2.5 Meßdatenauswertung mit Hilfe chemometrischer Methoden

Bei der Analyse komplexer Proben ist es mit Hilfe moderner Analysensysteme möglich, simultan eine Vielzahl von Eigenschaften zu bestimmen und so einen typischen „Fingerabdruck" von Merkmalen einer Probe zu erhalten. Dabei sammelt sich sehr schnell ein großer Datenbestand an. Bei der Interpretation dieser großen Datenmengen ist die Erkennung und Veranschaulichung von Zusammenhängen zwischen scheinbar weit auseinander liegenden Meßgrößen notwendig – mitunter bei sehr hohem Versuchsfehler. Um eine objektive Auswertung zu erreichen, greift man in zunehmendem Maße – unterstützt durch den Einsatz von Computersystemen mit leistungsstarken Prozessoren – auf chemometrische[1] Methoden zurück.

Eine besondere Form der Datenanalyse, die multivariate Datenanalyse, beschäftigt sich mit der Systematisierung und Klassifizierung von Merkmalsmustern. Der Mensch besitzt mit seinen Sinnesorganen die Fähigkeit, eine Vielzahl von Merkmalen wie Formen, Geräusche, Gerüche, Tastempfindungen usw. zu klassifizieren. Viele Menschen können Personen anhand ihrer Gesichter zuordnen und erreichen dabei oft eine erstaunliche Leistungsfähigkeit. Diese Zuordnung erfolgt dabei nicht aufgrund einzelner Merkmale wie Nasenlänge oder Mundbreite, sondern beruht auf gewissen, nicht direkt meßbaren, latenten Charakteristika (Formen), die als gewichtete Mischung der meßbaren Merkmale aufgefaßt werden können [88].

Ein Datensatz wird von zwei Größen aufgespannt: Von den Untersuchungsgegenständen, Proben, etc., die als *„Objekte"* bezeichnet werden, und den *„Variablen"*, d.h. den Meßgrößen, Eigenschaften oder Parametern [88]. In dem Ausdruck „Variable" liegt gleichzeitig auch schon die Information, daß die Größe fehlerbehaftet ist. Es ist üblich, die Objekte in Zeilen und die Variablen in Spalten in einer Tabelle darzustellen.

Bei der Einteilung der multivariaten Verfahren kann man eine Unterscheidung in primär *struktur-prüfende* Verfahren und primär *struktur-entdeckende* Verfahren vornehmen [89].

[1] Def. „Chemometrie": Einsatz der Informatik zur Verarbeitung analytisch-chemischer Fragestellungen [87].

Bei einem struktur-prüfenden Verfahren werden die Zusammenhänge zwischen Variablen überprüft. Der Anwender besitzt eine Vorstellung über die Zusammenhänge zwischen Variablen und möchte diese mit Hilfe multivariater Verfahren überprüfen. Zu der Gruppe der struktur-prüfenden Methoden gehört die Regressionsanalyse, die Varianzanalyse und die Diskriminanzanalyse.

Primäres Ziel der struktur-entdeckenden Verfahren, zu denen z.B. die Faktorenanalyse, die Clusteranalyse und die Hauptkomponentenanalyse zählen, ist die Entdeckung von Zusammenhängen zwischen Variablen oder zwischen Objekten.

2.5.1 Die Faktorenanalyse

Die Faktorenanalyse wird als struktur-entdeckendes Verfahren eingesetzt, wenn bei Fragestellungen, die eine Vielzahl von Einflußfaktoren (Variablen) beinhalten, eine Reduzierung der Variablen angestrebt wird. Besonders für empirische Untersuchungen lassen sich durch Anwendung der Faktorenanalyse erhebliche Vorteile realisieren: Nach Test einer Reihe von Einflußfaktoren kann entschieden werden, welche Variablen oder „Variablenbündel" (die sogenannten Faktoren) relevant zur Erklärung des Sachverhaltes sind [89].

Um die „hinter den Variablen" stehenden Faktoren ermitteln zu können ist in einem ersten Schritt eine Untersuchung der Variablen auf „Bündelungsfähigkeit" notwendig. Mit Hilfe der Korrelationsrechnung wird überprüft, inwieweit die Variablen voneinander unabhängig sind und ob sich Zusammenhänge zwischen einzelnen Variablen (d.h. ähnliche Eigenschaften von Variablen) erkennen lassen.

