

2 Grundlagen

2.1 Spektroskopie

In der Spektroskopie wird die Wechselwirkung der elektromagnetischen Strahlung mit Atomen oder Molekülen meßtechnisch genutzt. Beobachtet wird die Strahlungsabsorption oder -emission einer Probe in Abhängigkeit von der Wellenlänge oder deren reziproken Wert der Wellenzahl (cm^{-1}). In dieser Arbeit werden zwei verschiedene Techniken genutzt: die NIR-Spektroskopie und die Raman-Spektroskopie.

Bei der NIR-Spektroskopie wird die Absorption elektromagnetischer Strahlung im Nahen Infrarot (ca. 13000 cm^{-1} bis 4000 cm^{-1}) gemessen. Im Mittleren Infrarot (MIR) werden im wesentlichen Grundschwingungen in Form von Deformations- und Valenzschwingungen von Molekülen angeregt, im energiereicheren Nahen Infrarot Kombinationen und Obertöne dieser Schwingungen ^[15]. In Abbildung 1 (links) sind die Anregungen einer Schwingung aus dem Grundzustand in den ersten angeregten Zustand (Grundschwingung) und in zwei Obertöne dargestellt. NIR-Banden sind breiter und in der Regel nicht so stark konturiert wie im MIR. Im Mittleren Infrarot können die Spektren eher im Hinblick auf Teilstrukturen und funktionelle Gruppen interpretiert werden. Diese strukturelle Information ist im Prinzip auch im NIR enthalten, aber durch die zahlreichen Kombinationsmöglichkeiten der Grundschwingungen zu Oberton- und Kombinationsbanden im allgemeinen wesentlich schwieriger zu entschlüsseln.

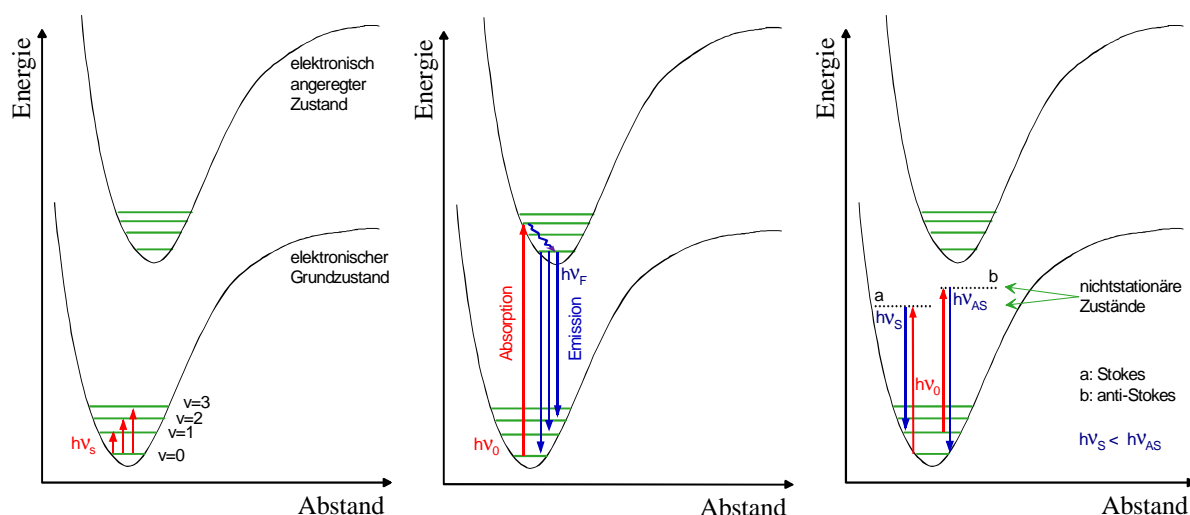


Abb. 1: Termschemata für Infrarot-Absorption (links), Fluoreszenz (mitte) und Raman-Streuung (rechts), Stokes (a) und Anti-Stokes (b).

¹⁵ Günzler, H.; Heise, H.M.; IR-Spektroskopie, eine Einführung; VCH, Weinheim (1996)

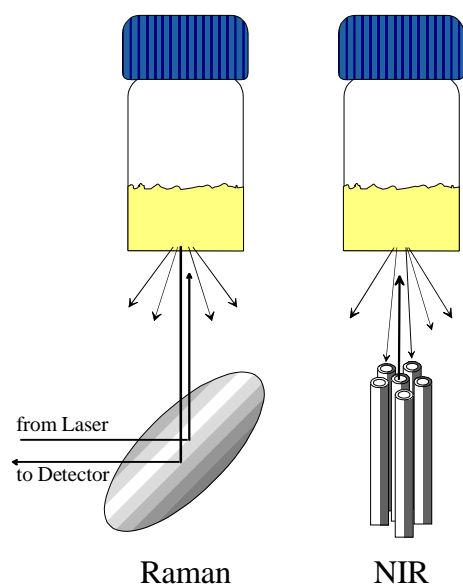


Abb. 2: Probengläser und Meßanordnung.

Im NIR sind die Absorptionen schwächer als im Mittleren Infrarot. Daher kann im NIR mit einer Schichtdicke von ca. 1 – 10 mm in Transmission gemessen werden, im Vergleich zu maximal 0,1 mm im MIR. Auch die Wahl der Küvetten- und Probengefäßmaterialien ist unkritisch. Während Glas und Quarz im MIR zu stark absorbieren, ist deren Absorption im NIR vernachlässigbar. Feststoffe können im Nahen Infrarot mittels diffuser Reflexion praktisch ohne Probenvorbereitung vermessen werden. Dies kann z. B. durch Eintauchen von Glasfasersonden in die Probe erfolgen. In dieser Arbeit wurde eine andere Aufnahmetechnik gewählt. Hierbei wurden Probengläschen wie sie auch in der Chromatographie verwendet werden mit der pulverförmigen Probe

gefüllt und von unten durch den Glasboden des Gefäßes hindurch vermessen (Abb. 2). Bei höheren Anforderungen, wie bei der quantitativen Analyse oder dem in dieser Arbeit beschriebenen Identitätstest sehr ähnlicher Substanzen, ist eine genauere Spezifizierung des Probengefäßes notwendig. Es wurden daher spezielle Glasgefäße mit planem Boden verwendet.

Bei der Raman-Spektroskopie wird die Probe mit dem monochromatischen Licht eines Lasers bestrahlt, wodurch der weiter unten beschriebene Raman-Effekt ausgelöst wird. Parallel kann es jedoch auch zu Fluoreszenz kommen. Ein Molekül geht dabei vom Grundzustand in einen elektronisch angeregten Zustand über (Abb. 1, Mitte). Dieser Vorgang ist unerwünscht, da die aufgrund der Fluoreszenz entstehende Emissionsstrahlung das Raman-Spektrum überlagert. Durch den in der FT-Raman-Spektroskopie üblichen Einsatz eines Neodym-YAG-Lasers können derartige Störungen durch Fluoreszenz in den meisten Fällen vermieden werden. Die Wellenzahl der Laserstrahlung des Nd-YAG-Lasers liegt im Nahen Infrarot ($940 \text{ cm}^{-1} = 1064 \text{ nm}$) und ist damit geringer als die bei der herkömmlichen Raman-Spektroskopie mit Gittergeräten eingesetzten Gaslasern, die meistens im sichtbaren Spektralbereich arbeiten (z. B. Argonlaser bei $19436 \text{ cm}^{-1} = 514,5 \text{ nm}$). Bei vielen Substanzen reicht die kleinere Anregungsenergie des Nd-YAG-Lasers für die elektronische Anregung nicht mehr aus, so daß keine Fluoreszenz mehr auftritt (Abb. 1, rechts)^[16].

¹⁶ Schrader, B.; Raman Spectroscopy of Mineral Oil Products. Part I: NIR/FT-Raman Spectra of Polycyclic Aromatic Hydrocarbons; Appl. Spectr. (1991); Vol. 45, No. 8; S. 1230-1232

Der Raman-Effekt selbst kommt durch inelastische Stöße der Photonen mit den Molekülen der Probe zustande (Raman-Streuung). Der Hauptanteil der Streustrahlung entsteht allerdings durch elastische Stöße, bei denen das Photon mit gleicher Energie in einen beliebigen Raumwinkel reflektiert wird (Rayleigh-Strahlung). Bei dem mit dem Raman-Effekt verbundenen inelastischen Stoß ändert sich die Energie des Photons. Das Molekül geht hierbei vom Schwingungsgrundzustand in einen nichtstationären elektronisch angeregten Zustand über. Kehrt das Molekül in den elektronischen Grundzustand zurück, dann wird ein Photon mit gleicher Energie in den Raum gestreut, wenn die Rückkehr wieder in den Grundzustand erfolgt (Rayleigh-Streuung) oder ein Photon mit geringerer Energie, wenn die Rückkehr auf ein höheres Schwingungsniveau erfolgt (Raman-Streuung). In diesem Fall sind die Raman-Linien gegenüber der Rayleigh-Linie ins Langwellige verschoben (Raman-Verschiebung). Die Differenz entspricht der Frequenz der Molekülschwingung und damit der Lage der entsprechenden Absorptionsbanden im MIR Spektrum (falls die entsprechenden Schwingungen IR-aktiv sind). Die Raman-Linien auf der langwelligen Seite der Rayleigh-Linie werden auch als Stokes-Linien bezeichnet. Anti-Stokes-Linien treten auf, wenn das Molekül von einem angeregten Schwingungszustand aus angeregt wird und dann in den Schwingungsgrundzustand zurückkehrt. Hieraus resultiert eine Ramanverschiebung in Richtung der kurzwelligen Seite der Rayleigh-Linie.

Das Intensitätsverhältnis von Stokes- und anti-Stokes-Linien ist abhängig von dem Verhältnis der Zahl der Moleküle im Grundzustand und angeregtem Schwingungszustand. Da bei Raumtemperatur die meisten Moleküle im Grundzustand vorliegen, ist die Intensität der Stokes-Linien größer als die der anti-Stokes-Linien^[17].

Die Intensität der Rayleigh-Strahlung ändert sich proportional zum Quadrat der Polarisierbarkeit des Moleküls. Die Intensität der Raman-Strahlung hingegen ändert sich proportional zum Quadrat der Änderung der Polarisierbarkeit mit der Normalschwingungskordinate und ist wesentlich geringer als die der Rayleigh-Strahlung. Sie beträgt nur ungefähr 10^{-8} der Ausgangsintensität. Daher sind im allgemeinen hohe Einstrahlleistungen erforderlich, was zu einem Aufheizen der Probe führt und bei empfindlichen Substanzen problematisch sein kann. Eine weitere Konsequenz der geringen Raman-Intensität ist die Tatsache, daß im Gegensatz zur NIR-Absorptionsspektroskopie in Raman-Spektren keine Oberton- und Kombinationsbanden beobachtet werden können.

Ein entscheidender Unterschied zwischen der Infrarot- und Raman-Spektroskopie ist die Auswahl der anregbaren Schwingungen. Bei der Infrarot-Spektroskopie werden Schwingungen angeregt, bei denen sich das Dipolmoment des Moleküls ändert. Für Raman-Aktivität hingegen ist eine Änderung der Polarisierbarkeit Voraussetzung. Betrachtet man z. B. ein CO_2 -Molekül, so gilt aufgrund dessen Symmetrie (Inversionszentrum) ein Alter-

¹⁷ Schrader, B. (Ed.); Infrared and Raman Spectroscopy, Methods and Applications; VCH, Weinheim (1995)

nativverbot^[18]. So bleibt der symmetrischen Streckschwingung (Abb. 3, links) z. B. das Dipolmoment konstant, es ändert sich jedoch die Polarisierbarkeit. Daher ist diese Schwingung ausschließlich Raman-aktiv. Das Umgekehrte gilt für die Deformationsschwingung (Abb. 3, rechts) und die antisymmetrische Streckschwingung^[19].

