeinere Drehungen, mit Verschiebung, betrachten; es bleiben dann nur die Längen von Differenzen $\underline{x}_i - \underline{x}_h$ erhalten.

o* Kommen wir auf die Frage nach weiteren Lösungen von $\boldsymbol{B}\boldsymbol{B}^T = \widehat{\boldsymbol{\Sigma}}$ zurück! Wenn $\boldsymbol{B}_c$ eine Lösung ist, beispielsweise die Cholesky-Wurzel, dann ist auch $\boldsymbol{B}_c\boldsymbol{B}_o$ mit jeder orthogonalen Matrix $\boldsymbol{B}_o$ eine Lösung, denn $\boldsymbol{B}_c\boldsymbol{B}_o(\boldsymbol{B}_c\boldsymbol{B}_o)^T = \boldsymbol{B}_c\boldsymbol{B}_o\boldsymbol{B}_o^T\boldsymbol{B}_c^T = \boldsymbol{B}_c\boldsymbol{I}\boldsymbol{B}_c^T = \boldsymbol{B}_c\boldsymbol{B}_c^T = \widehat{\boldsymbol{\Sigma}}$. Man kann auch das Umgekehrte ähnlich einfach beweisen: Zwei Lösungen unterscheiden sich immer um einen orthogonalen „Faktor“ $\boldsymbol{B}_o$.

Man kann sich leicht anschaulich klar machen, was dahinter steht: Wenn man bereits standardisierte Daten mit einer orthogonalen Matrix transformiert, also dreht oder spiegelt, dann bleiben sie standardisiert. Setzt man also eine „Standardisierungs-Transformation“ und eine orthogonale zusammen, so hat man wieder eine Standardisierungs-Transformation.

p **Basis-Transformation.** Statt an die Drehung aller Punkte um den Winkel β zu denken, können wir auch von einer Drehung des Koordinatenkreuzes um $-\beta$ reden – beide Vorstellungen führen zu den gleichen „neuen Koordinaten“ $\underline{y}_i$. In Abbildung 2.6.g (i) ist dieses neue Achsenkreuz gestrichelt eingezeichnet. Auch in mehr als zwei Dimensionen sind orthogonale Transformationen äquivalent zu Änderungen des Koordinatensystems. Die Längen von Vektoren bleiben bei dieser Änderung erhalten.

2.7 Projection Pursuit

a **Grundidee.** Explorative multivariate Statistik soll interessante Strukturen in den Daten finden. Wenn sich diese nicht in einer Streudiagramm-Matrix zeigen, kann man hoffen, dass sie bei einer geeigneten Veränderung des Koordinatensystems sichtbar werden. Wir suchen also nach „Richtungen im Raum“, die interessante Strukuren zeigen.

b **Manuelle Suche.** Dynamische grafische Programme erlauben es, die beiden Achsen, die für ein Streudiagramm verwendet werden, stetig und nach Belieben mit Hilfe irgendwelcher Bewegungen eines Joystick, von Cursor-Pfeilen oder ähnlichen Eingaben zu verändern (vergleiche 2.3.a). Das Auge kann dann nach Strukturen suchen.

Was auf dem Bildschirm jeweils dargestellt wird, ist ein Streudiagramm von zwei Projektionen $\underline{y}^{(1)} = \boldsymbol{x}\underline{b}_1$ und $\underline{y}^{(2)} = \boldsymbol{x}\underline{b}_2$ der Daten (2.6.d) mit zwei aufeinander senkrecht stehenden Richtungsvektoren $\underline{b}_1$ und $\underline{b}_2$. Anders ausgedrückt verwenden wir die ersten beiden Koordinaten der Punkte in einem neuen Koordinatensystem mit orthogonaler Transformationsmatrix (2.6.p).

c Interessante Projektionen können auch vom Rechner selbst mit numerischer Optimierung gesucht werden, wenn man ein quantitatives Mass für die „Interessantheit“ einer Projektion, einen **Projektionsindex** Q, angeben kann. Dieses Mass wird von der Anwendung abhängen.