Die Korrelation zwischen zwei Variablen x_1 und x_2 wird ausgedrückt durch den Koeffizienten r:

$$r_{x1,x2} = \frac{\sum_{k=1}^{K} (x_{1k} - \bar{x_1}) \cdot (x_{2k} - \bar{x_2})}{\sqrt{\sum_{k=1}^{K} (x_{1k} - \bar{x_1})^2 \cdot \sum_{k=1}^{K} (x_{2k} - \bar{x_2})^2}} \tag{2.9}$$

bzw.

$$r_{x1,x2} = \frac{S_{x1,x2}}{\sqrt{S_{x1}^2 \cdot S_{x2}^2}} \tag{2.10}$$

und

$$S_{x1,x2} = \frac{1}{K-1} \cdot \sum_{k=1}^{K} (x_{1k} - \bar{x_1}) \cdot (x_{2k} - \bar{x_2}) \tag{2.11}$$

mit:	$r_{x1,x2}$	=	Korrelationskoeffizient
	x_{k1}	=	Ausprägung der Variable 1 bei Objekt k
	$\bar{x}_1$	=	Mittelwert der Ausprägungen von Variabler 1 über alle Objekte k
	x_{k2}	=	Ausprägung der Variable 2 bei Objekt k
	$\bar{x}_2$	=	Mittelwert der Ausprägungen von Variabler 2 über alle Objekte k
	$S_{x1,x2}$	=	Kovarianz
	K	=	Anzahl der Objekte

Für eine bestehende Matrix von Ausgangsdaten werden die Korrelationskoeffizienten über alle Eigenschaften bestimmt und in einer Korrelationsmatrix festgehalten. Vor Durchführung der Korrelationsrechnung empfiehlt sich eine Standardisierung der Ausgangsdatenmatrix; dies ermöglicht eine Vergleichbarkeit von Variablen, die auf unterschiedlichen Skalen erhoben wurden (z.B. Konzentration eines Elementes gemessen in μg/g und Probeneinwaagemenge in mg). Eine Standardisierung der Datenmatrix erfolgt durch Bildung der Differenz zwischen dem Mittelwert und dem jeweiligen Beobachtungswert einer Variablen sowie anschließender Division durch die Standardabweichung. Durch die Standardisierung wird sichergestellt, daß der neue Mittelwert gleich Null und die neue Standardabweichung gleich Eins ist.

Die Werte von standardisierten Datenmatrizen werden – im Unterschied zur Ausgangsmatrix – nicht mehr mit x, sondern mit z bezeichnet.

Die standardisierte Variable lautet:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k} \tag{2.12}$$

mit:
- x_{ik} = Beobachtungswert der Variablen k bei Objekt i
- $\bar{x}_k$ = Durchschnitt aller Beobachtungswerte der Variablen k über alle Objekte
- s_k = Standardabweichung der Variablen k
- z_{ik} = Standardisierter Beobachtungswert der Variablen k bei Objekt i

Durch Berechnung der Korrelationskoeffizienten und Aufstellung der zugehörigen Matrix aus standardisierten Daten kann aufgezeigt werden, welche Variablen in Zusammenhang stehen. In einem nächsten Schritt wird in der Faktorenanalyse die Frage beantwortet, mit welchem Gewicht die Faktoren an der Beschreibung der beobachteten Zusammenhänge beteiligt sind. Eine größere oder geringere Bedeutung wird durch eine Gewichtszahl, die sogenannte „Faktorladung" charakterisiert. Die Faktorladung gibt an, wieviel ein Faktor mit einer Ausgangsvariablen zu tun hat.

Während mit der Faktorenanalyse eine Verdichtung auf *Variablenebene* vorgenommen wird, wird mit Hilfe der Clusteranalyse versucht, eine Reduzierung auf *Objektebene* zu erreichen.