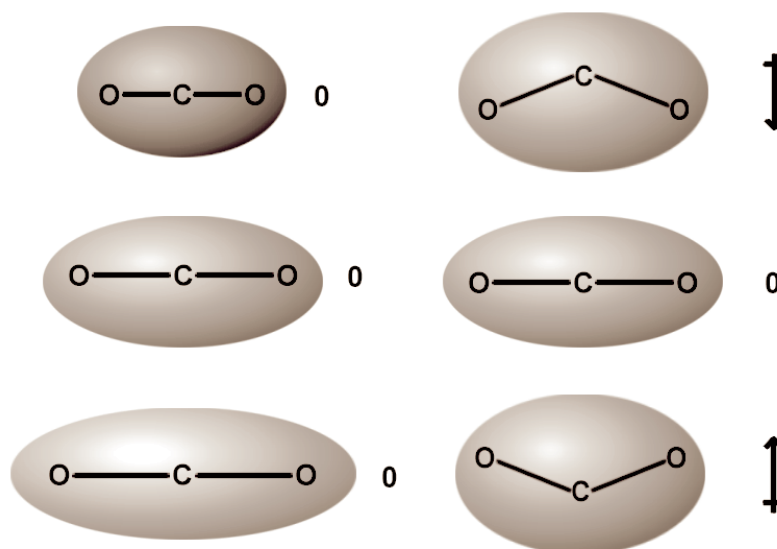


Abb. 3: Änderung der Polarisierbarkeit (links) und des Dipolmomentes (rechts) am Beispiel von CO₂.

Tabelle 1 zeigt eine Zuordnung der Grundschiwingungen und Obertönen zu den unterschiedlichen Symmetrierassen (CO₂ gehört zu der Punktgruppe D_{∞h})

IR- und Raman-Spektren können zwar einige Übereinstimmungen bezüglich der Bandenlagen aufweisen, doch durch die unterschiedlichen Anregungsmechanismen mit unterschiedlichen Auswahlregeln und Bandenintensitäten sind die beiden Techniken als komplementär zueinander anzusehen. Für komplexe Moleküle liefert die Kombination IR/Raman praktisch Informationen über sämtliche Molekülschwingungen. Da bei der Absorptionsspektroskopie von pulverförmigen Proben im Mittleren Infrarot eine Probenvorbereitung wie z. B. die KBr-Preßtechnik notwendig ist, wurde in diesem Projekt, dessen Schwerpunkt auf Routinemessungen liegt, statt dessen die NIR-Spektroskopie eingesetzt. Diese erlaubt das Vermessen pulverförmiger Proben praktisch ohne Probenvorbereitung in diffuser Reflexion.

¹⁸ Molt, K.; Grundlagen und Anwendungen der modernen NIR-Spektroskopie; GIT (1992) 36; S. 107-113;353-362

¹⁹ Brame, E.G. Jr.; Grasselli, J. (Ed.); Infrared and Raman Spectroscopy (Part A); Marcel Dekker Inc.; New York (1976)

Im NIR werden Kombinations- und Oberton-Schwingungsbanden eines Moleküls beobachtet. Tabelle 2 zeigt eine Auswahl der NIR-Banden von CO₂ mit entsprechender Zuordnung, wobei auch hier feste Auswahlregeln gelten.

Tab. 1: Auswahlregeln für die Obertöne von CO₂.

Schwingung	Symmetrierasse	$\tilde{\nu}$ [cm ⁻¹]	Bemerkung
v ₁	Σ_g^+	1286 1388	Raman-aktiv (Sym. Valenzschw.) Fermi-Resonanz
v _{2a} , v _{2b}	Π_u	667	IR-aktiv, \perp (Deformationsschw.)
v ₃	Σ_u^+	2349	IR-aktiv, \parallel (Asym. Valenzschw.)
2v ₂	Σ_g^+	1286 1388	Raman-aktiv
3v ₂	Π_u	1933 2067	IR-aktiv, \perp
4v ₂	Σ_g^+	*	Raman-aktiv
2v ₃	Σ_g^+	*	Raman-aktiv
3v ₃	Σ_u^+	6976	IR-aktiv, \parallel
4v ₃	Σ_g^+	*	Raman-aktiv
5v ₃	Σ_u^+	11497	IR-aktiv, \parallel

* Die gekennzeichneten Obertöne werden wegen zu geringer Intensität nicht beobachtet.

Tab. 2: NIR-Banden von CO₂.

$\tilde{\nu}$ [cm ⁻¹]	λ [nm]	Intensität	Angeregter Zustand		
			v ₁	v ₂	v ₃
4861	2057	mittel	0	4	1
4984	2007	mittel	1	2	1
5109	1957	mittel	2	0	1
6077	1646	schwach	0	6	1
6231	1605	schwach	1	4	1
6351	1575	schwach	2	2	1
6510	1536	schwach	3	0	1
6976	1433	schwach	0	0	3
8193	1221	sehr schwach	0	2	3
8293	1206	sehr schwach	1	0	3
11497	870	sehr schwach	0	0	5
12672	789	sehr schwach	0	2	5
12775	783	sehr schwach	1	0	5

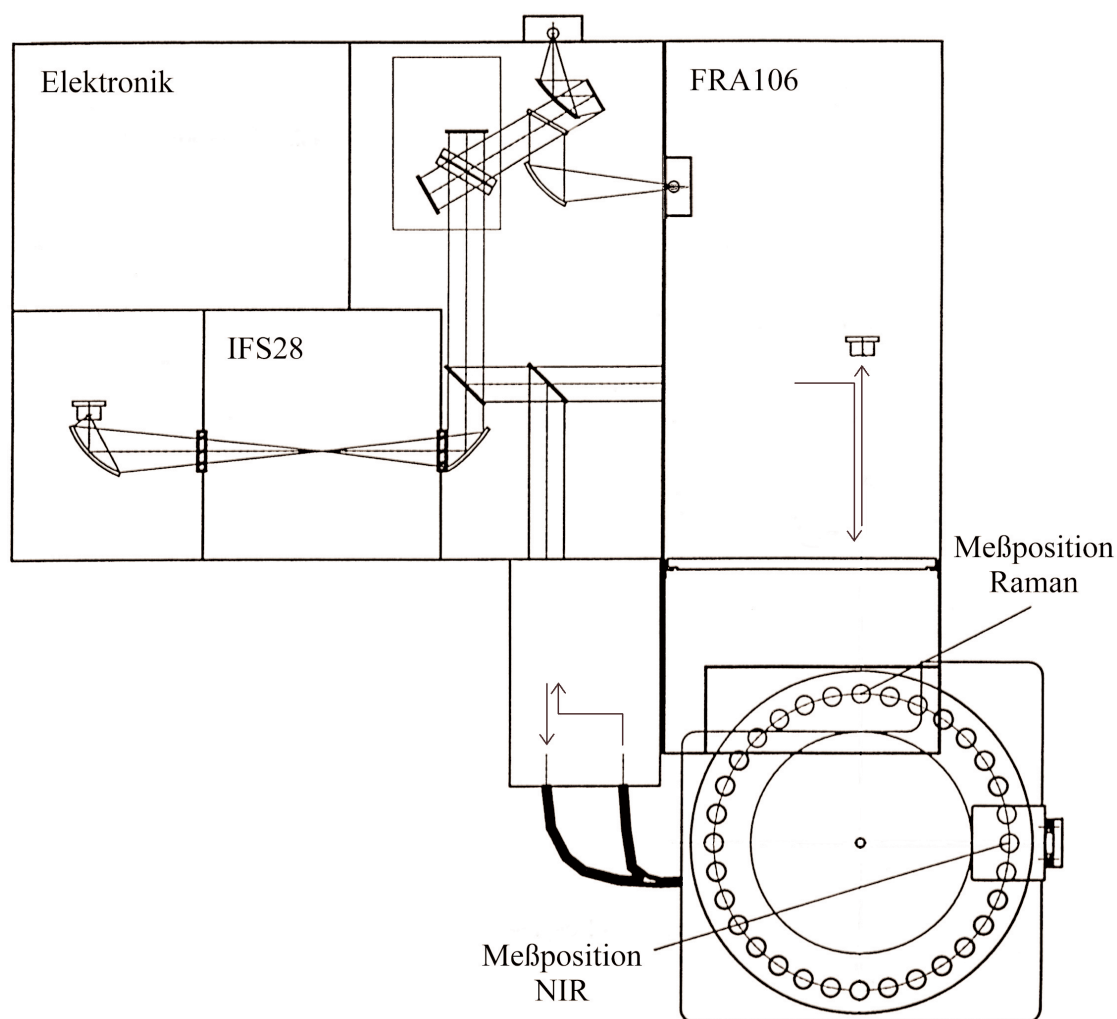


Abb. 4: Schematischer Aufbau des kombinierten NIR/Raman-Spektrometers (Monochromator: mitte, NIR-Detektor: links, Raman-Detektor und Laser: oben rechts, Autosampler: unten rechts).

Für die Aufnahme der in diesem Projekt untersuchten Spektren wurde bei der Firma Merck ein FT-NIR-Spektrometer der Firma Bruker eingesetzt, welches aus einem NIR-Interferometer mit jeweils einem Modul für Absorptionsmessungen und Raman-Emissionsmessungen besteht (Abb. 4). Als Lichtquelle für die in diffuser Reflexion durchgeführten Absorptionsmessungen diente eine Halogenlampe als Anregung für die Raman-Spektroskopie ein Nd-YAG-Laser. Sowohl bei der Aufnahme der NIR- als auch der Raman-Spektren wurden die Proben durch den Boden der Glasfläschchen hindurch vermessen. Desweiteren verfügt das Gerät über einen Autosampler mit einem Probenrad, welches 24 Meßpositionen umfaßt.

2.2 Spektrenbibliotheken

Spektrenbibliotheken sind Datenbanken, in denen eine Vielzahl von Spektren unterschiedlicher Stoffe gesammelt und einheitlich archiviert werden. Diese Bibliotheken können z. B. zur Identifizierung unbekannter Substanzen oder zur Identitätskontrolle von bereits gemessenen Substanzen dienen^[20]. Durch Strukturvergleich ist z. T. auch eine Aussage über neue Verbindungen möglich.

In der vorliegenden Arbeit wurden zwei spektrometrisch unterschiedliche Techniken eingesetzt, nämlich die NIR- und die Raman-Spektroskopie. Entsprechend wurden zwei Teilbibliotheken angelegt, von denen eine nur NIR-Spektren und die andere nur Raman-Spektren enthielt.

In der Praxis wird man aus Zeitgründen eine zu untersuchende Substanz nur mit *einer* spektrometrischen Technik identifizieren. Nur wenn eine eindeutige Identifizierung mit einer Technik alleine nicht möglich ist, wird die andere angewandt. Da eine NIR-Messung im allgemeinen wesentlich schneller verläuft als eine Raman-Messung, wird nach Möglichkeit erstere bevorzugt. Die einzelnen zu identifizierenden Substanzen werden dann in die Teilbibliothek aufgenommen, mit der die besten Resultate für die Identifizierung erzielt werden, wobei zu beachten ist, daß alle der voneinander nicht unterscheidbaren Substanzen in der gleichen Teilbibliothek verbleiben. Die Schnittmenge der in beiden Bibliotheken enthaltenen Artikel soll möglichst klein gehalten werden. Gefordert wird eine „robuste Bibliothek“, die möglichst gute Ergebnisse liefert. Um dies zu erreichen ist eine „Kalibrierung“ der Bibliothek erforderlich, wobei unterschiedliche und zum Teil gegensätzliche Kriterien berücksichtigt werden müssen:

- Es soll eine große Zahl chemisch deutlich unterschiedlicher Artikel möglichst problemlos erkannt werden. Da chemisch deutlich unterschiedliche Substanzen sich auch spektroskopisch deutlich unterscheiden, kann eine relativ grobe Diskriminierung verwendet werden. Dies macht die Bibliothek relativ unempfindlich gegenüber Störungen in Form spektraler Artefakte. Es besteht allerdings die Gefahr der Verwechslung ähnlicher Substanzen.
- Ähnliche Artikel sollen so gut wie möglich unterschieden werden. Dies können Substanzen mit unterschiedlicher technischer oder physikalischer Qualität oder geringen chemischen Unterschieden sein. Aufgrund der nur geringen spektralen Differenzen zwischen derart ähnlichen Artikel, müssen die Grenzen für die spektrometrische Diskriminierung enger gesteckt werden. Dies erhöht die Gefahr einer Störung durch spektrale Artefakte oder Schwankungen (z. B. Verfahrensumstellung, anderer Lieferant, Austausch von Gerätekomponenten). In diesen Fällen kann es dazu kommen, daß die

Probe als „nicht identisch“ befunden wird, obwohl sie noch innerhalb der Spezifikation liegt.

