1
Einführung in die multivariate Datenanalyse

1.1
Was ist multivariate Datenanalyse?

Die Welt, in der wir leben, ist nicht eindimensional, sondern in großem Maße mehrdimensional. Die menschlichen Sinnesorgane haben sich dieser mehrdimensionalen Welt in erstaunlichem Maße angepasst und besitzen deshalb die Fähigkeit mehrdimensionale Daten auszuwerten. Jeder Mensch vollzieht täglich viele solcher mehrdimensionalen Auswertungen, ohne sich dessen bewusst zu sein. Wir haben z. B. kein Problem Gesichter zu unterscheiden und wieder zu erkennen. Wir können im Straßenverkehr komplexe Situationen erkennen und richtig darauf reagieren. Die Information, die wir dabei verarbeiten, liegt uns in mehreren Dimensionen vor: wir sehen die Dinge in einem dreidimensionalen Raum, wir hören, wir riechen und können auch schmecken und tasten. All diese Information können wir dazu benutzen, um Dinge oder Situationen zu unterscheiden, einzuordnen und damit zu klassifizieren. Das bedeutet nichts anderes, als dass wir eine Mustererkennung durchführen. Das folgende Beispiel soll dies noch etwas verdeutlichen. Vor nicht all zu langer Zeit wurde folgende Meldung in den Zeitungen gebracht: *Ncah eneir Sutide der Cmabridge Uinervtistät, ist es eagl in wlehcer Riehenfloge die Bcuhstbaen in eneim Wrot sethen, Haputschae der esrte und ltzete Bcuhstbae snid an der rhcitgien Setlle.*

Beim Lesen denken wir zuerst, hier hätte sich der Druckfehlerteufel eingeschlichen, aber nach einigen Worten ist es uns möglich, die Mitteilung zu erkennen, dass es nach einer Studie der Cambridge Universität egal ist, in welcher Reihenfolge die Buchstaben in einem Wort stehen. Hauptsache der erste und letzte Buchstabe sind an der richtigen Stelle.

Nun können wir ohne große Probleme die Meldung bis zu Ende lesen: *Der Rset knan ttoaels Druchenianedr sien und man knan es torztedm onhe Porbelme lseen, wiel das mneschilhce Gherin nhcit jdeen Bcuhstbaen enizlen leist, snodren das Wrot als Gnazes.*

Ihc kntöne nun afannegn, den Rset des Bcuhes onhe Rcsikühct auf ingredwleche Orthografie zu schreiben, und wir könnten es alle (mehr oder weniger gut) lesen.

Was macht unser Gehirn mit der Information der verdrehten Buchstaben? Es versucht das unbekannte Wort in die in unserem Gehirn vorhandene Liste der

Multivariate Datenanalyse: für die Pharma-, Bio- und Prozessanalytik. Waltraud Kessler

ISBN: 978-3-527-31262-7

bekannten Wörter einzuordnen, also wird eine Mustererkennung und Klassifizierung durchgeführt. Man kann das ganze nun auch in Spanisch hinschreiben: *Sgeún un etsiudo de la uiniserdvad Cmarbigde no ipmrota el oedrn de las lretas en una parbala. Lo eceisnal es que la pmerira y la umitla lreta eétsn en el lgaur crerocto.* Aber nun können nur wenige der Leser etwas mit den Buchstaben und Worten anfangen, nämlich nur diejenigen Leser, die des Spanischen kundig sind. (Richtig heißt der Satz: *Según un estudio de la universidad Cambridge no importa el orden de las letras en una palabra. Lo esencial es que la primera y la ultima letra estén en el lugar correcto.*) Das bedeutet, wir können nur Informationen verarbeiten, die wir einem uns bekannten Muster zuordnen können.

Wir werden sehen, dass die Werkzeuge der multivariaten Datenanalyse ähnlich funktionieren. Die multivariate Datenanalyse wird uns Informationen aus der Menge (häufig der Unmenge) an Daten herausarbeiten, aber schließlich werden wir es sein, mit unserem Fachwissen, die diese Informationen einsortieren und beurteilen werden. Dazu ist Vorwissen über den Sachverhalt unverzichtbar und derjenige, der mit den Daten vertraut ist und über das entsprechende Hintergrundwissen auf dem Gebiet der Physik, Chemie, Biologie, Sensorik oder anderer Fachgebiete verfügt, wird bei der Interpretation der Ergebnisse aus der multivariaten Datenanalyse dem Statistiker oder Mathematiker überlegen sein.

Ein wichtiges Lernziel in diesem Buch wird sein, die mit Hilfe mathematischer Algorithmen herausgehobenen Informationen zu interpretieren und in ein für uns erklärbares wissenschaftliches Modell oder Gerüst einzuordnen. Nur wenn wir verstehen, welche Aussagen in den Daten stecken, können wir mit dem Ergebnis der multivariaten Datenanalyse etwas Sinnvolles anfangen.

Unser menschliches Gehirn ist perfekt in der Lage, komplizierte grafische Daten (z. B. Gesichter) zu verarbeiten. Probleme haben wir aber, wenn wir eine Mustererkennung aus umfangreichen Zahlenkolonnen machen müssen. Hier bringt uns die Fähigkeit der bildhaften Mustererkennung nicht weit. Nehmen wir zur Veranschaulichung ein ganz einfaches Beispiel aus sechs Zahlenpaaren (Tabelle 1.1). Hier sind für sechs Objekte jeweils zwei Koordinaten angegeben. Wenn wir nur die Zahlenwerte betrachten, ist es für uns nicht ohne weiteres möglich zu erkennen, dass es sich um zwei Gruppen von je drei Objekten handelt.

Tabelle 1.1 Zahlenwerte für sechs Zahlenpaare

	x1	x2
Objekt 1	3	1
Objekt 2	2	5
Objekt 3	3,5	2
Objekt 4	4	1
Objekt 5	3	5
Objekt 6	2,5	4

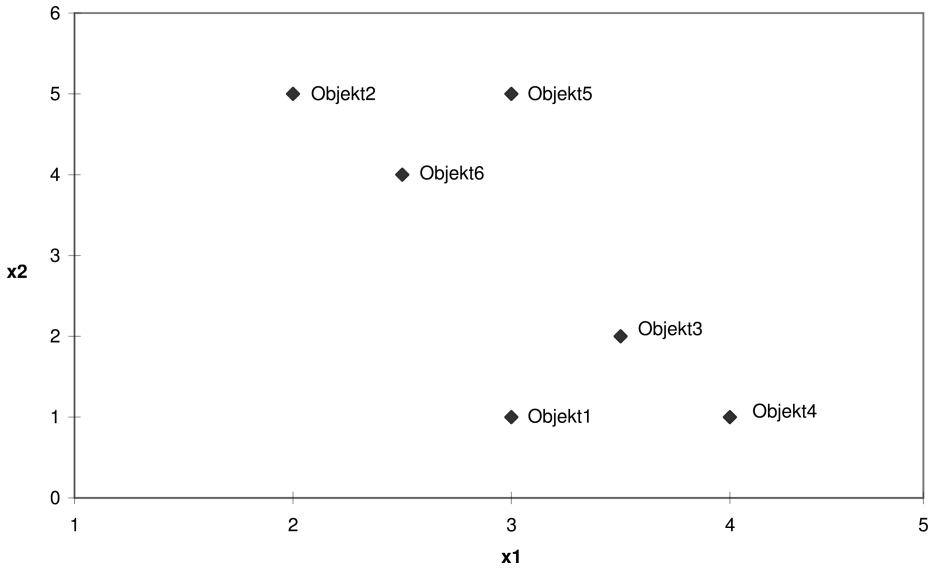

Abb. 1.1 Grafische Darstellung der Zahlenpaare aus Tabelle 1.1.