2.5.2 Die Clusteranalyse

Unter dem Begriff „Clusteranalyse" versteht man Verfahren zur Gruppenbildung. Das durch sie zu verarbeitende Material besteht im allgemeinen aus einer Vielzahl von Objekten, beispielsweise Personen, Unternehmen, Analysenproben etc.. Ausgehend von diesen Daten besteht die Zielsetzung der Clusteranalyse darin, Objekte mit weitgehend verwandten Eigenschaften (d.h. Objekte die sich möglichst ähnlich sind) zu Gruppen zusammenzufassen. Bei der Clusteranalyse geht es somit immer um die Analyse einer heterogenen Gesamtheit von Objekten, mit dem Ziel, homogene Teilmengen von Objekten aus dieser Gesamtheit zu identifizieren.

Der Ablauf einer Clusteranalyse unterteilt sich in zwei Schritte [89]:

1. Wahl des Proximitätsmaßes
   Man überprüft für jeweils zwei Objekte die Ausprägungen der Eigenschaftsmerkmale und versucht, durch einen Zahlenwert die Unterschiede bzw. Übereinstimmungen zu messen. Die berechnete Zahl ist ein Symbol für die Ähnlichkeit der Objekte hinsichtlich der untersuchten Merkmale.

   Es lassen sich zwei Arten von Proximitätsmaßen unterscheiden:

   - Ähnlichkeitsmaß:
     Je größer der Wert des Ähnlichkeitsmaßes zweier Objekte ist, desto ähnlicher sind sie sich.
   - Distanzmaß:
     Je größer die Distanz zweier Objekte ist, desto unähnlicher sind sich zwei Objekte.

2. Wahl des Fusionierungsalgorithmus
   Aufgrund der Ähnlichkeitswerte werden die Objekte so zu Gruppen zusammengefaßt, daß sich Objekte mit weitgehend übereinstimmenden Eigenschaftsstrukturen in einer Gruppe wiederfinden.

Charakteristisch für die Clusteranalyse ist die gleichzeitige Einbeziehung aller vorliegenden Eigenschaften der Objekte zur Gruppenbildung, sodaß für die Ermittlung der Ähnlichkeit zwischen zwei Objekten immer ein Paarvergleich angestellt wird.

Weitergehende, sehr verständliche Einführungen in die Clusteranalyse und ihre Methodik, bieten die Monographien von BACKHAUS et al. [89] sowie STEINHAUSEN und LANGER [90].

Die Clusteranalyse findet in der Chemie eine breite praktische Anwendung. Auf dem Gebiet der Umweltchemie, z.B. bei der Untersuchung von Spurenelementen in menschlichem Hirngewebe [91] oder der Analyse von Aerosolproben aus der Norwegischen Arktis [92] wurde die Clusteranalyse ebenso eingesetzt wie im Bereich der Lebensmittelchemie, wo beispielsweise die Elementgehalte in Torf bestimmt wurden [93].

2.5.3 Die Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (engl. Principal Component Analysis, PCA) zählt ebenfalls zu den struktur-entdeckenden Verfahren. Sie ist eine der wichtigsten multivariaten Techniken, weil sie einerseits die Möglichkeit bietet, die komplexen Strukturen eines Datensatzes graphisch gut zu veranschaulichen, und weil andererseits durch Reduktion der Datenmatrix und gleichzeitiger Bildung von Hauptkomponenten neue statistische Variablen mit günstigen Eigenschaften entstehen, die auch in einer Reihe anderer Problemstellungen (Regression, Klassifikation, Clusteranalyse) Verwendung finden.

Durch Linearkombinationen der vorhandenen, *realen* Variablen werden neue, *künstliche* Variablen erzeugt, die selbst nicht direkt meßbar sind, sondern aus den gemessenen Variablen errechnet werden. Hauptkomponenten sind optimale Linearkombinationen, d.h. eine gewichtete Summe der ursprünglichen gemessenen Variablen. Der Vorteil solcher künstlicher Größen besteht in ihrem hohen Informationsgehalt; sie tragen mehr Informationen als jede einzelne Ausgangsvariable.