In der Regel wird man versuchen beide der oben genannten Forderungen zu erfüllen, was aber nicht unbedingt in einem Schritt realisiert werden kann. Eine Möglichkeit ist das Zusammenfassen ähnlicher Substanzen in eine Substanzklasse. Die Zuordnung eines gemessenen Spektrums erfolgt zunächst zu einer bestimmten Klasse innerhalb der Gesamtbibliothek. In einem zweiten Schritt wird die gemessene Substanz dann durch schärfere Diskriminierung innerhalb der Klasse endgültig identifiziert.

Es gibt viele Faktoren, die sich auf die Erscheinungsform eines Spektrum auswirken können. Neben äußeren Einflüssen wie der Temperatur spielen Meßparameter wie die Auflösung und die Gerätecharakteristik eine Rolle. Während einige Meßparameter vorgegeben werden können, hat man z. B. auf die Detektorcharakteristik keinen Einfluß. Der Austausch einzelner Komponenten des Gerätes könnte im schlimmsten Fall dazu führen, daß die komplette Bibliothek neu vermessen werden muß. Um dies zu vermeiden werden in regelmäßigen Abständen Kalibrationsstandards (z. B. in Form wohldefinierter optischer Glasfilter oder Polymerfolien) zu vermessen, die zur Kontrolle des Gerätes dienen. Durch den Vergleich der Spektren dieser Kalibrationsstandards mit Sollwerten besteht die Möglichkeit, Geräteinflüsse zu minimieren. Durch entsprechende rechnerische Korrektur bzw. Normierungsverfahren ist dann in manchen Fällen sogar ein Transfer der Bibliotheken von einem Gerät auf andere Geräte möglich. So können Bibliotheken in einem zentralen Labor gepflegt und an verschiedene Standorte weitergeleitet werden. Solche Ansätze für Kalibrationstransfers sind bisher hauptsächlich für die quantitative NIR-Analyse beschrieben worden^{[21][22][23][24]}. Aber auch auf dem Gebiet der qualitativen Analyse und der Identitätskontrolle finden derzeit intensive Bemühungen statt^[25].

²⁰ Viktorin, M.; Spezielle FTIR-Techniken in der Anwendung; LaborPraxis (1995) 11; S.80-86

²¹ Bouveresse, E.; Massart, D.L.; Dardenne, P.; Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardisation samples; Anal. Chim. Acta (1994) 297, Issue: 3; S. 405-416

²² Sjöblom, J.; Svensson, O.; Josefson, M.; Kullberg, H.; Wold, S.; An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra; Chem.Lab.Sys (1998) 44; S. 229-244

²³ Quang, H. N.; Jouan, M.; Dao, N.Q. ; A simplified calibration model of spectral data for quantitative analysis with different Raman spectrometers; Anal. Chim. Acta (1998) 379; S. 159-167

²⁴ Lin, J.; Lo, S.-C.; Brown, C.W.; Calibration transfer from a scanning near-IR spectrophotometer to a FT-near-IR spectrophotometer; Anal. Chim. Acta (1997) 349, Issue: 1-3; S. 263-269

²⁵ Siesler, H.W.; NIR in der pharmazeutischen Identitätskontrolle: Gerätevergleich und Kalibrationstransfer: Gerätevergleich und Kalibrationstransfer; Vortrag Internationaler NIR-Kongreß: Pharmazeutische Anwendungen der NIR-Spektroskopie; 23.-24.4.97 Berlin

Der Schwerpunkt des vorliegenden Projektes lag jedoch nicht auf der Bearbeitung dieser Problematik, die hier zwar durchaus eine Rolle spielt, aber aus folgenden zwei Gründen etwas entschärft ist. Erstens wird nur ein bestimmter wohldefinierter Gerätetyp einer bestimmten Firma verwendet. Zweitens wird die interferometrische Fourier-Transform-Technik eingesetzt. Diese liefert eine sehr gute optische Auflösung, so daß man davon ausgehen kann, daß die Linienformen der Spektren naturgetreu wiedergegeben werden. Ferner garantiert die interne Laserreferenz dieser Geräte eine gute Wellenzahlgenauigkeit und -reproduzierbarkeit.

Spektrenbibliotheken können also einerseits eine große Bandbreite von Substanzen, die sich chemisch deutlich unterscheiden, andererseits aber auch viele Artikel umfassen, deren spektrale Unterschiede nur sehr gering sind. Letzteres kann daran liegen, daß die entsprechenden Substanzen sich chemisch geringfügig oder nur physikalisch z. B. durch unterschiedliche Korngrößen unterscheiden. Da bei Vermessung in diffuser Reflexion die technologisch-physikalischen Eigenschaften einen Einfluß auf die NIR-Spektren besitzen, können sich bei verschiedenen Chargen einer bestimmten Substanz Schwankungen ergeben, die eine Abgrenzung gegenüber anderen aber ähnlichen Substanzen erschwert. Um daher eindeutige Diskriminierungskriterien aufstellen zu können, sollte von jeder Substanz in der Bibliothek die größtmögliche Varianz enthalten sein, weshalb von jeder Substanz unterschiedliche Chargen aufgenommen werden. Die Pflege der Bibliothek umfaßt eine Überprüfung der langfristigen Richtigkeit der Such- und Identifizierungsergebnisse sowie die Aufnahme neuer Spektren, wenn es zu falschen Zuordnungen kommen sollte. Hierbei ist allerdings in jedem Fall zuvor sorgfältig die Ursache zu klären. Nichtkonformität kann auf Unterschiede in der chemischen Reinheit (Wassergehalt) oder unterschiedliche physikalische Eigenschaften (Partikelgröße und -verteilung, Polymorphie) zurückgeführt werden^[26]. Selbstverständlich sind die Meßbedingungen so zu optimieren, daß spektrale Artefakte ausgeschlossen werden können.

Nachdem alle Spektren aufgenommen worden sind und in Form von Rohdaten vorliegen, wird die Bibliothek kalibriert. Da der Begriff Kalibration sonst eher im Zusammenhang einer quantitativen Analyse verwendet wird, soll hier kurz definiert werden, was unter einer Bibliothekskalibration im Rahmen der Ausarbeitung eines Systems für die Identitätskontrolle zu verstehen ist:

Der Vorgang der Kalibration umfaßt im weitesten Sinne alle Maßnahmen, die das Bibliothekssystem in die Lage versetzen zu entscheiden, ob kleine spektrale Unterschiede zwischen Probe und Referenz echt, d. h. auf unterschiedliche Produkte oder nur auf kleinere

²⁶ Plugge, C. Van der Vlies, C.; The use of near infrared spectroscopy in the quality control laboratory of the pharmaceutical industry; J. Pharm.biom.anal. (1992) 10, Nr.10/12; S. 797-803

Schwankungen zwischen verschiedenen Chargen oder mangelnde Reproduzierbarkeit bei der Messung zurückzuführen sind.

Bei einer Bibliothekskalibration werden mehrere Schritte durchlaufen (Abb. 5). In der Regel wird eine Datenvorbehandlung durchgeführt, um Störeffekte wie Basislinieneffekte, Rauschen oder Intensitätsunterschiede zwischen den Spektren zu eliminieren. Je nach Anwendungsgebiet werden oftmals die Vektornormierung oder Berechnung von Derivativspektren mit unterschiedlicher Zahl von Filterpunkten eingesetzt^[27]. Eine Datenreduktion kann durch Umwandlung der Spektren in Binärstrings oder Verminderung der digitalen Auflösung erreicht werden.

Im nächsten Schritt kann eine Transformation der Spektren in einen anderen Datenraum durchgeführt werden. Sehr weit verbreitet ist die Principal Component Analysis (PCA). Rauschen oder andere Minoritätskomponenten schlagen sich hierbei in den höheren Faktoren nieder, die dann für die Kalibration nicht berücksichtigt werden. Über die Identitätsprüfung hinaus kann die PCA auch zur Analyse von Gemischen innerhalb einer Bibliothek angewendet werden^[28].

Eine neuere Methode ist die Wavelet-Transformation, bei der durch Variation verschiedener Parameter ebenfalls Störeffekte entfernt werden können.

Um die Leistungsfähigkeit der Bibliothek im Hinblick auf die Diskriminierung unterschiedlicher Artikel zu überprüfen, werden von den vorbehandelten Spektren nun die jeweiligen Distanzen der Einzelspektren einer Substanz zu deren Mittelwertspektrum und die Distanz der Mittelwertspektren aller Substanzen untereinander berechnet. Von den Einzelspektren werden nur diejenigen in die Bibliothek aufgenommen, die nicht zu weit vom zugehörigen Mittelwertspektrum entfernt sind, d. h. Ausreißer werden entfernt.

Nach Kalibration und endgültiger Fertigstellung der Bibliothek muß noch einmal getestet werden, ob alle enthaltenen Artikel eindeutig identifiziert werden können. Hierbei unterscheidet man zwischen einer internen und externen Validierung. Für die interne Validierung werden ausschließlich die in der Bibliothek befindlichen Spektren herangezogen. Die externe Validierung wird mit Spektren bzw. Proben durchgeführt, die unabhängig von der Bibliothek aufgenommen wurden. Bei der internen Validierung wird für jede Substanz, deren Mittelwertspektrum mit allen anderen Mittelwertspektren verglichen und geprüft, ob die Distanz zu einer eindeutigen Diskriminierung ausreicht. Zusätzlich kann ein entspre-

²⁷ Blanco, M.; Coello, J.; Iturriaga, H.; MasPOCH, S.; Pezuela, C.de la; Effect of Data Preprocessing Methods in Near-Infrared Diffuse Reflectance Spectroscopy for the Determination of the Active Compound in a Pharmaceutical Preparation; *Appl. Spectr.* (1997) 2; S. 240-246

²⁸ Chen, C.-S.; Li, Y.; Brown, C.W.; Searching a mid-infrared spectral library of solids and liquids with spectra of mixtures; *Vibr. Spec.* (1997); S. 9-17

chender Test aller Einzelspektren gegen alle Mittelwertspektren durchgeführt werden. Bei der externen Validierung wird das Spektrum der für die Validierung verwendeten Probe mit allen Mittelwertspektren verglichen. Der Gesamtablauf einer Bibliothekserstellung ist in Abbildung 5 schematisch dargestellt.

Ziel der Bibliothekserstellung ist eine eindeutige und richtige Identifizierung jedes einzelnen Spektrums bei einer späteren Analyse, die im Rahmen der hier vorliegenden Problemstellung als Test auf Identität verstanden werden kann. Hierbei werden folgende Definitionen benutzt:

- Wahr negativ: Das Spektrum der Probe, von der angenommen wird, daß es sich um die Substanz X handelt, wird dem Bibliotheksspektrum der Substanz X richtig zugeordnet. Dies bedeutet, daß die Probe unauffällig bzw. in Ordnung ist.
- Wahr positiv: Das Spektrum der Probe, von der angenommen wird, daß es sich um eine Substanz X handelt, stimmt nicht mit dem Bibliotheksspektrum von X überein. Die Probe ist auffällig bzw. nicht in Ordnung. Möglicherweise wird eine Übereinstimmung mit einer anderen Substanz Y festgestellt, dann ist von einer Verwechslung auszugehen.
- Falsch negativ: Es wird die Aussage getroffen, die unter *wahr negativ* beschrieben ist. Es stellt sich aber heraus, daß diese Aussage nicht zutrifft, also eine falsche Analyse durchgeführt wurde. Die Probe ist in Wirklichkeit nicht in Ordnung.
- Falsch positiv: Es wird die Aussage getroffen, die unter *wahr positiv* beschrieben ist. Es stellt sich aber heraus, daß diese Aussage nicht zutrifft, also eine falsche Analyse durchgeführt wurde. Die Probe ist also in Wirklichkeit in Ordnung.