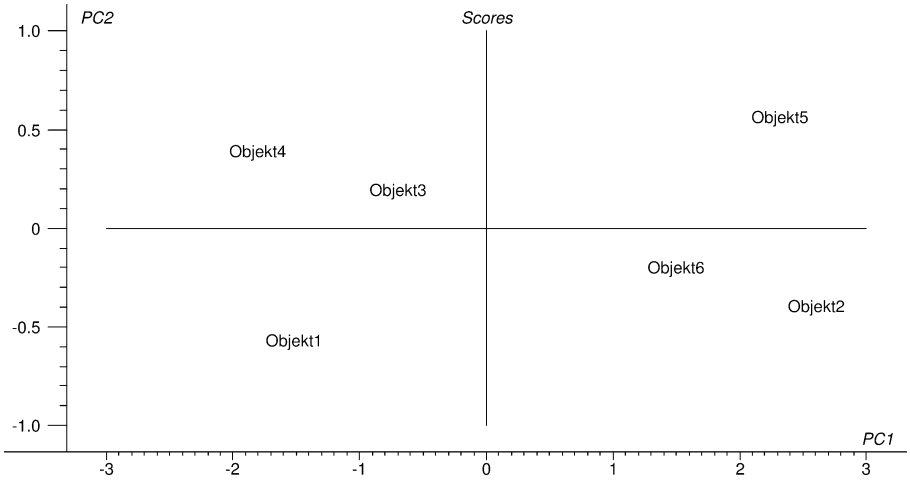

Abb. 1.2 Daten aus Tabelle 1.1 in der Darstellung nach einer Hauptkomponentenanalyse.

Betrachten wir aber die grafische Darstellung der Daten in Abb. 1.1, so erkennen wir sofort, dass es sich um zwei Gruppen handelt, die zudem noch symmetrisch angeordnet sind.

Die multivariate Datenanalyse soll genau diesen Zusammenhang der Daten herausarbeiten. Sie soll gleichzeitig beliebig viele Merkmale, die von mehreren Objekten gemessen wurden, im Zusammenhang untersuchen und das Ergebnis

dann so präsentieren, dass es leicht verständlich und klar zu erkennen ist. Dies geschieht in der Regel in grafischer Form und zwar meistens in einer zweidimensionalen grafischen Darstellung.

Nach einer Auswertung mit der Hauptkomponentenanalyse werden die Daten aus Tabelle 1.1 wie in Abb. 1.2 dargestellt. Man erkennt deutlich den (zugegebenermaßen sehr einfachen) Zusammenhang der Daten. Auffällig ist, dass die Koordinatenachsen anders angeordnet sind und nun auch andere Namen haben (PC1 und PC2). Warum das so ist, wird im nächsten Kapitel ausführlich besprochen.

1.2 Datensätze in der multivariaten Datenanalyse

Der Grund für den Einstieg in die multivariate Datenanalyse ist das Vorhandensein sehr vieler, manchmal zu vieler Daten. Meistens wurden von vielen Objekten viele verschiedene Eigenschaften gemessen. Die Beispiele in diesem Buch konzentrieren sich auf Anwendungen in der Bio- und Prozessanalytik. Die Daten werden sehr häufig spektroskopischer Art sein, denn die Spektroskopie gewinnt in der Prozessanalytik immer mehr an Bedeutung. Von verschiedenen Produkten werden Spektren aufgenommen, aus denen dann ein bestimmtes Qualitätsmerkmal für dieses Produkt berechnet werden soll. Man erhält hier sehr schnell eine sehr große Zahl an Daten. Nehmen wir z. B. ein NIR-Spektrum im Wellenlängenbereich von 1000 bis 1700 nm: Mit der Messung eines Spektrums liegen sofort 700 Werte vor, wenn die Absorption pro Nanometer gemessen wird. Macht man das für 20 verschiedene Produkte oder Produktvarianten und wird jede Messung nur zweimal wiederholt, so erhält man $20 \times 700 \times 2$ Messwerte, das sind bereits 28 000 Einzelwerte. Solch ein Datensatz ist typisch für die multivariate Datenanalyse und bezüglich der Größe durchaus noch als klein zu betrachten.

Man misst von N Objekten M Eigenschaften und erhält eine $N \times M$-Matrix, also eine Matrix mit N Zeilen und M Spalten. Üblicherweise wird in der multivariaten Datenanalyse pro Objekt eine Zeile verwendet und alle Messwerte, die zu diesem Objekt gehören, in diese Zeile geschrieben. Daten, die mit Hilfe des Tabellenkalkulationsprogramms *Excel*® erfasst werden, sind häufig genau anders herum angeordnet, so dass pro Objekt eine Spalte verwendet wurde. Das Programm *The Unscrambler*®, das in diesem Buch für die multivariate Datenanalyse verwendet wird, bietet die Möglichkeit, die Spalten in Zeilen umzuwandeln, also die Datenmatrix zu transponieren. Damit besteht keine Einschränkung bezüglich der vorhandenen Anordnung der Daten.

In diesem Buch werden als Datensätze ausschließlich zweidimensionale Datenmatrizen verwendet. Allerdings ist es prinzipiell möglich, diese Datenmatrizen um eine Dimension auf dreidimensionale Matrizen zu erweitern. Solche dreidimensionalen Matrizen erhält man z. B. in der Fluoreszenzspektroskopie, wenn für unterschiedliche Anregungswellenlängen die Emissionsspektren ge-

messen werden. Pro Messung ergibt sich eine $K \times L$-Matrix, wobei K die Anzahl der verschiedenen Anregungswellenlängen darstellt und L die Anzahl der gemessenen Emissionswellenlängen. Macht man dies für N Objekte, so ergibt sich ein Datensatz aus $K \times L \times N$ Werten. Auch HPLC (*High Performance Liquid Chromatography*) in Verbindung mit Spektroskopie ergibt solche dreidimensionalen Matrizen, ebenso die GC-Analyse (Gaschromatographie) kombiniert mit MS (Massenspektrometrie). Diese Datensätze können mit Hilfe spezieller dreidimensionaler multivariater Methoden ausgewertet werden.

Im Prinzip können mit diesen multivariaten Verfahren auch noch höher dimensionierte Datenmatrizen verarbeitet werden. In diesem Buch wird hierauf allerdings nicht eingegangen, da solche Datensätze doch recht selten sind. Eine ausführliche Abhandlung über die mehrdimensionalen Verfahren in der multivariaten Datenanalyse ist in [1] gegeben, hier wird z. B. auf eine Dreiwege-Regressionsmethode, die N-PLS, näher eingegangen.

