Waltraud Kessler

Multivariate Datenanalyse

für die Pharma-, Bio- und Prozessanalytik

Ein Lehrbuch

WILEY-VCH Verlag GmbH & Co. KGaA

Waltraud Kessler

Multivariate Datenanalyse

200 Jahre Wiley – Wissen für Generationen

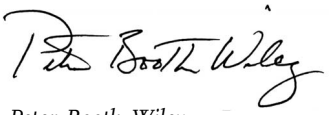
John Wiley & Sons feiert 2007 ein außergewöhnliches Jubiläum: Der Verlag wird 200 Jahre alt. Zugleich blicken wir auf das erste Jahrzehnt des erfolgreichen Zusammenschlusses von John Wiley & Sons mit der VCH Verlagsgesellschaft in Deutschland zurück. Seit Generationen vermitteln beide Verlage die Ergebnisse wissenschaftlicher Forschung und technischer Errungenschaften in der jeweils zeitgemäßen medialen Form.

Jede Generation hat besondere Bedürfnisse und Ziele. Als Charles Wiley 1807 eine kleine Druckerei in Manhattan gründete, hatte seine Generation Aufbruchsmöglichkeiten wie keine zuvor. Wiley half, die neue amerikanische Literatur zu etablieren. Etwa ein halbes Jahrhundert später, während der „zweiten industriellen Revolution" in den Vereinigten Staaten, konzentrierte sich die nächste Generation auf den Aufbau dieser industriellen Zukunft. Wiley bot die notwendigen Fachinformationen für Techniker, Ingenieure und Wissenschaftler. Das ganze 20. Jahrhundert wurde durch die Internationalisierung vieler Beziehungen geprägt – auch Wiley verstärkte seine verlegerischen Aktivitäten und schuf ein internationales Netzwerk, um den Austausch von Ideen, Informationen und Wissen rund um den Globus zu unterstützen.

Wiley begleitete während der vergangenen 200 Jahre jede Generation auf ihrer Reise und fördert heute den weltweit vernetzten Informationsfluss, damit auch die Ansprüche unserer global wirkenden Generation erfüllt werden und sie ihr Zeil erreicht. Immer rascher verändert sich unsere Welt, und es entstehen neue Technologien, die unser Leben und Lernen zum Teil tiefgreifend verändern. Beständig nimmt Wiley diese Herausforderungen an und stellt für Sie das notwendige Wissen bereit, das Sie neue Welten, neue Möglichkeiten und neue Gelegenheiten erschließen lässt.

Generationen kommen und gehen: Aber Sie können sich darauf verlassen, dass Wiley Sie als beständiger und zuverlässiger Partner mit dem notwendigen Wissen versorgt.

William J. Pesce
President and Chief Executive Officer



Peter Booth Wiley
Chairman of the Board

Waltraud Kessler

Multivariate Datenanalyse

für die Pharma-, Bio- und Prozessanalytik

Ein Lehrbuch

WILEY-VCH Verlag GmbH & Co. KGaA

Prof. Waltraud Kessler
Hochschule Reutlingen
STZ Prozesskontrolle und
Datenanalyse
STI Multivariate Datenanalyse
Herderstraße 47
72762 Reutlingen

Bibliografische Information
der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Printed in the Federal Republic of Germany
Gedruckt auf säurefreiem Papier

Satz K+V Fotosatz GmbH, Beerfelden
Druck betz-druck GmbH, Darmstadt
Bindung Litges & Dopf Buchbinderei GmbH, Heppenheim

ISBN: 978-3-527-31262-7

Inhaltsverzeichnis

Vorwort *XI*

1 Einführung in die multivariate Datenanalyse *1*

1.1 Was ist multivariate Datenanalyse? *1*
1.2 Datensätze in der multivariaten Datenanalyse *4*
1.3 Ziele der multivariaten Datenanalyse *5*
1.3.1 Einordnen, Klassifizierung der Daten *5*
1.3.2 Multivariate Regressionsverfahren *6*
1.3.3 Möglichkeiten der multivariaten Verfahren *7*
1.4 Prüfen auf Normalverteilung *8*
1.4.1 Wahrscheinlichkeitsplots *10*
1.4.2 Box-Plots *12*
1.5 Finden von Zusammenhängen *16*
1.5.1 Korrelationsanalyse *16*
1.5.2 Bivariate Datendarstellung – Streudiagramme *18*
Literatur *20*