d* Welche Eigenschaften soll ein Projektionsindex haben? Als interessant gelten meist Abweichungen von der Normalverteilung. Da eine solche Abweichung gleich bleibt, wenn man die Skala (Lage und Streuung) der Daten verändert, muss ein solches Q „**invariant**“ gegen Verschiebung und Streckung sein:

$$Q\langle a + b\underline{y}\rangle = Q\langle \underline{y}\rangle$$

Man definiert deshalb Q für standardisierte Grössen und verwendet $Q\langle \underline{y}\rangle = Q_0\langle(\underline{y} - \bar{y})/s_y\rangle$ wobei $s_y^2 = \widehat{\mathrm{var}}\langle Y\rangle$ die empirische Varianz von $\underline{y}$ ist. Zur Berechnung standardisiert man $\boldsymbol{X}$ (siehe 2.6.m). Dann wird $s_y = 1$ für alle Richtungsvektoren $\underline{b}$ mit Länge $|\underline{v}| = 1$.

e **Projektions-Indices.** Zwei Klassen sind üblich:

1. Funktionale für Dichten. Man schätzt die Dichte zuerst mit einem d-dimensionalen Dichteschätzer.
2. Höhere empirische Momente (drittes und viertes)

2.S S-Funktionen

a **Grafische Funktionen:** Die wichtigsten grafischen Funktionen zur Darstellung von mehreren Variablen sind `pairs, symbols, coplot`.

b **Funktion `pairs`.** erstellt eine Streudiagramm-Matrix der Daten. Abbildung 2.1.a entsteht durch

```
> pairs(iris[,1:4], pch=c(0,1,2)[iris[,5]])
```

Mit den Argumenten `diag.panel, lower.panel` und `upper.panel` lassen sich die Diagonale und die untere bzw. obere Dreiecksmatrix nach eigenem Gutdünken gestalten. So kann man sich z.B. in die obere Dreiecksmatrix, die ja nur die gespiegelten Scatterplots der unteren enthält, die paarweisen Korrelationen der Variablen schreiben.

```
> data(iris)
> pairs(iris, lower.panel=panel.smooth,
    upper.panel= function(x,y)
      text(mean(range(x)),mean(range(y)), round(cor(x,y),3) ) )
```

Für das untere Dreieck wurde hier die nützliche verfügbare Funktion `panel.smooth` benützt. (Im Beispiel ist das nur bedingt sinnvoll und für die unterste Zeile sogar sinnlos.)
`example(pairs)` zeigt, wie man für die Diagonale Histogramme erhält.

c **Funktion `symbols`.** Streudiagramm mit Symbolen.
Zusätzlich zu den zwei Variablen, die im Streudiagramm als horizontale und vertikale Koordinaten dargestellt werden, werden weitere Variablen durch die verwendeten Symbole auf vielfältige Art wiedergegeben.

```
Usage:
  symbols(x, y = NULL, circles, squares, rectangles, stars,
          thermometers, boxplots, inches = TRUE, add = FALSE,
          fg = 1, bg = NA, xlab = NULL, ylab = NULL, main = NULL,
          xlim = NULL, ylim = NULL, ...)
```

Von den Argumenten `circles, squares, rectangles, stars, thermometers` und `boxplots` wird jeweils nur eines verwendet.

- Mit `circles=z` werden Kreise mit einem Radius proportional zu `z` gezeichnet. (`z` ist ein Vektor mit einem Element für jede Beobachtung.) `squares=z` zeichnet entsprechende Quadrate.
- Mit `rectangles=cbind(z1,z2)` werden Rechtecke erzeugt, deren Seitenlängen die zwei Variablen `z1` und `z2` wiedergeben.
- Dem Argument `thermometers` kann man eine Matrix mit 3 oder 4 Spalten übergeben, die die Breite, Höhe und Füllung der „Thermometer“ angeben. Hat man nur 3 Variable, so wird man die dritte am ehesten als Füllung des Thermometers darstellen wollen. Man muss dazu dem Argument `thermometers` eine Matrix übergeben, die in den ersten zwei Kolonnen mit Einsen gefüllt ist, `thermometers=cbind(1,1,z)`
- Mit dem Argument `stars=mat` mit einer Matrix `mat` mit m Spalten zeichnet man Polygone (m-Ecke). Jede Richtung vom „Zentrum“ eines „Sterns“ zu einer Ecke hin symbolisiert eine zusätzliche Variable. Die Entfernung vom Zentrum zur Ecke gibt die relative Grösse dieser Variablen an.

Die Funktion `stars` ist jedoch weitaus flexibler und umfänglicher, um so eine Art Plot zu erstellen, siehe `par(ask=TRUE); example(stars)`.