1.3 Ziele der multivariaten Datenanalyse

Man kann die Ziele der multivariaten Datenanalyse im Wesentlichen in zwei Anwendungsbereiche einteilen.

1.3.1 Einordnen, Klassifizierung der Daten

Mit Hilfe der multivariaten Datenanalyse will man eine Informationsverdichtung oder auch Datenreduktion der Originaldaten erreichen. Aus einer großen Zahl von Messwerten sollen die relevanten Informationen herausgefunden werden. Messwerte, die den gleichen Informationsgehalt haben, werden zusammengefasst. Man kann damit die Objekte bezüglich mehrerer Messgrößen in Gruppen einteilen und erhält dabei Information über die Hintergründe, warum sich bestimmte Objekte in einer Gruppe befinden.

Mit Hilfe der Ermittlung von Zusammenhängen und Strukturen in den Daten bezüglich der Objekte und Variablen erhält man häufig Informationen über nicht direkt messbare Größen. Diese Information kann ausgenutzt werden, um z. B. Schwachstellen im Herstellungsprozess eines Produkts festzustellen und daraufhin eine gezieltere multivariate Qualitätskontrolle oder auch Prozesssteuerung aufzubauen. Auf die Methoden und Vorgehensweisen hierbei wird in diesem Buch ausführlich eingegangen. Das verwendete Verfahren für diese Datenevaluation ist die Hauptkomponentenanalyse (*Principal Component Analysis*, PCA), sie wird in Kapitel 2 ausführlich besprochen. Eine Weiterführung der Hauptkomponentenanalyse zur Klassifizierung unbekannter Objekte in bekannte Gruppen stellt das SIMCA-Verfahren dar (*Soft Independent Modelling of Class Analogy*), das in [2] besprochen wird. Außerdem gehört die Diskriminanzanalyse

dazu, die aufbauend auf Ergebnissen der PLS-Regression (*Partial Least Square Regression*) die unbekannten Objekte einordnet und ebenfalls in [2] besprochen wird.

1.3.2 Multivariate Regressionsverfahren

Die Hauptanwendung der multivariaten Verfahren besteht heutzutage in den Regressionsmethoden. Hierbei versucht man, leicht messbare Eigenschaften und schwer zu bestimmende Messgrößen, die häufig Zielgrößen genannt werden, über einen funktionalen Zusammenhang zu verbinden. Bei den Zielgrößen kann es sich z. B. um Qualitätsgrößen bei der Herstellung handeln. Immer häufiger wird bei der Produktionskontrolle oder der Überwachung einer Produkteigenschaft eine spektroskopische Kontrolle eingesetzt. Das heißt, es wird über einen bestimmten Wellenlängenbereich ein Spektrum des Produkts gemessen. Aus diesem Spektrum wird eine Zielgröße, z. B. die Konzentration eines Wirkstoffs, berechnet. Dazu benutzt man eine Kalibrierfunktion, die in einem vorausgegangenen Kalibrierprozess aufgestellt wurde und die den Zusammenhang zwischen Spektrum und Zielgröße enthält. Diese Vorgehensweise hat den Vorteil, die oft langwierig und aufwändig zu bestimmenden Zielgrößen durch einfachere, schnellere, damit meistens auch billigere spektroskopische Verfahren zu ersetzen.

Solche Regressionsverfahren können aber genauso gut in der Sensorik eingesetzt werden. Auch hier wird versucht, aufwändige Panel-Studien durch einfache und schnelle Messverfahren zumindest zum Teil zu ersetzen.

Das bekannteste Verfahren der multivariaten Regression ist die PLS-Regression (Partial Least Square Regression). Sie bietet die meisten Möglichkeiten aber auch die meisten Risiken. Denn bei unsachgemäßem Einsatz der PLS-Regression ist es möglich aus zufälligen oder unvollständigen Korrelationen Modelle zu erstellen, die in der Kalibrierung perfekt aussehen, aber über längere Zeit in der Praxis versagen. Ist man sich dieser Risiken bewusst, gibt es Wege sie zu umgehen und deshalb hat sich die PLS-Regression zusammen mit der NIR-Spektroskopie einen ersten Platz unter den multivariaten Verfahren erobert. Dieses Verfahren wird ausführlich in Kapitel 3, Abschnitte 3.9 bis 3.11 besprochen. Außer der PLS gibt es die multilineare Regression (Kapitel 3, Abschnitt 3.6) und die Hauptkomponentenregression (*Principal Component Regression*, PCR, Kapitel 3, Abschnitt 3.8). Diese Verfahren sind älter als die PLS-Regression, werden aber nicht so häufig eingesetzt, man hat sogar manchmal den Eindruck, dass sie (ungerechtfertigterweise) ganz in Vergessenheit geraten sind, da sie nicht ganz so flexibel einsetzbar sind.

1.3.3 Möglichkeiten der multivariaten Verfahren

Man kann die Möglichkeiten und Ziele der multivariaten Datenanalyse sowohl der Klassifizierungsmethoden als auch der Regressionsmethoden folgendermaßen zusammenfassen:

Ausgangspunkt der multivariaten Datenanalyse:

Datenmatrix mit vielen Objekten (N) und vielen zugehörigen Eigenschaften (M) pro Objekt.

Ziele der multivariaten Datenanalyse:

- ***Datenreduktion,***
- ***Vereinfachung,***
- ***Trennen von Information und Nicht-Information (Entfernen des Rauschens),***
- ***Datenmodellierung: Klassifizierung oder Regression,***
- ***Erkennen von Ausreißern,***
- ***Auswahl von Variablen (variable selection),***
- ***Vorhersage,***
- ***„Entmischen" von Informationen (curve resolution).***

An vielen Proben werden viele Eigenschaften gemessen (man nennt die Eigenschaften auch Attribute oder Merkmale oder man spricht einfach allgemein von Variablen). Daraus ergibt sich eine große Datenmatrix.

Wertet man diese Datenmatrix nur univariat aus, das bedeutet man schaut sich immer nur eine einzige Variable an, erhält man sehr viele Einzelergebnisse, die sich zum Teil gleichen, zum Teil widersprechen und man verliert sehr schnell den Überblick. Deshalb ist das erste Ziel der multivariaten Datenanalyse die *Datenreduktion*. Alle Variablen, die gleiche Information enthalten, werden in sog. Hauptkomponenten zusammengefasst. Damit erhält man eine Datenreduktion, da jedes Objekt dann nur noch mit den wenigen Hauptkomponenten beschrieben wird, anstatt durch die vielen einzelnen Variablen.

Mit dieser Datenreduktion erhält man eine *Vereinfachung*. Wurden z. B. in den Originaldaten 100 verschiedene Variablen verwendet, so können diese eventuell auf 10 Hauptkomponenten reduziert werden. Die Proben werden dann nur noch mit diesen 10 Hauptkomponenten beschrieben, was bedeutet, dass pro Probe nur noch 10 Hauptkomponentenwerte analysiert werden müssen, anstatt 100 Einzelmessungen.

Ein weiterer Effekt bei der multivariaten Analyse ist, dass beim Finden der Hauptkomponenten die Variablen, die Information enthalten, von den Variablen getrennt werden, die keine Information enthalten. Variable ohne Informationsgehalt erhöhen nur das Rauschen in den Daten. Die multivariate Datenanalyse trennt *Information* von *Nicht-Information* (Rauschen).