2 Hauptkomponentenanalyse *21*

2.1 Geschichte der Hauptkomponentenanalyse *21*
2.2 Bestimmen der Hauptkomponenten *22*
2.2.1 Prinzip der Hauptkomponentenanalyse *22*
2.2.2 Was macht die Hauptkomponentenanalyse? *24*
2.2.3 Grafische Erklärung der Hauptkomponenten *25*
2.2.4 Bedeutung der Faktorenwerte und Faktorenladungen (Scores und Loadings) *29*
2.2.5 Erklärte Varianz pro Hauptkomponente *35*
2.3 Mathematisches Modell der Hauptkomponentenanalyse *36*
2.3.1 Mittenzentrierung *37*
2.3.2 PCA-Gleichung *38*
2.3.3 Eigenwert- und Eigenvektorenberechnung *38*

2.3.4 Berechnung der Hauptkomponenten mit dem NIPALS-Algorithmus *40*
2.3.5 Rechnen mit Scores und Loadings *42*
2.4 PCA für drei Dimensionen *46*
2.4.1 Bedeutung von Bi-Plots *48*
2.4.2 Grafische Darstellung der Variablenkorrelationen zu den Hauptkomponenten (Korrelation-Loadings-Plots) *52*
2.5 PCA für viele Dimensionen: Gaschromatographische Daten *56*
2.6 Standardisierung der Messdaten *65*
2.7 PCA für viele Dimensionen: Spektren *72*
2.7.1 Auswertung des VIS-Bereichs (500–800 nm) *74*
2.7.2 Auswertung des NIR-Bereichs (1100–2100 nm) *81*
2.8 Wegweiser zur PCA bei der explorativen Datenanalyse *86*
Literatur *88*

3 Multivariate Regressionsmethoden *89*

3.1 Klassische und inverse Kalibration *90*
3.2 Univariate lineare Regression *92*
3.3 Maßzahlen zur Überprüfung des Kalibriermodells (Fehlergrößen bei der Kalibrierung) *93*
3.3.1 Standardfehler der Kalibration *93*
3.3.2 Mittlerer Fehler – RMSE *94*
3.3.3 Standardabweichung der Residuen – SE *95*
3.3.4 Korrelation und Bestimmtheitsmaß *96*
3.4 Signifikanz und Interpretation der Regressionskoeffizienten *97*
3.5 Grafische Überprüfung des Kalibriermodells *97*
3.6 Multiple lineare Regression (MLR) *99*
3.7 Beispiel für MLR – Auswertung eines Versuchsplans *100*
3.8 Hauptkomponentenregression (Principal Component Regression – PCR) *103*
3.8.1 Beispiel zur PCR – Kalibrierung mit NIR-Spektren *105*
3.8.2 Bestimmen des optimalen PCR-Modells *106*
3.8.3 Validierung mit unabhängigem Testset *110*
3.9 Partial Least Square Regression (PLS-Regression) *111*
3.9.1 Geschichte der PLS *112*
3.10 PLS-Regression für eine *Y*-Variable (PLS1) *113*
3.10.1 Berechnung der PLS1-Komponenten *114*
3.10.2 Interpretation der P-Loadings und W-Loadings bei der PLS-Regression *117*
3.10.3 Beispiel zur PLS1 – Kalibrierung von NIR-Spektren *117*
3.10.4 Finden des optimalen PLS-Modells *118*
3.10.5 Validierung des PLS-Modells mit unabhängigem Testset *121*
3.10.6 Variablenselektion – Finden der optimalen *X*-Variablen *122*
3.11 PLS-Regression für mehrere *Y*-Variablen (PLS2) *127*