Bei der Beurteilung der Größe einer Bibliothek orientieren sich Lo und Brown an dem Verhältnis von Spektrenzahl und enthaltener Information, welche vereinfacht durch die Anzahl der Datenpunkte wiedergegeben wird^[29]:

- kleine Bibliotheken: Alle Spektren sind linear unabhängig.

²⁹ Lo, S.C.; Brown, C.W.; Infrared Spectral Search for Mixtures in Medium-Size Libraries; Appl. Spectr. (1991) Vol. 45, Nr. 10; S. 1621-1627

- mittlere Bibliotheken: Die Spektren sind nicht linear unabhängig, die Anzahl der Spektren ist kleiner als die Anzahl der Datenpunkte eines Spektrums.
- große Bibliotheken: Die Spektren sind nicht linear unabhängig, die Anzahl der Spektren ist größer als die Anzahl der Datenpunkte eines Spektrums.

Nach dieser Definition ist die in dieser Arbeit erstellte Testbibliothek als klein, die derzeit in der Praxis eingesetzte Bibliothek (ca. 1000 Artikel) als mittelgroß einzustufen.

Aufgrund der Komplexität von NIR-Spektren (Überlagerung mehrerer Banden, keine eindeutige Zuordnung von Banden) können herkömmliche aus dem MIR bekannte Bibliotheks-Suchmethoden oftmals nicht mehr angewandt werden. So wird anstelle einzelner Banden meist das gesamte Spektrum ausgewertet. Die hierfür anzuwendenden mathematischen Modelle werden als Chemometrie bezeichnet^[2].

Abhängig von der Problemstellung können hierbei durchaus unterschiedliche Methoden eingesetzt werden. In dieser Arbeit wurde zunächst auf die von der OPUS-Software (Bruker) vorgegebene Distanzberechnung und die Möglichkeit der Faktorisierung zurückgegriffen.

In der Literatur werden für die Kalibration von Bibliotheken Mustererkennungs-Techniken (Pattern recognition) wie Clusteranalyse, neuronale Netze^{[30][31][32]} oder Fuzzy-Logic^{[33][34]} und SIMCA (Soft Independent Modelling of Class Analogy) beschrieben. SIMCA bietet den Vorteil, daß für jede Klasse ein eigenes Modell erstellt wird, was jedoch bei größeren Bibliotheken mit einem großen Rechenaufwand verbunden ist. Neuronale Netze können auch nichtlineare Zusammenhänge erkennen. Die in dieser Arbeit angewendeten Verfahren entsprechen am ehesten denen der Clusteranalyse.

³⁰ Schulz, H.; Derrick, M.; Stulik, D.; Simple encoding of infrared spectra for pattern recognition Part 2. Neural network approach using back-propagation and associative Hopfield memory; *Anal. Chim. Acta* (1995) 316; S. 145-159

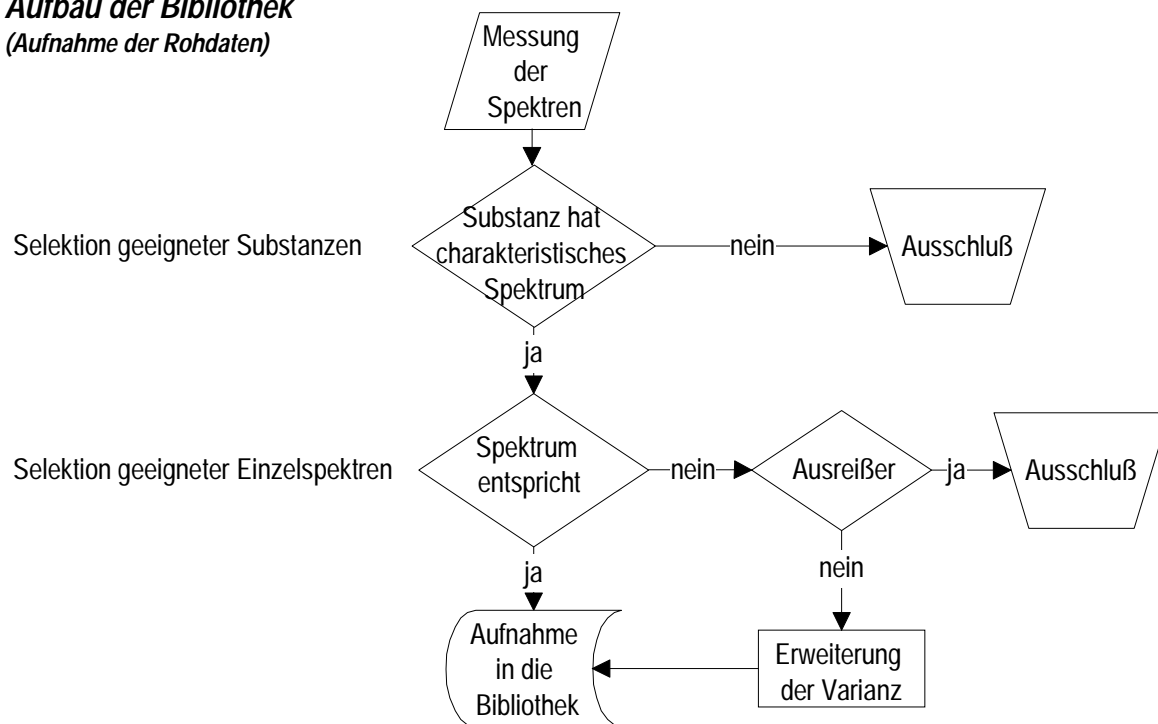
³¹ Wu, W.; Walczak, B.; Massart, D.L.; Heurding, S.; Erni, F.; Last, I.R.; Prebble, K.A.; Artificial neural networks in classification of NIR spectral data: Design of the training set; *Chem.Lab.Sys* (1998) 33, Issue: 1; S. 35-46

³² Klawun, C.; Wilkins, C.L.; Neural Network Assisted Rapid Screening of Large Infrared Spectral databases; *Anal. Chem.* (1995) 2; S. 374-378

³³ Linusson, A.; Wold, S.; Nordén, B.; Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry; *Chem.Lab.Sys* (1997) 44; S. 213-227

³⁴ Hörchner, U.; Otto, M.; Identification of UV/ VIS-spectra based on a fuzzy function; *Fres. J. anal. chem.* (1992) 344; S. 217-219

Aufbau der Bibliothek (Aufnahme der Rohdaten)



Kalibration der Bibliothek

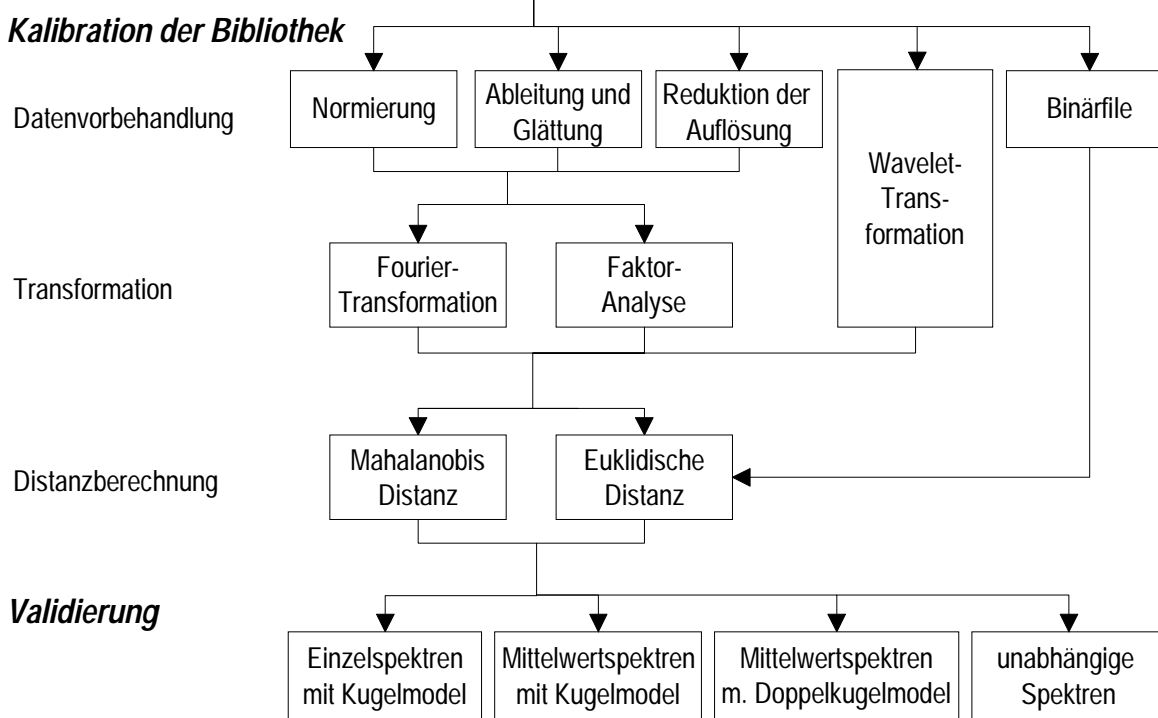


Abb. 5: Ablaufschema einer Bibliothekserstellung.

2.3 Datenvorbehandlung

Als Datenvorbehandlung wird hier die Bearbeitung der Spektren u. a. zur Verbesserung des Signal-zu-Rausch-Verhältnisses (*Signal to noise*, S/N) oder zur Verringerung der Zahl der Datenpunkte eingesetzt. Bei diesen Vorbehandlungen wird das ursprüngliche Koordinatensystem beibehalten. Mathematische Modelle, die darüber hinausgehen, werden in Abschnitt 2.4 behandelt.

2.3.1 Derivativspektren

Das Berechnen der ersten oder zweiten Ableitung eines Spektrums wird als Derivativspektroskopie bezeichnet. In der IR/NIR-Spektroskopie werden aufgrund des im Vergleich zu anderen Spektralbereichen (z. B. UV/VIS) eingeschränkten Signal-zu-Rausch-Verhältnisses höhere Ableitungen als die Zweite nicht verwendet.

Ein Bandenmaximum im ursprünglichen Spektrum führt zu einer Nullstelle in der ersten Ableitung. Durch die erste Ableitung der Spektren kann eine über den gesamten Spektralbereich konstante Verschiebung der Basislinie (Offset) eliminiert werden, da deren erste Ableitung gleich Null ist.

Die zweite Ableitung entspricht der Krümmung des Spektrums. Bei der Ableitungsbildung resultieren aus den ursprünglichen Banden hierbei negative Banden geringerer Halbwertsbreite, wodurch eine Erhöhung des Kontrastes bzw. bei überlappenden Banden eine Dekonvolution erreicht wird. Desweiteren kann durch die zweite Ableitung eine lineare Neigung der Basislinie beseitigt werden. Nachteilig ist jedoch eine Verschlechterung des S/N-Verhältnisses, denn das Rauschen nimmt mit jeder Ableitungsbildung zu.

Numerisch kann gleichzeitig die Ableitungsbildung wie auch die Glättung mit sogenannten „Filtern“ erfolgen. Hierzu wird ein Datenpunkt mit einer definierten Anzahl seiner linken und rechten Nachbarn in einem Fenster zusammengefaßt. Der Wert für den ursprünglichen mittleren Datenpunkt wird dann auf Basis einer mit allen Datenpunkten innerhalb des Fensters durchgeführten Rechenoperation ersetzt. Dann wird das Fenster punktweise weitergeschoben und der gleiche Vorgang ausgehend von den Rohdaten jeweils wiederholt. Wenn das Fenster auf diese Weise durch das ganze Spektrum hindurchgeschoben worden ist, ist der Filtervorgang beendet. Ein Beispiel hierfür ist die Glättung verwendete *gleitende Mittelwertbildung*. Die Zahl der Datenpunkte, die das Fenster umfaßt, entscheidet über den Grad der Glättung. Bei zu großen Fenstern, d. h. bei zu starker Glättung, können Details des Spektrums verlorengehen. Zum einen werden eng benachbarte Banden dann nicht mehr richtig aufgelöst, zum anderen nimmt die Signalintensität mit steigender Anzahl der Filterpunkte ab.