Wenn die Information aus der Vielzahl der Daten herausgefunden wurde, kann daraus ein *Modell* erstellt werden. Dieses Modell kann – abhängig von der Aufgabenstellung – ein *Klassifizierungsmodell* oder ein *Regressionsmodell* sein.

Wenn es möglich ist, für die Daten ein Modell zu berechnen, dann können die einzelnen Proben mit diesem Modell verglichen werden. Das bedeutet, dass *Ausreißer* bestimmt werden können und zwar sowohl für bereits vorliegende Proben als auch für neu hinzukommende Proben. Das ist vor allem in der Regressionsrechnung sehr wichtig. Hier kann es passieren, dass ganz salopp ausgedrückt ein Modell für Äpfel gemacht wird und hinterher Birnen untersucht werden. Dies erkennt die multivariate Datenanalyse und erklärt die Birnen zu Ausreißern.

Eine weitere optionale Möglichkeit der multivariaten Analyse ist die Auswahl von wichtigen Variablen. Da der Informationsgehalt jeder einzelnen Variablen in dem multivariaten Modell bekannt ist, können Variable, die wenig oder gar nicht zum Modell beitragen, von vornherein weggelassen werden. Damit spart man eventuell Messaufwand und die Modelle werden kleiner und robuster. Dieses Verfahren der *Variablenselektion* ist vor allem in der NIR-Spektroskopie sehr beliebt, um Bereiche mit wenig Information, die aber Einfluss auf das Signal-Rausch-Verhältnis haben, auszuschließen.

Die Modelle der multivariaten Datenanalyse können dann zur *Vorhersage* unbekannter Proben verwendet werden. Dabei spielt es keine Rolle, ob es sich um ein Klassifizierungsmodell oder ein Regressionsmodell handelt. Es werden die neuen „Rohdaten" in das Modell gegeben und je nach Modell erhält man die Klassenzugehörigkeit oder einen oder mehrere Werte für die Zielgrößen, für die das Modell aufgestellt wurde.

Die klassische multivariate Datenanalyse wurde in letzter Zeit durch viel versprechende Rotationsverfahren, sog. selbstmodellierende Kurvenauflösungsverfahren, erweitert (*Self-Modelling Curve Resolution*). Man will damit die klassischen Hauptkomponenten für den Benutzer anschaulicher darstellen. Vor allem in der Spektroskopie bietet das dem Anwender große Vorteile. Anstatt mathematisch orthogonaler Hauptkomponenten erhält man chemisch interpretierbare Spektren, die den beteiligten chemischen Komponenten entsprechen. Diese Verfahren eignen sich sehr gut zur Überwachung von Reaktionsprozessen und werden in [3] näher besprochen.

1.4 Prüfen auf Normalverteilung

Bevor man eine multivariate Datenanalyse beginnt, sollte man die Daten auf ihre statistische Zuverlässigkeit und Plausibilität überprüfen. Dazu gehört eine Überprüfung der Verteilung der Messgrößen. Handelt es sich allerdings um Spektren, muss die Verteilung nicht für jeden einzelnen Spektrumswert vorgenommen werden. Hier reicht es, sich die Spektren als ganzes grafisch anzeigen zu lassen. In der Regel erkennt man Unregelmäßigkeiten und Fehlmessun-

gen oder Extremwerte sofort spätestens nach Ausführung der Hauptkomponentenanalyse.

Nehmen wir zum Prüfen der Verteilung von Messgrößen ein Beispiel aus der Gaschromatographie (GC). Die Gaschromatographie wird häufig für die Trennung von Gasen oder verdampfbaren Flüssigkeiten und Feststoffen verwendet. Ein gasförmiges Stoffgemisch, das auch nur geringste Mengen der zu analysierenden Moleküle enthalten kann, wird mit Hilfe eines Trägergases (wie Wasserstoff, Helium, Stickstoff, Argon) durch eine Trennsäule geführt, die mit einem bestimmten Material (stationäre Phase) ausgekleidet ist. Durch unterschiedliche Verweildauern der einzelnen Komponenten in der Trennsäule aufgrund ihrer stoffspezifischen Adsorption erfolgt die analytische Trennung. Die getrennten Komponenten verlassen die Säule in bestimmten Zeitabständen und passieren einen Detektor, der die Signalstärke über der Zeit aufzeichnet. Man erhält damit ein Chromatogramm mit unterschiedlich hohen Banden (Peaks) zu bestimmten Zeiten, den sog. Retentionszeiten. Alle Banden eines Chromatogramms stehen für bestimmte Substanzen, die sich anhand ihrer Retentionszeiten bekannten Stoffen zuordnen lassen. Die Flächen der Banden (Peakflächen) sind proportional zu der Stoffmenge der jeweiligen Komponenten. Man kann mit dem GC-Verfahren also Stoffe in einem Gemisch identifizieren und über die Peakfläche auch quantitative Aussagen über diese Komponenten treffen. Der Gaschromatographie kommt in der analytischen Chemie und besonders auch in der Umweltanalytik eine breite Bedeutung zu.

Beispiel zum Prüfen von Verteilungen

In diesem Beispiel wurden 146 Obstbrände aus vier verschiedenen Obstsorten gaschromatographisch untersucht. Die Proben stammen aus vielen unterschiedlichen baden-württembergischen Brennereien aus den Jahren 1998 bis 2003. Sie wurden vom Chemischen und Veterinäruntersuchungsamt Karlsruhe mit einem Kapillar-Gaschromatographen mit Flammenionisationsdetektion auf folgende 15 Substanzen entsprechend der in [4, 5] beschriebenen Referenzanalysemethoden für Spirituosen untersucht[1]:

- Ethanol,
- Methanol,
- Propanol,
- Butanol,
- iso-Butanol,
- 2-Methyl-1-Propanol,
- 2-Methyl-1-Butanol,
- Hexanol,
- Benzylalkohol,
- Phenylethanol,

1) Mein besonderer Dank gilt hier Herrn Dr. Dirk Lachenmeier für die freundliche Überlassung der Daten.

- Essigsäuremethylester,
- Essigsäureethylester,
- Milchsäureethylester,
- Benzoesäureethylester,
- Benzaldehyd.

Für diese Substanzen wurden aus den gemessenen Peakflächen des Chromatogramms die Konzentrationen in g/hl r.A. (reiner Alkohol) bestimmt. Insgesamt wurden 54 Zwetschgenbrände, 43 Kirschbrände, 29 Mirabellenbrände und 20 Obstbrände aus Apfel&Birne untersucht. Die Daten sind auf der beiliegenden CD in der Datei „Obstbraende_GC.xls“ zu finden und im Anhang A aufgeführt.