3.11.1 Berechnung der PLS2-Komponenten *127*
3.11.2 Wahl des Modells: PLS1 oder PLS2? *129*
3.11.3 Beispiel PLS2: Bestimmung von Gaskonzentrationen in der Verfahrenstechnik *130*
3.11.4 Beispiel 2 zur PLS2: Berechnung der Konzentrationen von Einzelkomponenten aus Mischungsspektren *141*
Literatur *151*

4 Kalibrieren, Validieren, Vorhersagen *153*

4.1 Zusammenfassung der Kalibrierschritte – Kalibrierfehler *154*
4.2 Möglichkeiten der Validierung *155*
4.2.1 Kreuzvalidierung (Cross Validation) *156*
4.2.2 Fehlerabschätzung aufgrund des Einflusses der Datenpunkte (Leverage Korrektur) *157*
4.2.3 Externe Validierung mit separatem Testset *159*
4.3 Bestimmen des Kalibrier- und Validierdatensets *162*
4.3.1 Kalibrierdatenset repräsentativ für *Y*-Datenraum *164*
4.3.2 Kalibrierdatenset repräsentativ für *X*-Datenraum *164*
4.3.3 Vergleich der Kalibriermodelle *165*
4.4 Ausreißer *168*
4.4.1 Finden von Ausreißern in den *X*-Kalibrierdaten *169*
4.4.2 Grafische Darstellung der Einflüsse auf die Kalibrierung *172*
4.4.2.1 Einfluss-Grafik: Influence Plot mit Leverage und Restvarianz *172*
4.4.2.2 Residuenplots *174*
4.5 Vorhersagebereich der vorhergesagten *Y*-Daten *175*
4.5.1 Grafische Darstellung des Vorhersageintervalls *177*
Literatur *181*

5 Datenvorverarbeitung bei Spektren *183*

5.1 Spektroskopische Transformationen *183*
5.2 Spektrennormierung *185*
5.2.1 Normierung auf den Mittelwert *186*
5.2.2 Vektornormierung auf die Länge eins (Betrag-1-Norm) *186*
5.3 Glättung *187*
5.3.1 Glättung mit gleitendem Mittelwert *187*
5.3.2 Polynomglättung (Savitzky-Golay-Glättung) *187*
5.4 Basislinienkorrektur *190*
5.5 Ableitungen *193*
5.5.1 Ableitung nach der Differenzenquotienten-Methode (Punkt-Punkt-Ableitung) *193*
5.5.2 Ableitung über Polynomfit (Savitzky-Golay-Ableitung) *195*
5.6 Korrektur von Streueffekten *198*
5.6.1 MSC (Multiplicative Signal Correction) *198*
5.6.2 EMSC (Extended Multiplicative Signal Correction) *199*

5.6.3 Standardisierung der Spektren (Standard Normal Variate (SNV) Transformation) *202*
5.7 Vergleich der Vorbehandlungsmethoden *203*
Literatur *210*

6 Eine Anwendung in der Produktionsüberwachung – von den Vorversuchen zum Einsatz des Modells *211*

6.1 Vorversuche *211*
6.2 Erstes Kalibriermodell *217*
6.3 Einsatz des Kalibriermodells – Validierphase *220*
6.4 Offset in den Vorhersagewerten der zweiten Testphase *224*
6.5 Zusammenfassung der Schritte bei der Erstellung eines Online-Vorhersagemodells *227*