Bei dem Savitzky-Golay-Filter werden die Datenpunkte vor der Mittelung gewichtet. Die Gewichtungskoeffizienten können dabei so gewählt werden, daß das Resultat einer Appro-

ximation des Fensterbereiches an ein Polynom n-ten Grades entspricht. Durch die gegenüber der gleitenden Mittelwertbildung bessere Anpassung an die Originaldaten wird eine effektivere Glättung erzielt, weil Rauschen weitgehend entfernt wird, Bandenkonturen aber besser erhalten werden^[2]. Auch Ableitungen werden mit Savitzky-Golay-Filtern berechnet, wobei die Gewichtungskoeffizienten so gewählt werden, daß eine Approximation an die erste oder höhere Ableitung des Polynoms erfolgt.

2.3.2 Normierung

Durch die Normierung von Spektren können verschiedene Effekte erzielt werden. Bei NIR-Spektren haben Parameter wie Korngröße und Schüttdichte Einfluß auf die Basislinie. Verschiebungen können durch eine geeignete Normierung entfernt werden. Raman-Spektren weisen oftmals stark schwankende Peakintensitäten auf. Durch die Normierung wird z. B. die Intensität der intensivsten Bande festgelegt, wobei die Intensitätsverhältnisse zwischen den verschiedenen Banden des Spektrums erhalten bleiben.

Es gibt eine Reihe verschiedener Ansätze der Spektrennormierung, bei denen Bandenintensitäten oder Flächen berücksichtigt werden. Eine einfache Methode ist beispielsweise eine Minimum-Maximum-Normierung, bei der die geringste Absorption gleich 0 und die Bande mit der größten Absorption gleich 1 gesetzt werden. Schwankungen einer sehr intensiven Bande fallen hierbei jedoch stark ins Gewicht und beeinflussen das gesamte Spektrum. Geringer sind solche Einflüsse bei einer Normierung der Fläche. Die Summe über alle Datenpunkte e_i wird hierbei mittels Normierungsfaktor F gleich 1 gesetzt:

$$\frac{1}{F} \sum_{i=1}^n y_i = 1 \quad \text{Gleichung 1}$$

Eine andere Möglichkeit ist die Vektornormierung, die auch bei der OPUS-Software eingesetzt wird. Von einem Spektrum wird der Mittelwert berechnet:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{Gleichung 2}$$

Dieser wird vom Spektrum subtrahiert, wodurch man ein „zentriertes“ Spektrum erhält:

$$z_i = \bar{y}_i - \bar{y} \quad \text{Gleichung 3}$$

Die Vektorlänge dieses Spektrums ergibt sich wie folgt:

$$\|z\| = \sqrt{\sum_{i=1}^n z_i^2} \quad \text{Gleichung 4}$$

Das Spektrum z wird durch diesen Wert geteilt. Dadurch erhält man das Spektrum x :

$$x_i = \frac{z_i}{\|z\|} \quad \text{Gleichung 5}$$

X ist das vektornormierte Spektrum mit der vektoriellen Länge 1. Die Vektoren aller so normierten Spektren liegen auf einem Einheitskreis (Einheitskugel im n -dimensionalen Fall). Abbildung 6 zeigt dies für den zwei 2-dimensionalen Fall. Der Abstand zweier Vektoren (A, B) ist die Länge des Differenz-Vektors ($D_{A,B}$, s. 2.5). Der maximale Abstand der beiden Vektoren entspricht dem doppelten Radius des Einheitskreises, also 2.

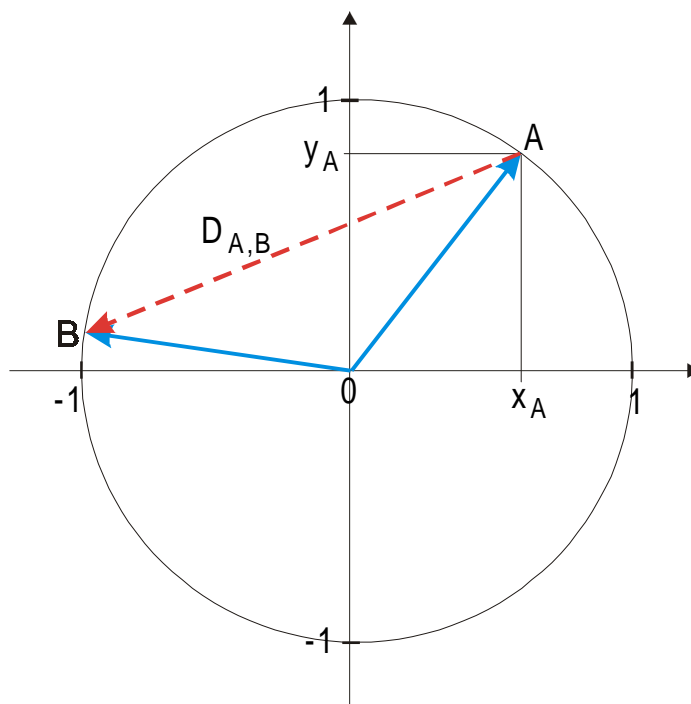


Abb. 6: Distanz $D_{A,B}$ zwischen zwei 2-dimensionalen Vektoren.

2.3.3 Datenkompression

Vor wenigen Jahren war die Speicherkapazität der computergestützten Datenverarbeitung noch stark begrenzt (und teuer). Um größere Bibliotheken erstellen zu können, mußte somit die Datenmenge drastisch reduziert werden, da Speicherplatz und Rechengeschwindigkeit eine viel kritischere Rolle als heute spielten. So wurden z. B. von IR-Spektren meistens lediglich die Bandenlagen und eventuell die Bandenintensitäten gespeichert. Die digitale Auflösung wurde ebenfalls beschränkt (4 Ziffern für die Bandenlage, 2 oder 3 Ziffern für die Intensität). Ein Spektrum war ca. 100 Byte groß (60 Byte Daten, 40 Byte Sub-

stanzname!)^[35]. Bedingt durch die rasante technische Entwicklung ist es nun möglich die gesamten Spektren mit rund 4000 Datenpunkten zu speichern und zu verarbeiten. Bei einer Validierung der Bibliothek mittels Einzelspektren kann hierbei die Rechenzeit auf ca. 1 h anwachsen. Fraglich ist jedoch, ob ein derartiger Aufwand gerechtfertigt ist, wenn mit weniger Information gleichwertige Ergebnisse erzielt werden können.

Eine Datenreduktion kann durch verschiedene Ansätze erreicht werden. Betrachtet man nur den ursprünglichen Datenraum, so kann man z. B. die digitale Auflösung der Spektren verringern oder eine einfachere Kodierung wählen (z. B. Binärstrings). Denkbar ist auch eine Annäherung (Approximation) der Spektren durch einfache Funktionen, deren Parameter stellvertretend für das ursprüngliche Spektrum gespeichert werden.

Unabhängig von der Art der Datenvorbehandlung und der erzielten Datenkompression gibt es eine untere theoretische Grenze, ein Minimum an Information (H_{\min}), welches mit dem Signal (Spektrum) übertragen werden muß, damit eine eindeutige Zuordnung möglich ist. Dieses Minimum, welches durch ein Signal übertragen werden muß, ergibt sich aus dem Logarithmus der Anzahl der Zustände n :

$$H_{\min} = {}^2\log n$$

Gleichung 6

Zur Unterscheidung von 2 Zuständen (z. B. 2 Substanzen) reicht 1 Bit aus, für 4 Zustände 2 Bits usw. Dabei muß man annehmen, daß die Informationen optimal kodiert sind und eine störungsfreie Übertragung vorliegt. Allerdings enthalten Spektren immer ein gewisses Maß an Meßfehler und Rauschen. Durch Übertragung von mehr Informationen als in Gleichung 6 gefordert, erhält man eine Redundanz an Information. Gleiche Informationen sind dann mehrfach in dem Signal kodiert, was zu einer geringeren Fehleranfälligkeit führt^[36].

³⁵ Tanabe, K.; Hiraishi, J.; Saeki, A.; Suzuki, I.; Tasumi, M.; SIRIUS: an efficient infrared spectra search system; *Comp. enh. Spectr.* (1984) 2, No.1; S. 13-15

³⁶ Molt, K.; How safe are NIR-library search systems? Information-theoretical and practical aspects; *Fres. J. anal. chem.* (1997) 359; S. 67-73

2.3.3.1 Approximation

Die Spektren werden in kleinere Teile zerlegt, welche durch Funktionen wiedergegeben werden können^[37]. Das Spektrum wird durch die Parameter der Funktion beschrieben. Es kann eine Datenreduktion und Minderung des Rauschens erreicht werden^[38]. Abhängig sind diese Effekte von der Wahl der Fenstergröße und der Art der verwendeten Funktionen.

2.3.3.2 Bandentabellen

Die Spektren werden nach Absorptionsmaxima durchsucht. Hieraus resultieren Tabellen, in denen die Bandenlagen der Absorptionsmaxima aufgelistet werden. Zusätzlich können Intensität und Bandenform berücksichtigt werden. Dies entspricht in etwa einem visuellen Vergleich der Spektren ohne Computerunterstützung.

Bei Raman-Spektren kann die Lage der relativ schmalen Banden gut bestimmt werden. NIR-Spektren sind jedoch weniger stark strukturiert und weisen breitere Banden auf. Von Meßfehlern und der Auflösung abgesehen, kann es hier zu weiteren Fehlern kommen, welche von dem angewandten Verfahren der Maximumerkennung abhängen (z. B. Bestimmung der Nullstelle bei 1. Ableitung). Bei einer visuellen Auswertung kann dies, abhängig vom Operator, zu erheblichen Unterschieden führen. Durch die computerunterstützte Auswertung können diese Unterschiede vermindert werden, der Operator muß aber letztendlich noch Parameter wie die minimale auszuwertende Bandenintensität festlegen.

2.3.4 Binärstrings

Eine gegenüber einer Bandentabelle noch weitergehende Reduktion der Spektrendaten stellt die Erzeugung von Binärstrings dar. Geeignet hierfür sind die Banden von Raman-Spektren, die in der Regel sehr schmal sind und daher stark vereinfacht als Linien betrachtet werden können. Verzichtet man dann noch auf die Intensitätsinformation, so werden für ein Spektrum nur noch die Bandenlagen abgespeichert. Dies kann in Form eines Binärstrings erfolgen. Zur Erzeugung des letzteren wird das Spektrum in Intervalle eingeteilt. Jedem Intervall entspricht eine Stelle im Binärstring. Die einzelnen Intervalle werden auf Banden untersucht, deren Intensität größer als ein vorgegebener Threshold ist. Tritt in dem jeweiligen Intervall eine Bande auf, so wird an dieser Stelle im Binärstring der Wert des Intervalls auf 1, anderenfalls auf 0 gesetzt (Abb. 7).