Für die multivariate Datenanalyse gilt wie für fast alle statistischen Auswerteverfahren die Annahme normalverteilter Proben. Allerdings sind normalverteilte Daten keine zwingende Voraussetzung für die multivariaten Verfahren. Liegen keine normalverteilten Werte vor, so kann die multivariate Datenanalyse durchaus Ergebnisse liefern, häufig sind diese aber schwerer zu interpretieren und benötigen mehr Komponenten für das Modell, als dies mit normalverteilten Daten der Fall wäre. Deshalb ist es ratsam, die Verteilung vorher zu prüfen und gegebenenfalls auf eine Normalverteilung anzunähern. Dies kann durch Transformation der Messwerte erreicht werden. Sehr oft ist dabei eine Log-Transformation hilfreich (auf alle Werte wird der log, also der Logarithmus zur Basis 10 oder der ln, also der Logarithmus zur Basis e angewandt). Schiefe Verteilungen, die zu kleinen Werten verschoben sind, werden damit normalverteilt. Die transformierten Werte sind die Ausgangsdaten für die multivariate Datenanalyse.

Wichtiger als die Normalverteilung der Originaldaten ist aber eine Normalverteilung im späteren Hauptkomponentenraum. Wir werden dies bei der Analyse der Hauptkomponentenmodelle berücksichtigen und auf diese Weise eine Ausreißererkennung durchführen.

1.4.1 Wahrscheinlichkeitsplots

Ein einfaches grafisches Verfahren für die Prüfung auf Normalverteilung sind die Wahrscheinlichkeitsplots. Man trägt die gemessenen Werte auf der y-Achse auf und vergleicht sie mit der theoretischen Verteilung dargestellt als Quantile der Normalverteilung auf der x-Achse. Entspricht die untersuchte Verteilung einer Normalverteilung, liegen die Punkte auf einer Geraden.

Die Abb. 1.3 und 1.4 zeigen solche Wahrscheinlichkeitsplots für die Variablen Methanol und Hexanol.

Bei der Variablen Methanol könnte man noch eine Normalverteilung annehmen, aber bei Hexanol sind erhebliche Abweichungen von der Normalverteilung festzustellen. Doch hier ist bei der Ablehnung der Normalverteilung Vorsicht geboten. Die Daten stammen von vier verschiedenen Obstbränden, die sich ja durchaus unterscheiden können, also von verschiedenen Grundgesamt-

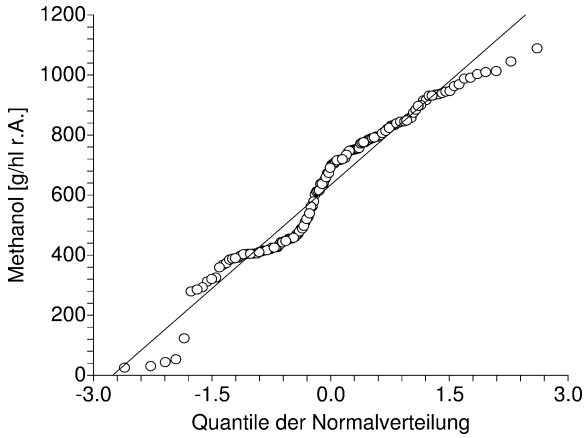

Abb. 1.3 Wahrscheinlichkeitsplot für alle Messwerte der Variable Methanol, annähernd normalverteilt.

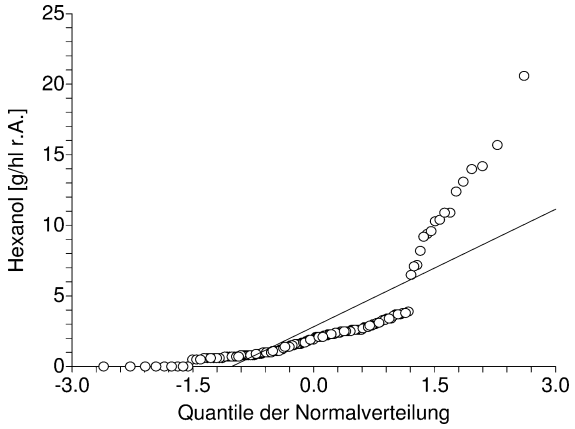

Abb. 1.4 Wahrscheinlichkeitsplot für alle Messwerte der Variable Hexanol, nicht normalverteilt.

heiten abstammen können. Deshalb ist die einfache Prüfung auf Normalverteilung mit allen Proben irreführend. Man muss die Gruppen einzeln betrachten. Dies ist in den Abb. 1.5 und 1.6 für die beiden Variablen gemacht. Man erkennt deutlich, dass die Verteilung innerhalb einer Gruppe sehr wohl normal ist. Lediglich bei Methanol weichen einige Werte für den Apfel&Birnen-Brand von der geraden Kurve ab, aber die Abweichung ist nicht so groß, als dass Anpassungsbedarf besteht.

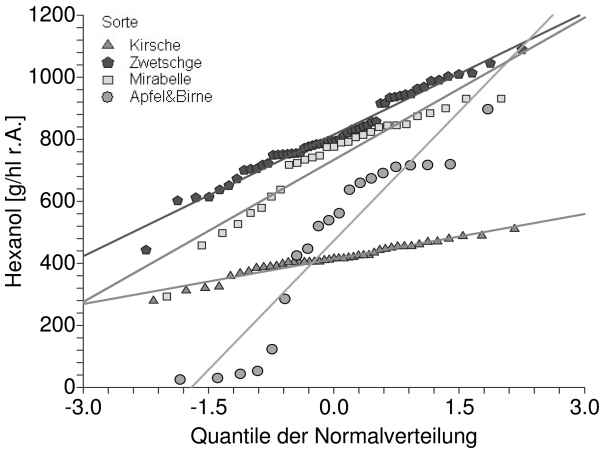

Abb. 1.5 Wahrscheinlichkeitsplot für alle Messwerte für die Variable Methanol nach Obstbrandsorten getrennt, normalverteilt.

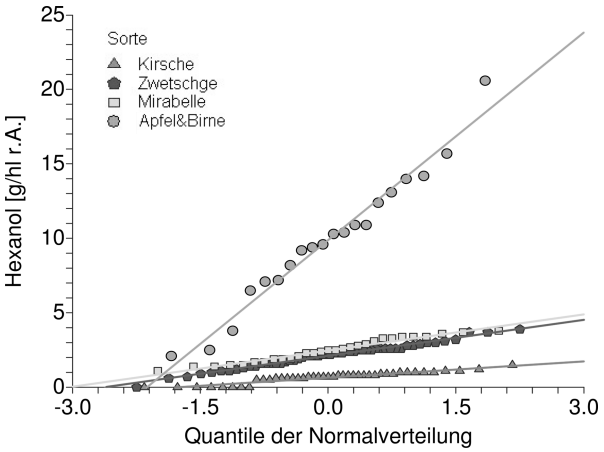

Abb. 1.6 Wahrscheinlichkeitsplot für alle Messwerte für die Variable Hexanol nach Obstbrandsorten getrennt, normalverteilt.

1.4.2 Box-Plots

Auch die Box-Plots dienen dazu, die Verteilungen der verschiedenen Variablen miteinander zu vergleichen. Man erkennt, ob die Verteilung symmetrisch ist, ob es Ausreißer bzw. extreme Werte gibt und wie groß die Streuung innerhalb der Messreihe ist. Der Box-Plot stellt eine Häufigkeitsverteilung dar und reduziert diese Häufigkeitsverteilung auf die Angabe von fünf wichtigen Werten, die die Verteilung beschreiben: Median, 1. und 3. Quartil, unterer und oberer Whisker.