7 Tutorial zum Umgang mit dem Programm „The Unscrambler" der Demo-CD *229*

7.1 Durchführung einer Hauptkomponentenanalyse (PCA) *229*
7.1.1 Beschreibung der Daten *229*
7.1.2 Aufgabenstellung *230*
7.1.3 Datendatei einlesen *230*
7.1.4 Definieren von Variablen- und Objektbereichen *231*
7.1.5 Speichern der Datentabelle *232*
7.1.6 Plot der Rohdaten *233*
7.1.7 Verwendung von qualitativen Variablen (kategoriale Variable) *235*
7.1.8 Berechnen eines PCA-Modells *238*
7.1.9 Interpretation der PCA-Ergebnisse *241*
7.1.9.1 Erklärte Varianz (Explained Variance) *241*
7.1.9.2 Scoreplot *242*
7.1.9.3 Loadingsplot *247*
7.1.9.4 Einfluss-Plot (Influence Plot) *250*
7.2 Datenvorverarbeitung *253*
7.2.1 Berechnung der zweiten Ableitung *253*
7.2.2 Glättung der Spektren *256*
7.2.3 Berechnen der Streukorrektur mit EMSC *257*
7.3 Durchführung einer PLS-Regression mit einer Y-Variablen *261*
7.3.1 Aufgabenstellung *261*
7.3.2 Interpretation der PLS-Ergebnisse *266*
7.3.2.1 PLS-Scoreplot *266*
7.3.2.2 Darstellung der Validierungsrestvarianzen (Residual Validation Variance) *269*
7.3.2.3 Darstellung der Regressionskoeffizienten *270*
7.3.2.4 Darstellung der vorhergesagten und der gemessenen Theophyllinkonzentrationen (Predicted versus Measured Plot) *271*
7.3.2.5 Residuenplot *273*

7.4 Verwenden des Regressionsmodells – Vorhersage des Theophyllingehalts für Testdaten *276*
7.5 Export der Unscrambler-Modelle zur Verwendung in beliebigen Anwendungen *278*
7.5.1 Kalibriermodell für Feuchte erstellen *279*
7.5.2 Export des PLS-Regressionsmodells für die Feuchte *283*
7.5.2.1 Umwandeln der Grafikanzeige in numerische Daten *283*
7.5.2.2 Export des Regressionsmodells als Text-Datei (ASCII Model) *285*
7.5.2.3 Berechnung der Feuchte in Excel *286*
7.6 Checkliste für spektroskopische Kalibrierungen mit dem Unscrambler *287*
Literatur *290*

Anhänge A–D *291*

Anhang A *292*
Anhang B *302*
Anhang C *304*
Anhang D *310*

Stichwortverzeichnis *313*

Vorwort

Multivariate Methoden sind seit vielen Jahren ein wichtiges Hilfsmittel bei der Analyse großer Datenmengen. Die Verfahren waren allerdings häufig nur „Chemometrie-Insidern“ bekannt. In den letzten 10 Jahren, vor allem durch den Einsatz der Spektroskopie in der chemischen Analytik, ist der Bekanntheitsgrad der multivariaten Verfahren beträchtlich gestiegen. Die pharmazeutische und chemische Industrie bewies in vielen Anwendungen die Leistungsfähigkeit dieser Methoden und demonstrierte damit einem größeren Publikum in den Ingenieur- und Naturwissenschaften deren Alltagstauglichkeit. Heutzutage werden die Verfahren in fast allen Industriezweigen angewandt. Dazu gehören neben der chemischen und pharmazeutischen Industrie die Lebensmittelindustrie, die Geowissenschaften, die Biowissenschaften sowie die Medizinwissenschaften. Auch in den Sozialwissenschaften und im Marketingbereich gewinnen die multivariaten Analysemethoden immer mehr Anwender.

Das vorliegende Buch soll einen einfachen Einstieg in die multivariate Datenanalyse ermöglichen. Es wendet sich an Studierende in naturwissenschaftlichen und ingenieurwissenschaftlichen Fächern sowie an Praktiker aus allen Bereichen der Industrie und der Forschung. Dem Nutzer soll ein ausreichender mathematischer Hintergrund der multivariaten Verfahren vermittelt werden. Gleichzeitig wird viel Wert auf Anschaulichkeit und Interpretation gelegt. An Beispielen aus der industriellen Praxis wird die Theorie verdeutlicht und es gibt viele Hinweise und Tipps für die Anwendung der Verfahren beim Auswerten großer Datenmengen.