Es gibt verschiedene Möglichkeiten, die Problemstellung der Hauptkomponentenanalyse zu formulieren. Die verschiedenen historischen Ausgangspunkte führten aber alle zu demselben abstrakten Kern. Ausgangspunkt sind nicht die Rohdaten selbst, sondern zentrierte bzw. standardisierte Werte. Es sei ein Datensatz $\mathbf{X}$ mit n Objekten (Zeilen) und p Variablen (Spalten) gegeben, wobei $p \leq n$ ist. Dazu lassen sich beispielhaft drei Problemstellungen formulieren, die alle zueinander äquvalent sind [88]:

1. Geometrischer Standpunkt (PEARSON, 1901 [94]):
   Die Objekte (Zeilen) eines Datensatzes $\mathbf{X}$ sind Punkte in einem p-dimensionalen euklidischen Raum. Gesucht wird ein r-dimensionaler linearer Unterraum, der sich der Punktmenge optimal anpaßt.

2. Statistischer Standpunkt (HOTELLING, 1933 [95]):
   Es werden $r \leq p$ normierte, untereinander unkorrellierte Linearkombinationen (die Hauptkomponenten) der p Originalvariablen gesucht, die sukzessive maximale Varianz ausschöpfen.

3. Standpunkt der mehrdimensionalen Skalierung (GOWER, 1966 [96]):
   Es wird eine Projektion der Objektpunkte auf einen r-dimensionalen Unterraum gesucht, so daß dabei alle paarweise Objektabstände möglichst gut erhalten bleiben.

Die Gemeinsamkeit aller Problemstellungen liegt darin, daß eine Reduktion der Dimension p des Variablenraumes auf die im allgemeinen erheblich kleinere Dimension r (z.B. $r = 2$ für graphische Darstellungen) erreicht werden soll, ohne dabei allzuviel Information zu verlieren.

Der erste Schritt bei einer Hauptkomponentenanalyse besteht in einer geeigneten Datenvorbehandlung. Im Regelfall erfolgt zunächst eine Transformation der Rohdaten, die in einer Zentrierung und anschließenden Standardisierung der Datensätze besteht. Im Anschluß hat jede Spalte den gleichen Mittelwert Null (Zentrierung) und die gleiche Standardabweichung und auch Streuung Eins (Standardisierung), wie es Abbildung 2.1 veranschaulicht.

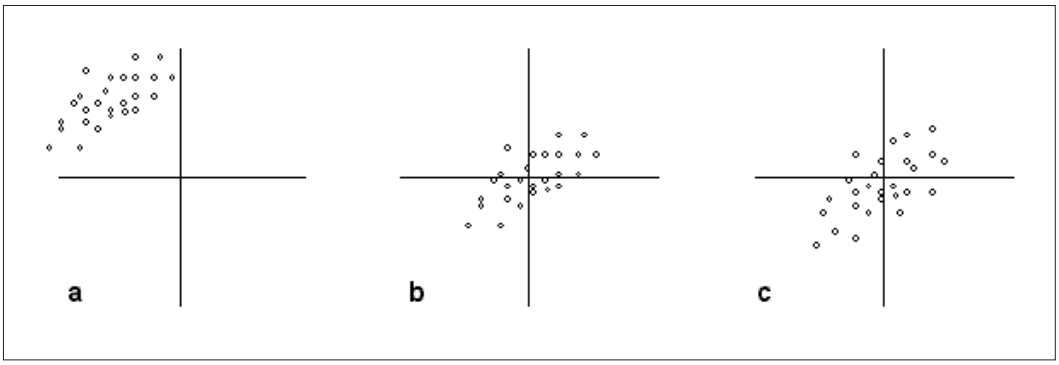

Abbildung 2.1: *Veranschaulichung der Datenvorbehandlung: (a) ursprünglicher, (b) zentrierter und (c) standardisierter Datensatz [88]*

Zur Interpretation der Resultate der PCA ist die Entscheidung darüber, wieviele Hauptkomponenten notwendig bzw. signifikant sind, um die Datenstruktur - abgesehen von Meßwertrauschen oder Experimentalfehlern - adäquat zu reproduzieren, von großer Bedeutung. Zu den am häufigsten eingesetzten Schnellverfahren zählen das Eigenwert-Eins-Kriterium nach KAISER [97] und der Scree-Test nach CATTELL [98].