³⁷ Nikitin, O.Y.; Gribov, L. A.; Information-Search System with Self-Correction; J. appl. spec. (1994) 60, Nr.1/2; S. 147-150

³⁸ Divis, R.A.; White, R.L.; Deconvoluted Band Representation for Infrared Spectrum Compression; Anal. Chem. (1989) 1; S. 33-37

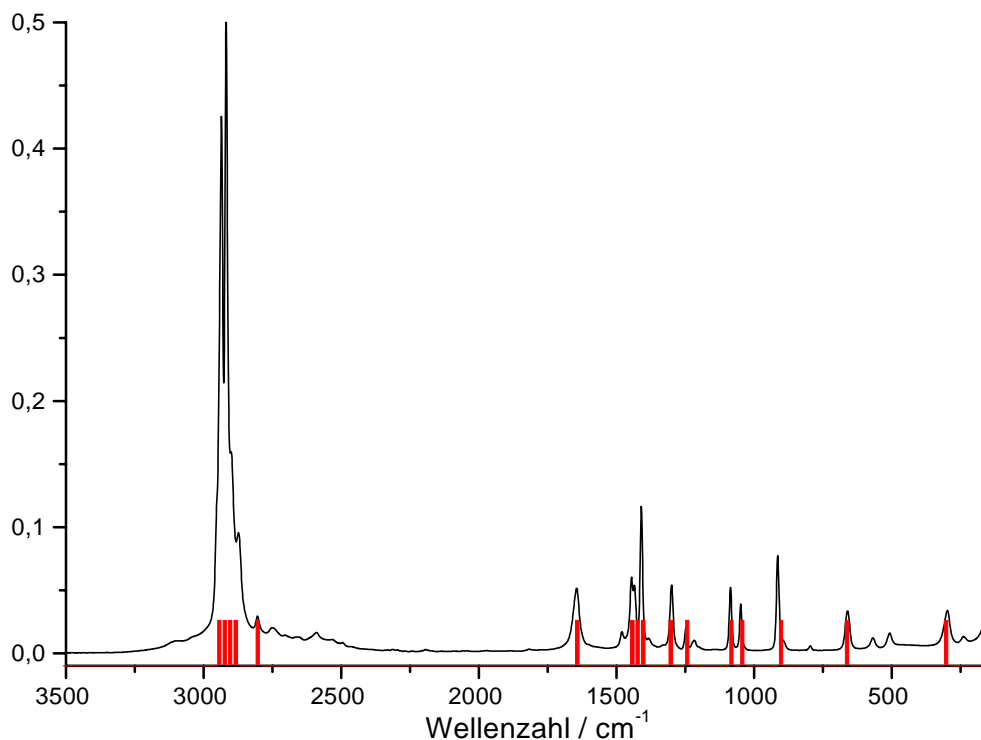


Abb. 7: Ramanspektrum und der resultierende Binärstring als Balkendarstellung ($\Delta\nu=20\text{ cm}^{-1}$).

Diese Art der Kodierung wurde in der Literatur beschrieben, allerdings wurde früher die Kodierung auf Lochkarten von Hand vorgenommen. Die Interpretation der Spektren war jedoch abhängig vom Operator und dadurch ungenau^[39]. Eine genauere Bearbeitung kann mittels Computer erreicht werden, da hier der Algorithmus zur Bestimmung der Bandenlage sowie der Grenzwert der minimalen Bandenintensität vorgegeben werden.

Der Vergleich von binär kodierten Spektren läßt sich sehr einfach mit entsprechenden Operatoren durchführen. Einige gängige Operatoren zeigen Tabelle 3 bis 6. Während man mit dem Und-Operator die Ähnlichkeit zwischen zwei Spektren erhält, läßt sich mit dem Operator „Exklusives Oder“ die Distanz berechnen. Diese ergibt sich als Quersumme

Spektrum	
↓	
Bandentabelle	
↓	
Binär String	
Probenspektrum:	1 0 0 0 1 1 0
Referenzspektrum:	1 0 0 1 0 1 0
Exclusives Oder (XOR):	0 0 0 1 1 0 0
Distanz:	$\Sigma = 2$

Abb. 8: Binärstring-Erstellung und Distanzberechnung.

³⁹ Lowry, Huppler, Anderson; Database development and search algorithms for automated infra-red spectral identification; (1985); S. 235-241

des Ergebnisstrings, bei der Anwendung des Operators auf die einzelnen Bits der beiden Primärstrings x und y (Abb. 8). Während bei dem Und-Operator nur das Vorhandensein von Banden in den beiden Spektren bewertet wird, üben bei dem XOR-Operator die Bereiche, in denen keine Banden auftreten, ebenfalls einen Einfluß auf das Ergebnis aus.

Tab. 3: Binärer Oder-Operator

x	y	$x \vee y$
0	0	0
0	1	1
1	0	1
1	1	1

Tab. 4: Binärer Und-Operator

x	y	$x \wedge y$
0	0	0
0	1	0
1	0	0
1	1	1

Tab. 5: Exklusives Oder

x	y	$x \nabla y$
0	0	0
0	1	1
1	0	1
1	1	0

Tab. 6: Binäres Nicht

x	$\neg x$
0	1
1	0

Die Berechnung der Distanzen erfolgte mittels eines C-Programmes. Da für dieses kein XOR-Operator zur Verfügung stand, wurde Gleichung 7 in Gleichung 8 umgewandelt und die Berechnung entsprechend durchgeführt.

$$D_{X,Y} = \sum_{i=1}^n (X_i \nabla Y_i) \quad \text{Gleichung 7}$$

$$D_{X,Y} = \sum_{i=1}^n \neg(\neg X_i \wedge \neg Y_i) \wedge \neg(X_i \wedge Y_i) \quad \text{Gleichung 8}$$

Dieses Rechenverfahren geht von konstanten Bandenlagen aus, was durch die begrenzte Meßreproduzierbarkeit nur bedingt zutrifft. Besonders problematisch ist dies bei Banden, die am Rand eines Intervalls liegen, da kleine Abweichungen der Bandenlage dann zu einer unterschiedlichen Intervallzuordnung führen können.

Eine mögliche Lösung ist die Einbeziehung der benachbarten Intervalle in den Vergleich der Spektren. So ergibt sich in Abbildung 9 für das Exklusive Oder ein Wert von 0, wenn in einem der drei Intervalle der Referenz (I_{x-1} bis I_{x+1}) eine Bande vorhanden ist.

Eine Alternative wäre, ganz von festen Intervallen abzurücken und für jede Bande eines Bibliotheksspektrums auf der Basis von Wiederholungsmessungen, bei denen auch unterschiedliche Chargen ein und derselben Substanz berücksichtigt werden, die Streuung bezüglich der Lage auf der Wellenzahlachse zu ermitteln. Ähnlich wie in der Fuzzy-Theorie^[2] kann auch eine Wahrscheinlichkeit für die Übereinstimmung berechnet werden (Abb. 10). Beide Ansätze wurden im Rahmen dieser Arbeit nicht weiterverfolgt, d. h. es wird vereinfacht mit festen Lagen für die binär kodierten Banden gearbeitet.

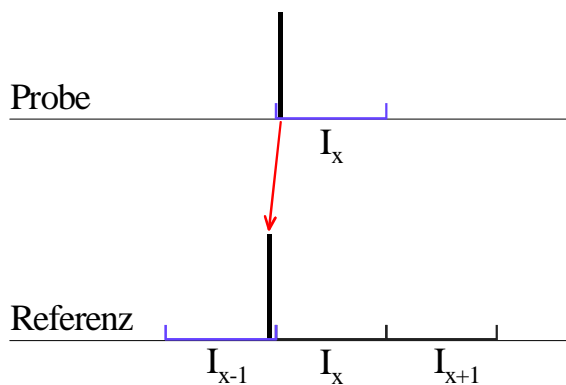


Abb. 9: Kodierungsfehler durch geringfügige Verschiebung der Bandenlage an der Intervallgrenze.

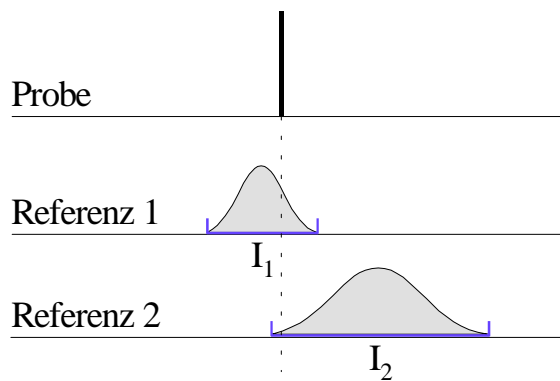


Abb. 10: Variable Intervallgröße, Bestimmung einer Wahrscheinlichkeit für die Übereinstimmung.

2.4 Datentransformation

In vielen Fällen kann es nützlich sein, die Daten nicht im ursprünglichen Datenraum zu betrachten, sondern vorher eine Transformation in eine andere Domäne, d. h. in ein anderes Koordinatensystem vorzunehmen. Dies soll im folgenden an zwei Beispielen gezeigt werden. An einer Standardmethode, nämlich der Hauptkomponentenanalyse und an einem neueren, im Arbeitskreis in jüngster Zeit intensiv untersuchten Verfahren, der Wavelet-Analyse.

2.4.1 Hauptkomponentenanalyse

Bei der Hauptkomponentenanalyse (Principal Component Analysis, PCA) werden die Spektren als Vektoren in einem n-dimensionalen Raum betrachtet (Abb. 11). Die Datenmatrix X hat hierbei die Dimension $i \times n$. Die Spektren bilden die Spalten und die Zahl der Datenpunkte pro Spektrum entspricht der Zeilenzahl:

$$X = \begin{pmatrix} x_{1,1} & x_{2,1} & \dots & x_{i,1} \\ x_{1,2} & x_{2,2} & \dots & x_{i,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \dots & x_{i,n} \end{pmatrix}$$

Gleichung 9

wobei: $i =$ Anzahl der Spektren Spektrum i : $x_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{pmatrix}$
 $n =$ Anzahl der Datenpunkte

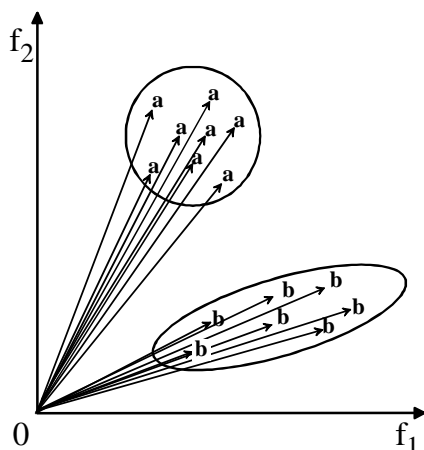


Abb. 11: Darstellung von zweidimensionalen Daten als Vektoren.

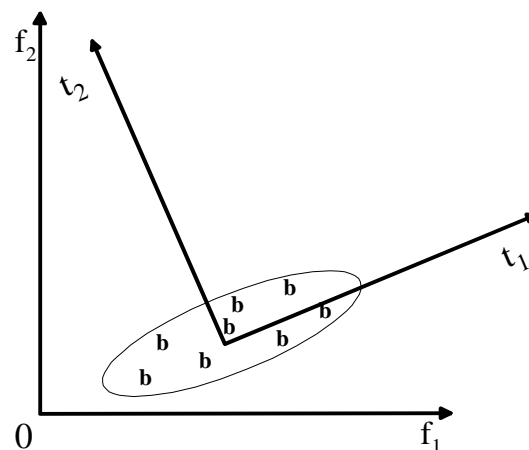


Abb. 12: Transformation eines zweidimensionalen Bezugssystems.