Dass dies eine „Gratwanderung“ zwischen Wissenschaftlichkeit, Anschaulichkeit und Praxisnähe ist und damit auch Konflikte in sich birgt, liegt auf der Hand. Ich bin deshalb jedem Leser dankbar für Hinweise, Kritiken, Anregungen und Vorschläge zu Inhalt und Darstellungen dieses Buches.

Seit vielen Jahren lehre ich an der Hochschule Reutlingen die Fächer Statistik, Statistische Versuchsplanung (Design of Experiments) und Multivariate Datenanalyse im Bereich der chemischen Ingenieurwissenschaften und in zahlreichen Kursen für die Industrie. Der Mangel an deutschsprachiger Literatur auf diesem Gebiet und die wiederholte Bitte, das Skriptum der Vorlesung bzw. der Kurse ausführlicher zu gestalten, führte schließlich zur Erstellung dieses Buches. Es gliedert sich im Wesentlichen in fünf Teile:

- Explorative Datenanalyse mit Hilfe der Hauptkomponentenanalyse
- Multivariate Regressionsmethoden wie die MLR, PCR und PLS
- Methoden der Kalibrierung, Validierung und Vorhersage
- Datenvorverarbeitung bei Spektren
- Anwendung und Durchführung multivariater Methoden mit Hilfe spezieller Software

Der erste Teil des Buches widmet sich der Hauptkomponentenanalyse. Es wird anhand eines Beispiels der Lebensmittelanalyse und anhand von NIR-Spektren erklärt, wie eine explorative Datenanalyse durchzuführen ist, um Wissen aus unübersichtlich erscheinenden Daten herauszuarbeiten.

Der zweite Teil grenzt die unterschiedlichen multivariaten Regressionsmethoden wie MLR, PCR und PLS voneinander ab, zeigt die Vor- und Nachteile auf und demonstriert deren Anwendung an zahlreichen Beispielen aus der Industrie.

Nicht minder wichtig ist das richtige Vorgehen bei der Kalibrierung und Validierung von Regressionsmodellen. Dies wird im dritten Teil des Buches ausführlich diskutiert. Anhand einer Anwendung in der Produktionsüberwachung wird von den ersten Vorversuchen bis zum Einsatz des Modells gezeigt, wie ein robustes Regressionsmodell erstellt, validiert und gegebenenfalls korrigiert wird.

Die Spektroskopie erlebt in den letzten Jahren in der chemischen Analytik und in der Pharmazeutischen Prozesskontrolle einen regelrechten Boom, wobei zur Auswertung vorwiegend multivariate Regressionsmethoden zum Einsatz kommen. Aus diesem Grunde wurde der Spektrenvorverarbeitung ein eigenes Kapitel gewidmet.

Eine wichtige Motivation für dieses Buch war, dem Leser die Möglichkeit zu geben, sich im Selbststudium oder studienbegleitend in das komplizierte Gebiet der multivariaten Datenanalyse einzuarbeiten. Deshalb liegt dem Buch eine CD mit einer Trainingsversion des Programmpakets „The Unscrambler“ bei, die von der Fa. CAMO Software AS freundlicherweise zur Verfügung gestellt wurde, wofür ich Frau Valerie Lengard ganz besonders danke. „The Unscrambler“ ist eines der am häufigsten benutzten Programme für diese Methoden. Alle Beispiele des Buches können anhand der CD selbständig nachvollzogen werden. Der Umgang mit der professionellen Software „The Unscrambler“ wird dem Leser in einem Tutorial am Ende des Buches vermittelt.