Beim Eigenwert-Eins-Kriterium werden solche Hauptkomponenten als signifikant klassifiziert, deren zugehörige Eigenwerte der Korrelationsmatrix über dem mittleren Eigenwert liegen, d.h. deren Streuung überdurchschnittlich ist. Für standardisierte Daten ist die Summe aller Eigenwerte gleich der Variablenzahl, der durchschnittliche Eigenwert ist folglich stets Eins.

Der Scree-Test geht von der Annahme aus, daß sich die Eigenwerte zu Korrelationsmatrizen von Datensätzen, die nur Zufallszahlen enthalten, typischerweise asymptotisch der Abszisse nähern. Bei der Analyse realer Datensätze macht sich dieses Verhalten erst ab einem bestimmten Eigenwert bemerkbar. Dieser kennzeichnet den wesentlichen vom unwesentlichen Teil der Datenstruktur. In Abb. 2.2 sind exemplarisch 10 Eigenwerte einer hypothetischen Korrelationsmatrix dargestellt.

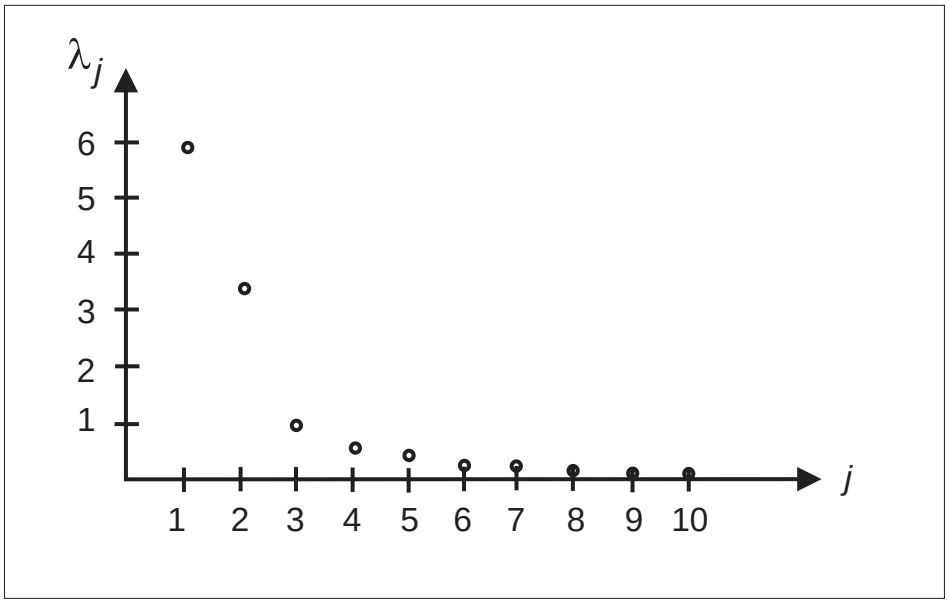

Abbildung 2.2: *Eigenwert-Diagramm - Plot der Eigenwerte λj in Abhängigkeit von den Variablen j eines hypothetischen Datensatzes [88]*

Abb. 2.2 zeigt, daß sich die zuvor erwähnte Asymptotik erst vom dritten Eigenwert an entwickelt. Die gedachte Verbindungskurve ist an dieser Stelle durch einen deutlichen Knick gekennzeichnet. Dies deutet - wie schon beim Eigenwert-Eins-Kriterium - auf eine 2-Komponentenlösung hin.

In der Chemie findet die Hauptkomponentenanalyse auf verschiedenen Feldern (u.a. Umweltchemie, Biochemie oder Geochemie) breite Anwendung. Fragestellungen aus dem Bereich der Lebensmittelchemie nehmen eine herausragende Stellung ein. So wurden z.B. ähnliche Sorten einer Klasse von Nahrungsmitteln mit Hilfe der PCA unterschieden oder es wurden Produkte analysiert, die bei Herstellung bzw. Lagerung unterschiedlichen Bedingungen unterworfen waren. Exemplarisch sei an dieser Stelle auf Untersuchungen von Fisch [99], Kuh-Milch [100] oder Oliven-Öl [101] verwiesen. Im Bereich der Umweltchemie wurden z.B. Schwermetallmuster in Abwässern [102] klassifiziert oder Belastungsparameter in Muscheln bestimmt [103].