Durch Farben und Strichqualitäten können weitere Variable angezeigt werden. So kann man sehr dichte grafische Darstellungen – auch überladene – erzeugen, sofern nicht zu viele Beobachtungen vorliegen.

d **Funktion `coplot`.** : Matrizen von „bedingten Streudiagrammen“.

```
> coplot(lat ~ long | depth * mag, data = quakes)
```

teilt die Beobachtungen des Datensatzes `quakes` entsprechend den Variablen `depth` und `mag` in Untergruppen ein (die sich überlappen) und zeichnet für jede Untergruppe ein Streudiagramm der Variablen `lat` gegen `long`. Das erste Argument ist eine `formula`, die Variable des Datensatzes enthält, der mit dem Argument `data` angezeigt wird (oder allenfalls andere Vektoren). Links vom Bedingungs-Zeichen | stehen die beiden Variablen, die für das Zeichnen der Streudiagramme verwendet werden, rechts die bedingenden Variablen.

e **Dynamische Grafik und Projection Pursuit.** Die beiden Packages `xgobi` und Rggobi bilden eine Brücke zu zwei Versionen eines entsprechenden Programms für dynamische Grafik, das auch Projection Pursuit-Methoden enthält.

f **Kennzahlen**
`t.x <- as.matrix(iris[1:50,1:4])` : Umwandlung in eine Matrix, nur erste Pflanzen-Art (Beobachtungen 1 bis 50)
`(t.mn <- apply(t.x, 2, mean))` : Mittelwerte für die Spalten
`(t.var <- var(t.x))` : Varianz-Kovarianz-Matrix

g **Standardisierung**
`t.xc <- scale(t.x, scale=FALSE)` : Zentrierte Beobachtungen.

`scale(t.x)` mit default-Wert `scale=TRUE` standardisiert die einzelnen Variablen. `(t.bt <- t(chol(t.var)))` : „Faktorisierung der Kovarianzmatrix“.
`t.bt%*%t(t.bt)` gibt die Kovarianzmatrix wieder.
`(t.b <- solve(t.bt))` : Inversion der Matrix $\widetilde{B}$.
`t.z <- t.xc %*% t(t.b)` : Standardisierung. Überprüfung des Ergebnisses mit `apply(t.z,2,mean)` und `var(t.z)`

h **QR-Zerlegung** Die vorhergehende Art der Standardisierung ist vom numerischen Gesichtspunkt her fahrlässig. Die gute Lösung geht über eine QR-Zerlegung:
`t.qr <- qr(t.xc)`; `t.z <- qr.Q(t.qr)`
`qr.R(t.qr)` : Transformations-Matrix

i **Verteilung der Längen der standardisierten Beobachtungen**
`t.d2 <- apply(t.z^2,1,sum)` : Längen
Quantil-Quantil-Diagramm:

```
qqplot(qchisq(ppoints(length(t.d2)),ncol(t.z)),t.d2,
  xlab="Quantiles of the Chisq. Distr.", ylab="Ordered Mahalanobis Dist.",
  main="QQ-plot for Mahalanobis Distances")
```

Literaturverzeichnis

Anderberg, M. R. (1973). *Cluster Analysis for Applications*, Academic Press, N. Y.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, N. Y.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer-Verlag, N. Y.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The S Language; A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole, Pacific Grove.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*, Springer Texts in Statistics, Springer-Verlag, New York.

Bock, H. H. (1974). *Automatische Klassifikation*, Vandenhoeck & Rupprecht, Göttingen.

Bollmann, J., Henderiks, J. and Brabec, B. (2002). Global calibration of gephyrocapsa coccolith abundance in holocene sediments for paleotemperature assessment, *Paleoceanography* **17**(3): 1035.

Bortz, J. (1977). *Lehrbuch der Statistik für Sozialwissenschaftler*, Springer Lehrbücher, Springer, Berlin.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*, Clarendon Press, Oxford, U.K.

Chambers, J. M. (1998). *Programming with Data; A Guide to the S Language*, Springer-Verlag, New York.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole.

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*, Science Paperbacks, Chapman and Hall, London.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, Summit, New Jersey. *2 Ex.*

Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.

Cooley, W. W. and Lohnes, P. R. (1971). *Multivariate Data Analysis*, Wiley, New York.

Deichsel, G. and Trampisch, H. J. (1985). *Clusteranalyse und Diskriminanzanalyse*, VEB Gustav Fischer Verlag (Stuttgart).

Everitt, B. (1980). *Cluster Analysis, Second Edition*, Halsted Press, Wiley.

Everitt, B. S. (1978). *Graphical Techniques for Multivariate Data*, Heinemann Educational Books.

Fahrmeir, L., Hamerle, A. and Tutz, G. (eds) (1996). *Multivariate statistische Verfahren*, 2nd edn, de Gruyter, Berlin.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics* **7**: 179–184.