Zwischen dem 1. und 3. Quartil wird ein Kasten aufgebaut (das ist der Quartilsabstand, engl. *Interquartile Range*, IRQ). In diesen Bereich fallen 50% der Messwerte. Die seitlich angrenzenden Whisker vermitteln einen Eindruck, wie weit die restlichen 50% der Werte streuen. Bevor also ein Box-Plot gezeichnet werden kann, müssen die Werte der Größe nach sortiert werden und dann die fünf die Verteilung charakterisierenden Werte bestimmt werden. Zur Übersicht sind diese Werte im Folgenden noch einmal aufgeführt. Außerdem sind die Endmarken des oberen und unteren Whiskers für den einfachen und den modifizierten Box-Plot angegeben. Beide Varianten werden verwendet. Beim modifizierten Box-Plot werden die Extremwerte klarer erkennbar.

Werte für Box-Plot, die charakteristisch für die Verteilung sind:

- ***Median: unterhalb und oberhalb des Medians liegen je 50% der Messwerte.***
- ***1. Quartil: unterhalb des 1. Quartils liegen 25% der Messwerte, damit liegen 75% darüber.***
- ***3. Quartil: unterhalb des 3. Quartils liegen 75% der Messwerte und 25% darüber.***
- ***Quartilsabstand (IQR): innerhalb des Quartilsabstands liegen 50% der Messwerte.***
- ***Whisker: die senkrechten Linien werden Whisker genannt.***

Standard-Box-Plot

- ***Endmarke für oberen Whisker: größter Wert der Datenreihe.***
- ***Endmarke für unteren Whisker: niedrigster Wert der Datenreihe.***
- ***Ausreißer: Ausreißer werden nicht gekennzeichnet.***

Modifizierter Box-Plot

- ***Endmarke des oberen Whisker: größter Messwert, der kleiner oder gleich dem 3. Quartil ist plus 1,5·IRQ.***
- ***Endmarke des unteren Whiskers: kleinster Messwert, der größer oder gleich dem 1. Quartil ist minus 1,5·IQR.***
- ***Innerhalb der Whisker des modifizierten Box-Plots befinden sich ca. 95% der Daten, wenn die Whiskerlänge 1,5·IQR beträgt.***
- ***Ausreißer: alle Werte größer bzw. kleiner als die Endmarke der Whisker werden als Ausreißer mit einem Kreis gekennzeichnet.***

Die Abb. 1.7 und 1.8 zeigen die Box-Plots für die Variablen Methanol und Hexanol.

Die Verteilung aller Methanolwerte ist nicht perfekt normalverteilt, denn der Median ist nicht genau in der Mitte der Box. Wir erhalten also das gleiche Ergebnis wie mit dem Wahrscheinlichkeitsplot. Die Unterschiede zwischen den unteren 50% und den oberen 50% der Daten sind aber auch für diesen Box-Plot nicht zu groß. Die Daten sind also nicht zu weit von einer Normalverteilung entfernt. Ganz anders sieht es bei den Hexanolwerten aus. Der Median liegt

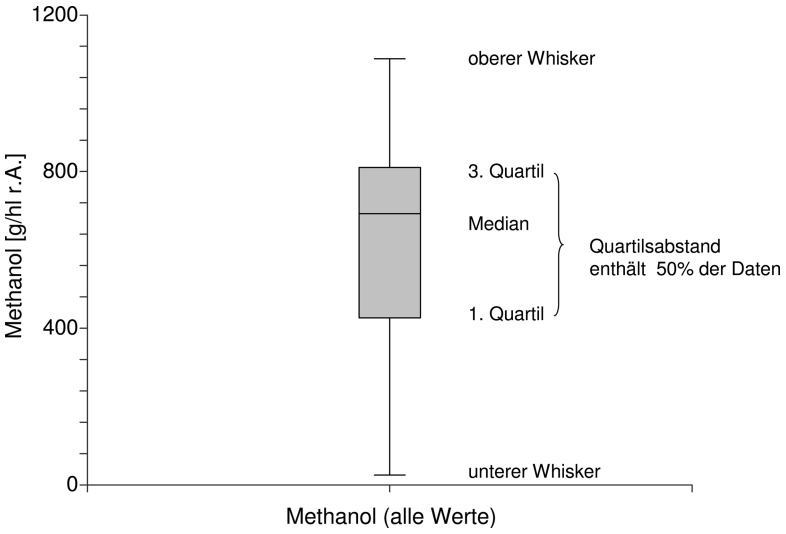

Abb. 1.7 Box-Plot für Methanol für alle Werte.

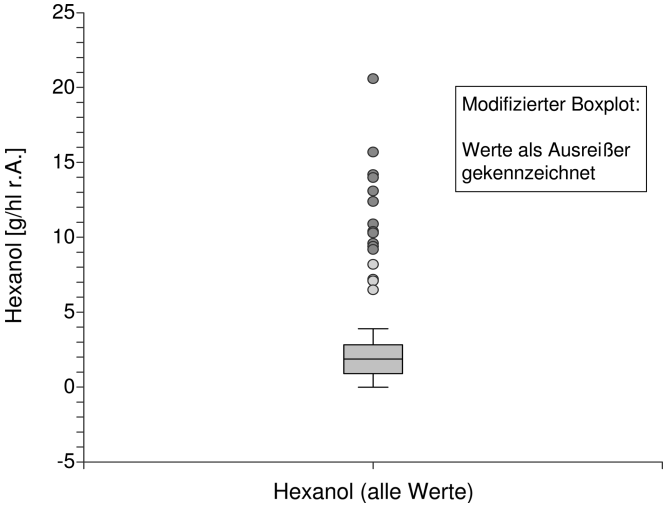

Abb. 1.8 Box-Plot für Hexanol für alle Werte.

zwar ziemlich genau in der Mitte der Box, aber es gibt oberhalb sehr viele Messwerte, die als Ausreißer gekennzeichnet sind. Damit ist der Median auch nicht annäherungsweise in der Mitte aller Daten, sondern sehr stark zu kleinen Werten verschoben. Diese Verteilung ist eindeutig nicht normalverteilt. Wie aus dem Wahrscheinlichkeitsplot zu sehen war, handelt sich in Wirklichkeit um mehrere Verteilungen.

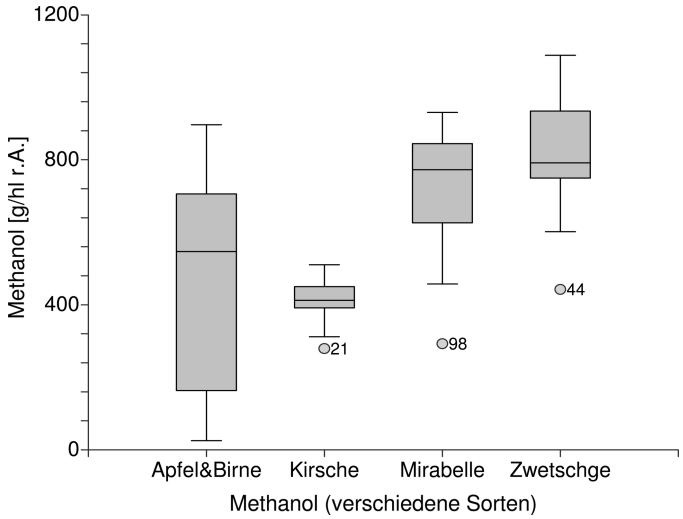

Abb. 1.9 Box-Plots für Methanol nach Obstbrandsorten getrennt.