Ganz herzlich danken möchte ich Herrn Dr. Dirk Lachenmeier, Herrn Dr. Christian Lauer, Herrn Joachim Mannhardt, Frau Anke Roder und Frau Kerstin Mader für die Aufbereitung und Bereitstellung einiger Datensätze. Vielen Dank auch den Firmen, dass ich aktuelle Projektbeispiele und Daten in diesem Buch veröffentlichen darf, was nicht immer selbstverständlich ist. Weitere Daten wurden im Rahmen von Forschungsprojekten innerhalb der Abteilung Prozessanalytik des Instituts für Angewandte Forschung der Hochschule Reutlingen erhalten. Für die Bereitstellung dieser Daten und die vielen fruchtbaren Diskussionen bezüglich deren Auswertung und Interpretation möchte ich mich ganz besonders bei Herrn Prof. Dr. Rudolf Kessler bedanken. Bedanken möchte ich mich auch bei meiner Tochter Wiltrud für die Durchsicht der Manuskripte aus

der Sichtweise des Studierenden, bei Herrn Dr. Dirk Lachenmeier, der die Anwenderseite vertrat und bei Herrn Prof. Dr. Claus Kahlert für die Überprüfung auf mathematische Korrektheit.

Ausdrücklich danke ich Frau Renate Dötzer und Frau Claudia Grössl vom Verlag Wiley-VCH für die bereitwillige Unterstützung und große Geduld, die sie stets für mich aufbrachten. Insbesondere gilt mein Dank aber meiner Familie und meinen Freunden, die mich in all den vergangenen Monaten in vielfältiger Hinsicht unterstützt haben, und vor allem viel Verständnis dafür aufbrachten, dass meine Prioritäten vorwiegend zugunsten des Buches ausgefallen sind.

Reutlingen, im September 2006 *Waltraud Kessler*

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

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

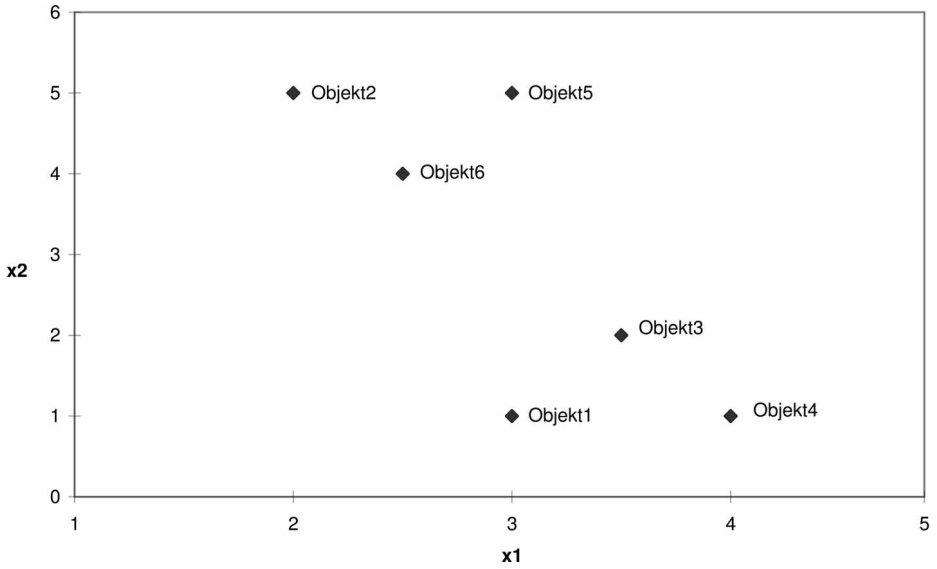


Abb. 1.1 Grafische Darstellung der Zahlenpaare aus Tabelle 1.1.