Bei der PCA wird eine Transformation vom ursprünglichen Koordinatensystem F in das Koordinatensystem T (Abb. 12) durchgeführt^[40]. Ziel der Transformation ist eine Anordnung der Daten, bei der die 1. Koordinate (erster Faktor) so liegt, daß mit ihr ein maximaler Anteil der Varianz der Daten beschrieben wird. Jede weitere Koordinate, die orthogonal zu allen anderen sein muß, wird so gewählt, daß sie einen möglichst großen Anteil der Restvarianz beschreibt^[41]. Zur Zerlegung der Datenmatrix können die Singulärwertzerlegung oder Eigenwertzerlegung eingesetzt werden^[2].

$$\begin{array}{c} \text{i} \\ \boxed{\text{T}} \\ \text{n} \end{array} = \begin{array}{c} \text{i} \\ \boxed{\text{X}} \\ \text{n} \end{array} \begin{array}{c} \text{i} \\ \boxed{\text{L}} \\ \text{i} \end{array} \quad \text{Gleichung 10}$$

Die Faktoren t sind orthogonal zueinander und können gemäß Gleichung 11 als Linearkombination der Originalspektren dargestellt werden^[42]:

$$t_{11} = x_{11} l_{11} + x_{12} l_{21} + x_{13} l_{31} + \dots + x_{1n} l_{n1} \quad \text{Gleichung 11}$$

Umgekehrt kann die Datenmatrix X nun durch die zwei Matrizen T und L^T wiedergegeben werden (Gl. 12). Spalten der Matrix L entsprechen den Zeilen der Matrix L^T . Die Zeilenvektoren λ_i von L^T werden als Eigenvektoren oder Principal Components bezeichnet.

$$\begin{array}{c} \text{i} \\ \boxed{\text{X}} \\ \text{n} \end{array} = \begin{array}{c} \text{i} \\ \boxed{\text{T}} \\ \text{n} \end{array} \begin{array}{c} \text{i} \\ \boxed{\text{L}^T} \\ \text{i} \end{array} \quad \text{Gleichung 12}$$

oder: $X = T L^T$

⁴⁰ Malinowski, E.R.; Factor analysis in chemistry; J. Wiley & Sons Ltd.; 2. Aufl. (1991)

⁴¹ Gnanadesikan, R.; Methods for Statistical Data Analysis of Multivariate Observations; J. Wiley & Sons Ltd. (1977), S. 7 ff.

⁴² Martens, H.; Naes, T.; Multivariate Calibration; Wiley & Sons Ltd., Chichester (1989)

bzw.:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} t_{1,1} l_{1,1}^T & + t_{1,2} l_{2,1}^T & + \dots & + t_{1,i} l_{i,1}^T \\ t_{2,1} l_{1,1}^T & + t_{2,2} l_{2,1}^T & + \dots & + t_{2,i} l_{i,1}^T \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} l_{1,1}^T & + t_{n,2} l_{2,1}^T & + \dots & + t_{n,i} l_{i,1}^T \end{pmatrix} \quad \text{Gleichung 13}$$

Das neue, *transformierte* Bezugssystem T hat zunächst die gleichen Dimensionen wie die Datenmatrix. Die Matrix T wird als Faktormatrix, die Spalten als Faktoren oder Eigenspektren bezeichnet. Die Eigenspektren sind orthogonal zueinander. Die Zeilen der transponierten Eigenvektormatrix L^T sind die *Principal Components*.

$$\begin{array}{c} \text{i} \\ \boxed{\text{X}} \\ \text{n} \end{array} = \begin{array}{c} \text{k} \\ \boxed{\text{T}} \\ \text{n} \end{array} \quad \begin{array}{c} \text{i} \\ \boxed{\text{L}^T} \\ \text{k} \end{array} \quad \text{Gleichung 14}$$

Die Faktoren, die die signifikanten Informationen des Datensatzes enthalten, werden als primäre Faktoren bezeichnet. Die sekundären Faktoren beschreiben hauptsächlich Rauschen. Werden durch eine bestimmte Anzahl von Hauptkomponenten die Daten hinreichend beschrieben, kann auf die restlichen Faktoren (sekundären Faktoren) verzichtet und somit eine Reduzierung von n auf k Faktoren erreicht werden (Gl. 14).

Bei der PCA können selektiv Faktoren entfernt werden, die nur Rauschen oder geringe Informationsanteile enthalten. Im Prinzip können statistische Verfahren zur Selektion der Faktoren eingesetzt werden. Bei der OPUS-Software muß die Auswahl allerdings von Hand erfolgen^[43]. Übertrifft die Anzahl der Bibliotheksspektren, die der Faktoranalyse unterworfen werden, die Anzahl der Datenpunkte pro Spektrum, so sind die Spektren nicht mehr linear unabhängig. Das kann dazu führen, daß nicht mehr alle Substanzen eindeutig identifiziert werden können und zu Substanzklassen zusammengefaßt werden müssen^[44].

⁴³ Elbergali, A.; Nygren, J.; Kubista, M.; An automated procedure to predict the number of components in spectroscopic data; Anal. Chim. Acta (1999) 379; S. 143-158

⁴⁴ Lo, S.C.; Brown, C.W.; Infrared Spectral Search for Mixtures in Large-Size Libraries; Appl. Spectr. (1991) 10; S. 1628-1632

In der vorliegenden Arbeit gingen die Mittelwertspektren aller Bibliothekssubstanzen in die Faktoranalyse ein. Die Zahl der resultierenden Faktoren entspricht dann der Zahl der Substanzen. Im Rahmen der Bibliothekskalibration wurde anschließend überprüft, mit wie vielen primären Faktoren eine optimale Diskriminierung der Bibliothekssubstanzen erreicht werden kann.

2.4.2 Wavelet-Transformation

Bei der Wavelet-Transformation wird ein Signal in seine Frequenzanteile zerlegt, wobei im Gegensatz zur Fourier-Transformation die Lokalisierung der Anteile in der Zeitdomäne erhalten bleibt. Das Ergebnis der Wavelet-Transformation ist die Zerlegung des Signals in Frequenzbänder. In dem untersten Band befinden sich die niedrigsten Frequenzen wie Basislinienanteile, im obersten Band die höchsten Frequenzen, die hauptsächlich Rauschanteile repräsentieren. Nach der Zerlegung erhält man für jedes Spektrum einen Satz von Wavelet-Koeffizienten. Jeder Koeffizient repräsentiert einen Signalanteil mit definiertem Frequenzbereich und definierter Lokalisierung im Spektrum. Im Gegensatz zur gefenster-ten Fourier-Transformation ist die Auflösung hier nicht konstant, sondern unterschiedlich für die einzelnen Bänder. Aufgrund der Heisenbergschen Unschärferelation kann das Produkt aus Auflösung der Zeit- und der Frequenzdomäne nicht beliebig klein werden, d. h. bei Verbesserung der Auflösung der Zeitdomäne verschlechtert sich die Auflösung in der Frequenzdomäne und umgekehrt. Die Wavelet-Transformation arbeitet selbstadaptiv, so daß ein Signal in verschiedenen Auflösungen in beiden Domänen analysiert wird^{[45][46][47]}. Die Basisfunktionen dieser Transformation sind sogenannte Wavelets, welche im Gegensatz zur unendlichen Sinus- und Cosinus-Funktion begrenzt sind und außerhalb eines endlichen Intervalls einen Wert nahe oder gleich Null annehmen^[48]. Es gibt eine Reihe verschiedener Wavelets (Haar-, Daubechies-, Sturm-Liouville-Wavelets, u. a.^[49]), welche sich in ihren Eigenschaften und damit in ihren Anwendungsgebieten unterscheiden. Auf eine Beschreibung dieser Typen und der tieferen mathematischen Zusammenhänge wird jedoch an dieser Stelle verzichtet.

Ähnlich der gefensterten Fourier-Transformation wird das Signal innerhalb eines Fensters untersucht. Das Wavelet wird verschoben und durch Stauchung bzw. Streckung des Wavelets wird die Fensterbreite und somit auch die Auflösung variiert.

⁴⁵ Alsberg, B.K.; Woodward, A.M.; Kell, D.B.; An introduction to wavelet transforms for chemometricians. A time-frequency approach; Chem.Lab.Sys (1997) 37; S. 215-239

⁴⁶ Niemöller, A.; Die Wavelet-Analyse als chemometrisches Werkzeug: Analytische Anwendungen in der NIR-Spektrometrie; Dissertation; Universität Duisburg (1999)

⁴⁷ Cohen, L.; Time-Frequency Analysis; Prentice-Hall PTR, Upper Saddle River, New Jersey (1995)

⁴⁸ Walczak, B.; Massart, D.L.; Wavelets - something for analytical chemistry?; Trends anal. chem. (1997) 16, Issue: 8; S. 451-462

Veranschaulichen lässt sich dies bei der schnellen Wavelet-Transformation nach Mallat durch das Schema des Pyramiden-Algorithmus (Abbildung 13)^[50]. Diese Multiskalen-Analyse ist mit einer sukzessiven Hoch- und Tiefpassfilterung vergleichbar^[46]. Ein Funktionsraum V_j wird hierbei in Teilräume zerlegt:

$$V_j = V_{j-1} \oplus W_{j-1} \quad \text{Gleichung 15}$$

Alle Elemente eines Teilraumes sind auch im übergeordnetem Raum enthalten und es bleiben alle Elemente des ursprünglichen Raumes erhalten. Das ursprüngliche Signal wird also in mehrere Teilräume zerlegt, wobei in V_0 , dem untersten Band, die niedrigsten Frequenzen enthalten sind. Bedingt durch die geringe Auflösung enthalten die untersten Bänder weniger Wavelet-Koeffizienten als die höheren, was in Abbildung 13 durch die verschiedenen breiten Boxen angedeutet wird. Abbildung 14 zeigt diese Zerlegung an einem Beispielspektrum.

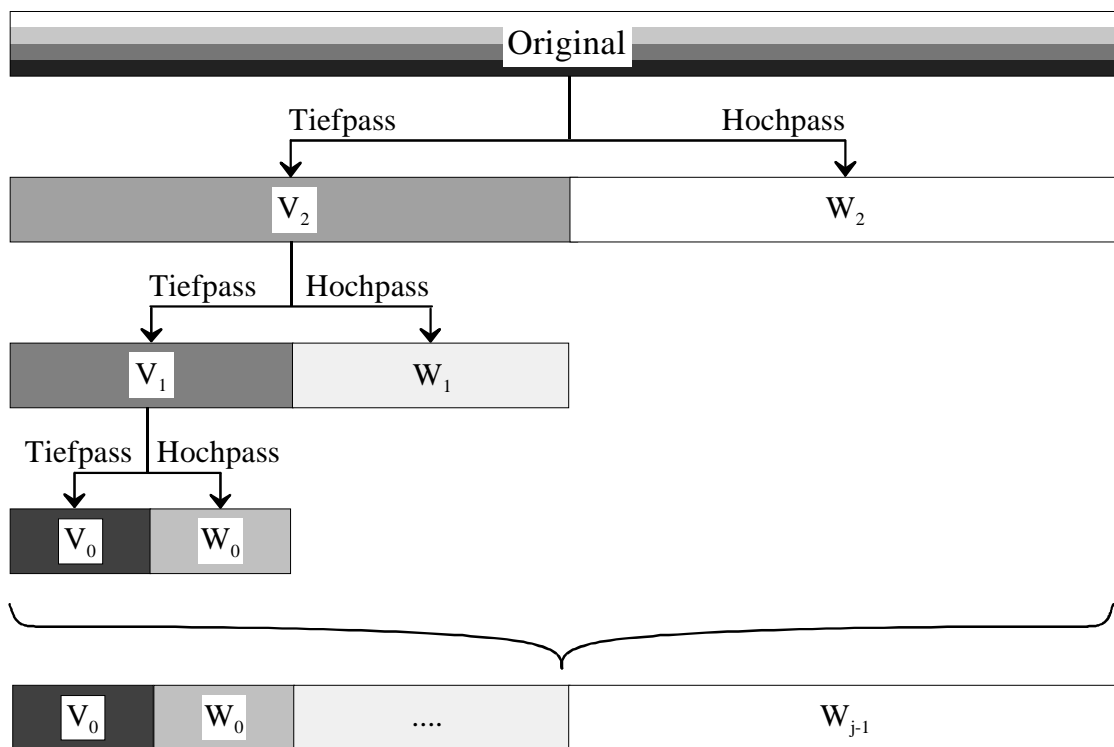


Abb. 13: Wavelet-Zerlegung eines Signals nach S. Mallat mit Hoch- und Tiefpassfilter.