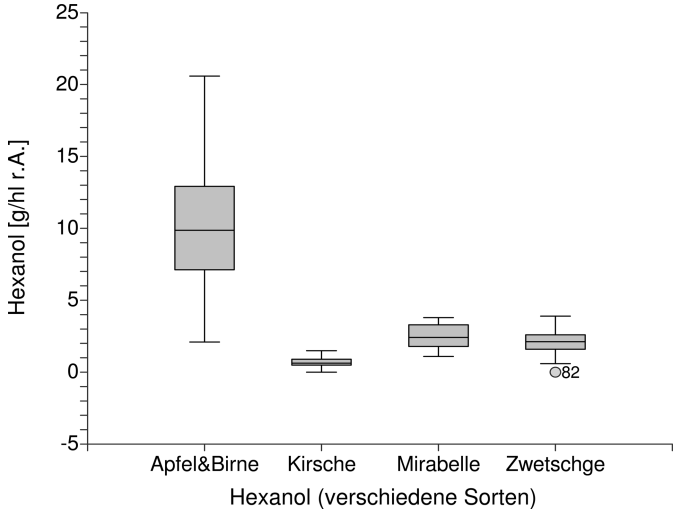

Abb. 1.10 Box-Plots für Hexanol nach Obstbrandsorten getrennt.

Die Abb. 1.9 und 1.10 stellen die Box-Plots nach Obstbrandsorten getrennt dar. Wir erkennen, dass „Apfel&Birne“ für das Methanol einen sehr großen Bereich abdeckt, während „Kirsche“ nur geringe Unterschiede in den Werten aufweist. Die Werte von „Mirabelle“ und „Zwetschge“ sind deutlich höher als die von „Kirsche“. Bei allen drei letztgenannten Sorten gibt es einen Ausreißer. Die Zahl neben dem Punkt gibt die Proben-Nummer an, die in der Tabelle 1.2 ver-

wendet wird. Die Sorte „Apfel&Birne“ zeigt bei Hexanol (Abb. 1.10) genauso wie bei Methanol die größte Varianz in den Messwerten. Es fällt auf, dass die Unterschiede in den Hexanolwerten bei den übrigen drei Sorten nur einen Bruchteil der Sorte „Apfel&Birne“ betragen. Auch hier gibt es bei „Zwetschge“ einen Wert (Probe 82), der außerhalb des 95%-Datenbereichs liegt.

1.5 Finden von Zusammenhängen

1.5.1 Korrelationsanalyse

Mit den Wahrscheinlichkeitsplots erhält man Information über die Verteilung der Messwerte. Über die Zusammenhänge der Messwerte untereinander wird aber noch nichts ausgesagt. Man kann nun mit einfachen grafischen Mitteln versuchen, erste Zusammenhänge in den Daten zu erkennen. Besonders gut geeignet dazu sind die Streudiagramme, auch Scatterplots genannt. Man trägt die Werte einer unabhängigen Variablen x über den Werten einer anderen unabhängigen Variablen y auf. Dabei können die Korrelationen der Daten untereinander sichtbar werden. Man kann vor allem auch nicht lineare Zusammenhänge erkennen, die bei einer reinen linearen Korrelationsrechnung nicht berücksichtigt werden. Allerdings werden die Streudiagramme ab einer Variablenzahl von etwa 20 relativ unübersichtlich, denn man muss sich dann bereits durch 400 Streudiagramme „durcharbeiten“. Deshalb macht es Sinn, auch eine Korrelationsmatrix für die Daten zu erstellen.

Tabelle 1.3 zeigt die Korrelationstabelle für die Obstbrände. Es wurde für jedes Variablenpaar (x_i, y_i) der Pearsonsche Korrelationskoeffizient r nach Gl. (1.1) für I Variablenpaare berechnet. Die Summe im Zähler wird Kovarianz genannt, sie bestimmt das Vorzeichen des Korrelationskoeffizienten. Da durch die Standardabweichung aller x_i und y_i Werte geteilt wird, ist der Wertebereich auf –1 bis +1 beschränkt. Ein positives Vorzeichen bedeutet, die beiden Variablen korrelieren in der gleichen Richtung, d. h. wenn x_i größer wird, wächst auch y_i, während ein negatives Vorzeichen auf einen gegenläufigen Zusammenhang hinweist, wenn x_i wächst, nimmt y_i ab.

$$r_{xy} = \frac{\sum_{i=1}^{I}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{I}(x_i - \overline{x})^2 \sum_{i=1}^{I}(y_i - \overline{y})^2}} \tag{1.1}$$

Tabelle 1.2 Korrelationstabelle für gaschromatographisch bestimmte Peakflächen der 15 Messvariablen der Obstbrände.

	Methanol	Propanol	Butanol	iso-Butanol	2-Methyl-1-Propanol	2-Methyl-1-Butanol	Hexanol	Benzylalkohol	Phenylethanol	Essigsäuremethylester	Essigsäureethylester	Milchsäureethylester	Benzoesäureethylester	Benzaldehyd
Methanol	1.00	–0.34	0.40	–0.15	0.32	–0.07	0.08	–0.23	–0.19	**0.53**	0.01	–0.32	–0.12	0.43
Propanol		1.00	–0.42	0.37	–0.25	–0.24	–0.30	0.33	–0.22	–0.03	**0.51**	0.49	0.35	–0.16
Butanol			1.00	0.07	0.13	0.28	0.56	–0.34	0.27	0.19	–0.23	–0.28	–0.31	0.12
iso-Butanol				1.00	–0.04	0.28	0.34	–0.06	0.25	0.02	0.24	0.13	–0.20	–0.23
2-Methyl-1-Propanol					1.00	**0.60**	0.32	–0.31	0.16	0.19	–0.09	–0.24	–0.38	0.13
2-Methyl-1-Butanol						1.00	**0.75**	–0.28	**0.68**	–0.04	–0.21	–0.14	–0.52	–0.17
Hexanol							1.00	–0.32	**0.63**	0.06	–0.19	–0.17	–0.55	–0.20
Benzylalkohol								1.00	–0.05	–0.08	0.18	**0.79**	0.48	0.05
Phenylethanol									1.00	–0.07	–0.14	0.04	–0.47	–0.24
Essigsäuremethylester										1.00	**0.71**	0.01	–0.05	0.10
Essigsäureethylester											1.00	0.33	0.21	–0.11
Milchsäureethylester												1.00	0.36	–0.12
Benzoesäureethylester													1.00	0.23
Benzaldehyd														1.00

Die Korrelation kann in folgende Grenzen eingeteilt werden:

0	<	$\|r\|$	<	0,2	sehr geringe Korrelation
0,2	<	$\|r\|$	<	0,5	geringe Korrelation
0,5	<	$\|r\|$	<	0,7	mittlere Korrelation
0,7	<	$\|r\|$	<	0,9	hohe Korrelation
0,9	<	$\|r\|$	<	1	sehr hohe Korrelation

Aus Gründen der Übersichtlichkeit sind die Korrelationskoeffizienten nur in die obere Hälfte der Tabelle 1.2 eingetragen, die dazu symmetrischen Werte unterhalb der Diagonalen sind weggelassen.