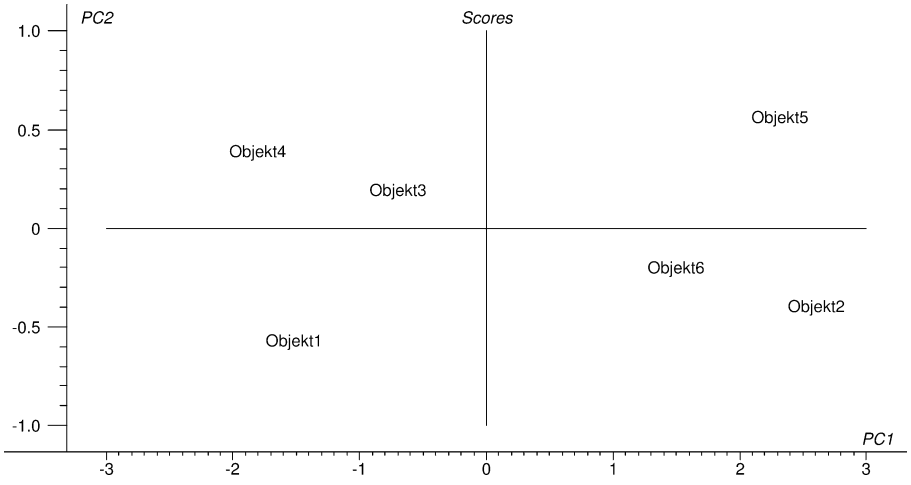


Abb. 1.2 Daten aus Tabelle 1.1 in der Darstellung nach einer Hauptkomponentenanalyse.

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-

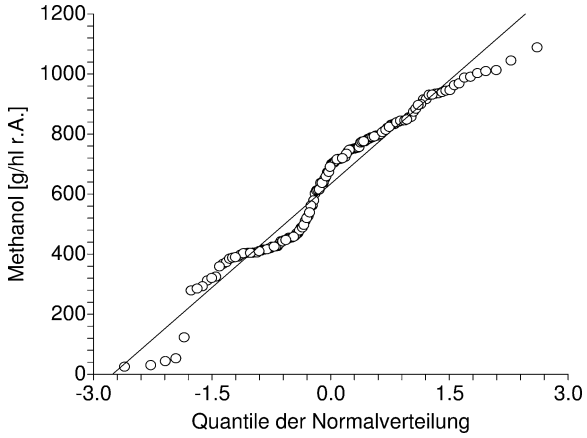


Abb. 1.3 Wahrscheinlichkeitsplot für alle Messwerte der Variable Methanol, annähernd normalverteilt.

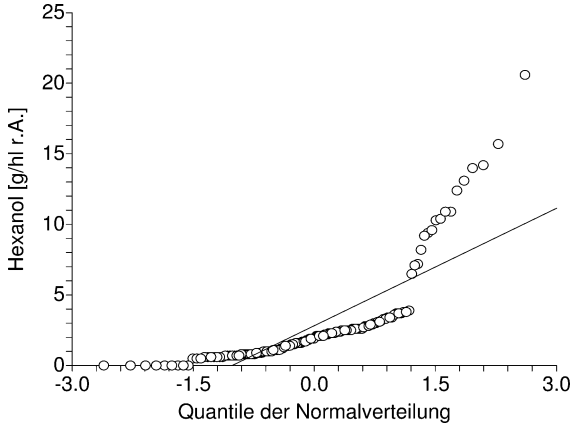


Abb. 1.4 Wahrscheinlichkeitsplot für alle Messwerte der Variable Hexanol, nicht normalverteilt.

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.

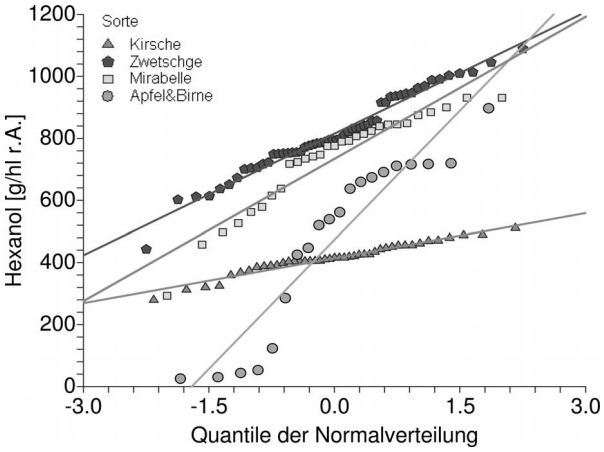


Abb. 1.5 Wahrscheinlichkeitsplot für alle Messwerte für die Variable Methanol nach Obstbrandsorten getrennt, normalverteilt.