⁴⁹ Daubechies, I.; Ten Lectures on Wavelets; SIAM Publ.; Philadelphia (1992)

⁵⁰ Depczynski, U.; Jetter, K.; Molt, K.; Niemoller, A.; The fast wavelet transform on compact intervals as a tool in chemometrics. I. Mathematical background; Chem. Lab. Sys. (1997) 39; S. 19-27

Jedes Band repräsentiert den gesamten Spektralbereich des Spektrums. In dem untersten Band sind in der Regel Basislinienanteile enthalten. Da z. B. bei Raman-Spektren die Basislinie relativ flach verläuft, sind die Beträge der Wavelet-Koeffizienten in den untersten beiden Bändern relativ klein. Basislinieneffekte von NIR-Spektren können durch Beschneiden des untersten Bandes vermindert werden. Bei den mittleren Bändern sind Übereinstimmungen mit den Bandenlagen des Ursprungssignals zu erkennen. Das Vorzeichen der Koeffizienten kann bei diesen Betrachtungen vernachlässigt werden. Rauschen kann durch Entfernen des obersten Bandes vermindert werden. Dieses Beschneiden der Bänder wird als „Wavelet-smoothing“ bezeichnet. Eliminieren weiterer hochfrequenter Bänder kann insbesondere bei Raman-Spektren schnell zu Verschlechterungen führen, da Raman-Banden sehr schmal sind und deren hochfrequenten Anteile daher deutlich ausgeprägt sind.

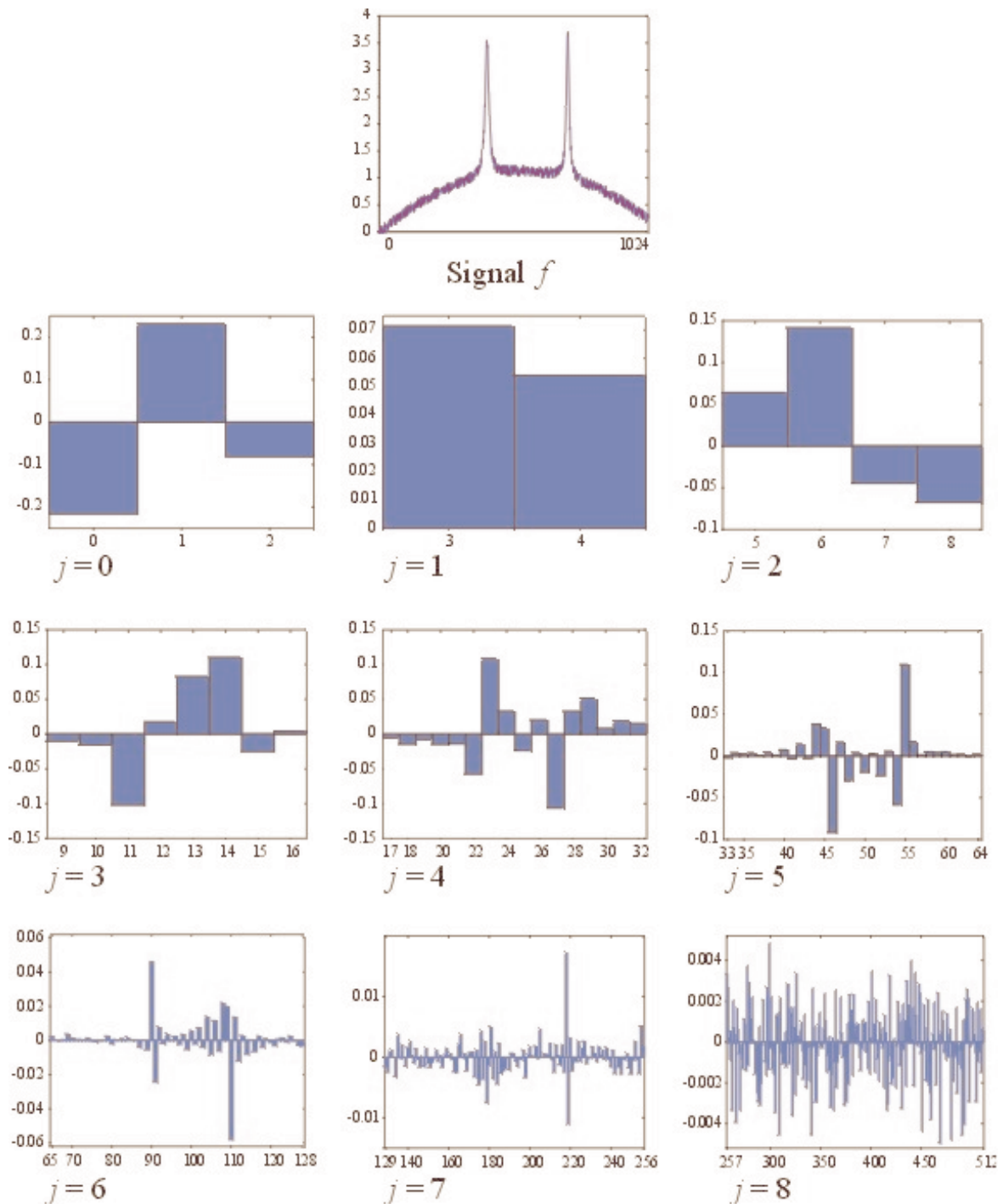


Abb. 14: Wavelet-Zerlegung eines Beispiel-Spektrums (Bänder von links oben nach rechts unten).

Generell können mit der Wavelet-Zerlegung verschiedene Ziele angestrebt werden. Ein Nebeneffekt des oben beschriebenen Einsatzes von Wavelets im Sinne einer Datenvorbehandlung zur Rauschunterdrückung und Glättung ist eine Datenreduktion. Diese Komprimierung von Daten ist in vielen Fällen (z. B. Bild- oder Tonverarbeitung) ein zentraler Anwendungsaspekt von Wavelets.

Während bei der oben beschriebenen Methode hierzu ganze Bänder weggelassen werden, besteht auch die Möglichkeit die einzelnen Wavelet-Koeffizienten zu betrachten und zu überprüfen, ob ihr Wert einen bestimmten Mindestbetrag überschreitet (Thresholding).

Hierbei wird die Tatsache genutzt, daß Koeffizienten mit sehr kleinem Betrag nur wenig Information enthalten. Bei dieser Art der Datenkomprimierung mit „Wavelet-denoising“ oder „thresholding“ wird daher ein Grenzwert berechnet und alle Koeffizienten gleich Null gesetzt, die kleiner als dieser Grenzwert sind. Nach Donoho kann mit einem „universal threshold“ weißes Rauschen entfernt werden. Zur Berechnung des entsprechenden Grenzwertes wird aus dem obersten Band die Standardabweichung σ des Rauschens bestimmt und in Gleichung 16 eingesetzt^[45]:

$$\text{Threshold} = \sqrt{2 \log N} * \sigma \qquad \text{Gleichung 16}$$

wobei: N= Anzahl der Datenpunkte

In *sämtlichen* Bändern werden schließlich die Koeffizienten, die kleiner als dieser Threshold sind, gleich Null gesetzt. Es kann somit Rauschen unabhängig von der Frequenz entfernt werden.

Auf diese Art und Weise läßt sich z. B. bei der Raman-Testbibliothek (Mittelwertspektren) eine Datenreduktion auf 20 % erreichen. Ferner wurde in dieser Arbeit überprüft, inwieweit die Wavelet-Transformation genutzt werden kann, um die Spektren in einer NIR- bzw. Raman-Bibliothek anstelle der herkömmlichen Datenvorbehandlung zu diskriminieren.

Es können sowohl die Wavelet-Koeffizienten, als auch die rekonstruierten Spektren für die Analyse genutzt werden, wobei die Information in den Wavelet-Koeffizienten anders verteilt ist, was unter Umständen zu einer Verbesserung des Ergebnisses führen kann. Bei den rekonstruierten Spektren dient die Wavelet-Transformation lediglich als Datenvorbehandlung, eine Komprimierung findet hier nicht statt.

2.5 Distanz und Ähnlichkeit

Eine Methode der Mustererkennung ist der Punkt zu Punkt Vergleich der Spektren. Bei dem Vergleich der Vollspektren wird zwischen zwei verschiedenen Maßen unterschieden, dem bereits diskutierten Distanz- oder Abstandsmaß und dem Ähnlichkeitsmaß.

Die jeweilige Maßzahl wird für das gesamte Spektrum berechnet. Liegen zwei identische Spektren vor, so haben diese die Distanz 0. Die Ähnlichkeit ist maximal und wird in der Regel auf einen Wert von 1 oder 2 normiert^[51].

Häufig angewandte Abstandsmaße sind die Summe der absoluten Differenzen, die Euklidische- und die Mahalanobis-Distanz^[52]. Betrachtet man die Spektren als Vektoren in einem n -dimensionalen Raum, wobei n die Anzahl der Datenpunkte ist, so ist die Euklidische Distanz (Gleichung 17) die Länge des resultierenden Differenz-Vektors^[53].

$$D_{i,j} = \sqrt{\sum_{i=1}^n (x_i - x_j)^2} \quad \text{Gleichung 17}$$

Die Länge $D_{i,j}$ hängt hierbei von den Intensitäten der beteiligten Spektren ab. Für eine bessere Vergleichbarkeit der berechneten Distanzwerte ist daher eine vorangestellte Normierung der Spektren sinnvoll.

Bei der Mahalanobis-Distanz $D_{i,j}$ (Gleichung 18) wird durch die Kovarianz-Matrix C (Gleichung 19) eine Skalierung der Daten erzielt. In diese fließen die Varianzen (Gleichung 20) und die Kovarianzen (Gleichung 21) der Meßvariablen n ein^[2].

$$D_{i,j} = \sqrt{(x_i - x_j)^T C^{-1} (x_i - x_j)} \quad \text{Gleichung 18}$$

$$C = \begin{pmatrix} s_{1,1}^2 & \text{COV}_{1,2} & \dots & \text{COV}_{1,n} \\ \text{COV}_{2,1} & s_{2,2}^2 & \dots & \text{COV}_{2,n} \\ \dots & \dots & \dots & \dots \\ \text{COV}_{n,1} & \text{COV}_{n,2} & \dots & s_{n,n}^2 \end{pmatrix} \quad \text{Gleichung 19}$$

wobei: $\text{cov}_{j,k}$ = Kovarianz
 s_{jj}^2 = Varianz
 n = Anzahl der Variablen

⁵¹ OPUS-Ident; Handbuch zur Identity Test Software; Spektroskopiesoftware OPUS (Vers. 2.2); Bruker Analytische Meßtechnik GmbH (1995)

⁵² Luinge, H.J.; Leussink, E.D.; Visser, T.; Trace-level identity confirmation from infrared spectra by library searching and artificial neural networks; Anal. Chim. Acta (1997) 0; S. 173-184

⁵³ Graham, R.C.; Data Analysis for the Chemical Sciences – A Guide to Statistical Techniques; VCH Weinheim (1993)

$$\text{cov}_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad \text{für } j, k \in \{1, 2, \dots, n\}, j \neq k$$

Gleichung 20

$$s_{jj}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \quad \text{für } j \in \{1, 2, \dots, n\}$$

Gleichung 21

In dieser Arbeit wird hauptsächlich die Euklidische Distanz verwendet. Weitere Distanzmaße sind der „Conformity Index“ (CI) und der „Spectral Match Value“ (SMV)^[26]. Bei dem Conformity Index wird das Differenzspektrum berechnet und durch die Standardabweichung an jedem einzelnen Punkt geteilt. Das Maximum der resultierenden Werte ist der Conformity Index. Bei vollkommener Übereinstimmung ergibt sich ein CI von 0. Der Spectral Match Value ist der Cosinuswert zwischen den beiden Spektren, die als Vektoren im n-dimensionalen Raum betrachtet werden. Bei perfekter Übereinstimmung ist der SMV gleich 1. In der Literatur werden weitere Algorithmen zum Vergleich von Spektren beschrieben, welche im wesentlichen auf Distanz- oder Ähnlichkeitsmaße zurückzuführen sind^{[54][55]}.

⁵⁴ Kwiatkowski, J.; Riepe, W.; A Quality Index for Spectra Library Search Methods; Fresenius (1980) 302; S. 300-303

⁵⁵ Stein, S.E.; Scott, D.R.; Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification; J. Am. Soc. Mass Spectr. (1994) 5; S. 859-866