Man erkennt nur wenige Variable (x_i, y_i), die untereinander mit einem $r > 0{,}5$ korreliert sind. Den größten Korrelationskoeffizienten hat Milchsäureethylester und Benzylalkohol mit $r = 0{,}79$, während z. B. Methanol mit Hexanol so gut wie gar nicht korreliert ist ($r = 0{,}08$).

1.5.2 Bivariate Datendarstellung – Streudiagramme

Die Korrelationen dieser beiden Variablenpaare sind in den Abb. 1.11 und 1.12 gezeigt. In Abb. 1.11 erkennt man deutlich die hohe positive Korrelation von Milchsäureethylester und Benzylalkohol. Diese Korrelation ist unabhängig von der Obstbrandsorte; hoher Benzylalkoholgehalt bedeutet auch einen hohen Milchsäureethylestergehalt (Ausnahme „Apfel&Birne“).

Wirft man nur einen flüchtigen Blick auf Abb. 1.12, so stimmt man mit der Aussage $r = 0{,}08$, also keine Korrelation und damit kein Zusammenhang zwischen

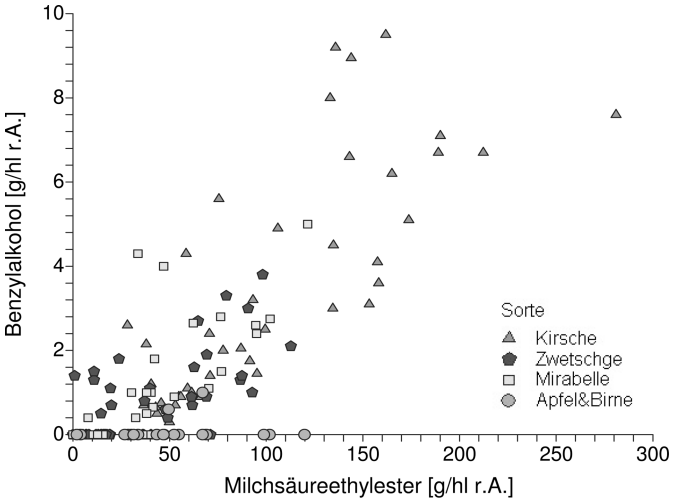

Abb. 1.11 Streudiagramm nach Sorten gekennzeichnet für Milchsäureethylester und Benzylalkohol ($r = 0.79$).

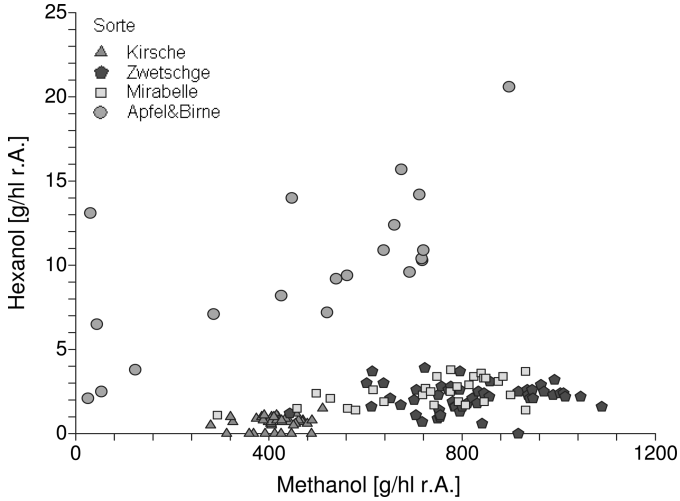

Abb. 1.12 Streudiagramm nach Sorten gekennzeichnet für Methanol und Hexanol ($r = 0.08$).

den beiden Variablen Hexanol und Methanol, überein. Schaut man aber genauer hin, so erkennt man, dass die Obstbrandsorten anhand dieser zwei Variablen bereits in Gruppen eingeteilt werden. Die Proben der „Apfel&Birne"-Sorte haben fast alle höhere Hexanolkonzentrationen, während die „Zwetschgen" und „Mirabellen" höhere Methanolkonzentrationen haben als die „Kirschen". Auch mehrere „Apfel&Birne"-Proben haben hohe Methanolwerte, aber gleichzeitig sind auch deren Hexanolwerte höher als bei „Zwetschge" und „Mirabelle", damit ist bei gleichzeitiger Betrachtung beider Messwerte eine eindeutige Unterscheidung möglich. Schaut man dagegen nur auf eine Variable allein oder auf die Korrelationen der beiden Variablen, ist keine Unterscheidung der Sorten möglich.

Eine ausführliche verständliche Besprechung dieser grundlegenden statistischen Betrachtungen und Darstellungen von Daten findet sich in dem Buch von Clarke und Cooke [6] und speziell für den Bereich der Biologie in dem Buch von Sokal und Rohlf [7].

Was hat uns die bisherige Datenbetrachtung an Information über die GC-Werte Methanol und Hexanol der vier verschiedenen Obstbrandsorten gebracht? Wir wissen nun, dass sich die vier Sorten in den Mittelwerten und den Varianzen unterscheiden, die Verteilungen sind innerhalb der Sorten normalverteilt, Benzylalkohol und Milchsäureethylester sind am stärksten korreliert und Hexanol und Methanol gemeinsam betrachtet teilen die Sorten in recht eindeutige Gruppen ein, allerdings lassen sich „Mirabelle" und „Zwetschge" nicht unterscheiden.

Diese ganzen Aussagen beruhen aber immer nur auf dem Vergleich von maximal zwei Variablen. Dies soll im Folgenden geändert werden. Wir wollen alle Variablen gleichzeitig betrachten. Dazu werden wir die Hauptkomponentenanalyse verwenden.

Literatur

1 A. Smilde, R. Bro and P. Geladi, Multi-way analysis with applications in the chemical sciences. John Wiley & Sons Inc., Chichester, 2004.
2 J.-H. Jiang, R. Tsenkova and Y. Ozaki, Principal Discriminant Variate Method for Classification of Multicollinear Data: Principle and Applications, Analytical Sciences (2001) 17, 471–474.
3 R. Tauler, A. Smilde and B. R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. J Chemom (1995) 9, 31–58.
4 Referenzanalysemethoden für Spirituosen. EG-Verordnung Nr. 2870/2000 vom 19. 12. 2000.
5 D. W. Lachenmeier und F. Musshoff, Begleitstoffgehalte alkoholischer Getränke, Verlaufskontrollen, Chargenvergleich und aktuelle Konzentrationsbereiche. Rechtsmedizin (2004) 14, 454–462.
6 G. M. Clarke and D. Cooke, A Basic Course in Statistics. Arnold Publishers, London, 2005.
7 R. R. Sokal and F. J. Rohlf, Biometry – The Principles and Practice of Statistics in Biological Research. Freeman and Co., New York, 2000.