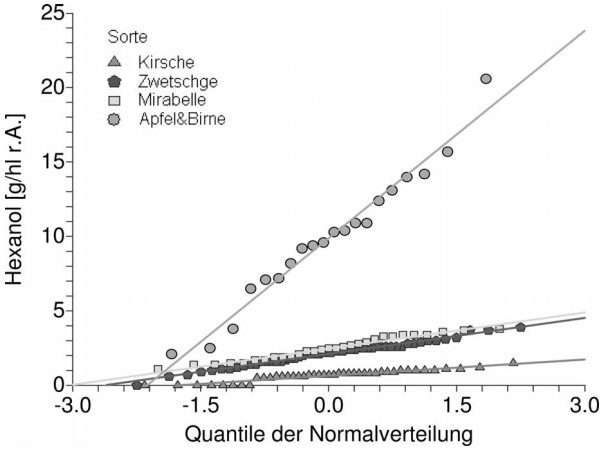


Abb. 1.6 Wahrscheinlichkeitsplot für alle Messwerte für die Variable Hexanol nach Obstbrandsorten getrennt, normalverteilt.

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

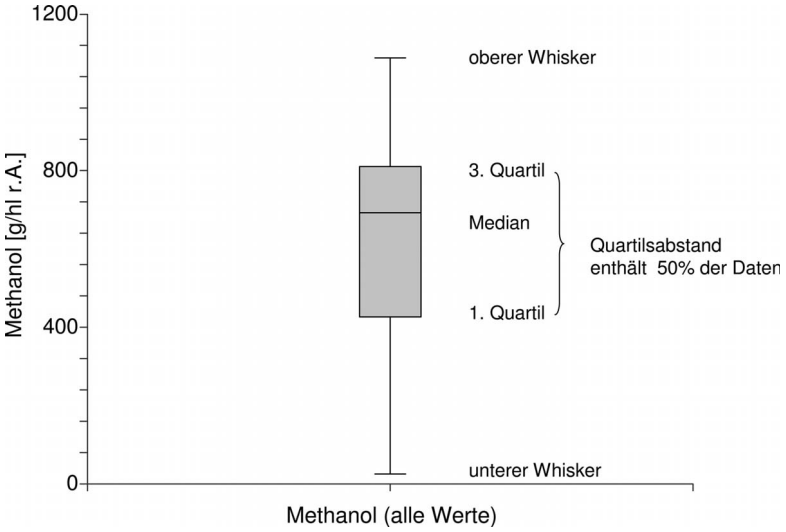


Abb. 1.7 Box-Plot für Methanol für alle Werte.

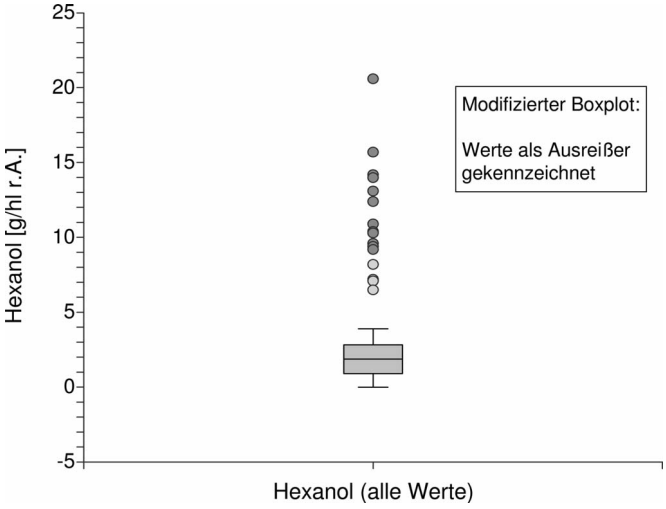


Abb. 1.8 Box-Plot für Hexanol für alle Werte.

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.