Flury, B. (1997). *A first course in multivariate statistics*, Springer texts in statistics, Springer-Verlag, NY.

Flury, B. und Riedwyl, H. (1983). *Angewandte multivariate Statistik*, Gustav Fischer, Stuttgart.

Friedman, Hastie and Tibshirani (2000). Additive logistic regression: a statistical view of boosting, *Annals of Statistics* **28**: 377–386.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N. Y.

Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, Series in Probability and Statistics, 2nd edn, Wiley, NY.

Gordon, A. D. (1981). *Classification. Methods for the Exploratory Analysis of Multivariate Data*, Chapman & Hall, London.

Green, P. E. and Caroll, J. D. (1976). *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York.

Harman, H. H. (1960, 1967). *Modern Factor Analysis*, 2nd edn, University of Chicago Press.

Harris, R. J. (1975). *A Primer of Multivariate Statistics*, Academic Press, New York.

Hartigan, J. A. (1975). *Clustering algorithms*, Wiley.

Hastie, T. and Tibshirani, R. (1994). Discriminant analysis by gaussian mixtures, *Journal of the Royal Statistical Society B* **?**: ?

Jewell, P. L., Güsewell, S., Berry, N. R., Käuferle, D., Kreuzer, M. and Edwards, P. (2005). Vegetation patterns maintained by cattle grazing on a degraded mountain pasture. *Manuscript*

Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions*, A Wiley Publication in Applied Statistics, Wiley, New York.

Johnson, R. A. and Wichern, D. W. (1982, 1988, 1992). *Applied Multivariate Statistical Analysis*, Prentice Hall Series in Statistics, 3rd edn, Prentice Hall Int.,Englewood Cliffs,N.J.,USA.

Karson, M. J. (1982). *Multivariate Statistical Methods*, The Iowa State University Press, Ames.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, N. Y.

Kendall, M. G. (1957, 1961). *A Course in Multivariate Analysis*, Griffin's Statistical Monographs & Courses, No.2, 2nd edn, Charles Griffin, London.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis; A User's Perspective*, Oxford statistical science series; 3, 2nd edn, Oxford University Press, Oxford, UK.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths Mathematical Texts, 2nd edn, Butterworths, London.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, N. Y.

Manly, B. F. J. (1986, 1990). *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

Maxwell, A. E. (1977). *Multivariate Analysis in Behavioural Research*, Monographs on Applied Probability and Statistics, Chapman and Hall, London.

Morrison, D. F. (1967, 1976). *Multivariate Statistical Methods*, McGraw-Hill Series in Probability and Statistics, 2nd edn, McGraw-Hill Book Co., New York.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, N. Y.

Rencher, A. C. (1995). *Methods of Multivariate Analysis*, Wiley, N. Y.

Rencher, A. C. (1998). *Multivariate Statistical Inference and Applications*, Wiley, N. Y.

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions, *Applied Statistics — Journal of the Royal Statistical Society C* **42**: 615–631.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge UK.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, number 72 in *Monographs on Statistics and Applied Probability*, Chapman and Hall.

Seber, G. A. F. (1984). *Multivariate Observations*, Wiley, N. Y.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*, Freeman, San Francisco.

Späth, H. (1977). *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion*, Oldenbourg; München, Wien.

Späth, H. (1983). *Cluster-Formation und -Analyse: Theorie, FORTRAN-Programme und Beispiele*, Oldenbourg; München, Wien.

Srivastava, M. S. and Carter, E. M. (1983). *An Introduction to Applied Multivariate Statistics*, North Holland.

Stahel, W. A. (2002). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 4. Aufl., Vieweg, Wiesbaden.

Steinhausen, D. and Langer, K. (1977). *Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation*, de Gruyter, Berlin.

Tatsuoka, M. M. (1971). *Multivariate Analysis: Techniques for Educational and Psychological Research*, Wiley, New York.

Timm, N. H. (2002). *Applied Multivariate Analysis*, Springer-Verlag, N. Y.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire.

Tufte, E. R. (1990). *Envisioning Information*, Graphics Press, Cheshire.

Tufte, E. R. (1997). *Visual Explanations; Images and quantities, evidence and narrative*, Graphics Press, Cheshire.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, Statistics and Computing, 3rd edn, Springer-Verlag, New York.

Venables, W. N. and Ripley, B. D. (2000). *S Programming*, Statistics and Computing, Springer-Verlag, New York.