3 Datenanalyse und Regression

In den folgenden Abschnitten werden die Grundlagen der *Spektrochemometrie* [15–17] und der *Principal Component Regression* [18–20] erläutert. Hierbei handelt es sich stets um sogenannte multivariate Verfahren.

Grundsätzlich wird zwischen *univariaten* und *multivariaten* Analysenmethoden unterschieden. Während durch ein univariates Verfahren ein Zusammenhang zwischen einer Eigenschaft und nur einer Meßgröße hergestellt wird, werden bei multivariaten Verfahren mehrere Meßgrößen zur Beschreibung der Eigenschaft herangezogen. Multivariate Kalibrationsverfahren finden immer dann Anwendung, wenn das Optimum der mit einer Probeneigenschaft in bezug stehenden Information aus sehr großen Datenmengen herausgefiltert werden soll [1]. Diese Problemstellung tritt bei der Auswertung von Schwingungsspektren auf, wenn z.B. die Konzentration einer Probe in einer chemisch komplexen Matrix bestimmt werden soll [21, 22].

3.1 Lambert–Beersches Gesetz

Univariate und auch multivariate spektroskopische Auswerteverfahren bauen auf dem Gesetz von *Lambert–Beer* [23, 24] auf, das einen linearen Zusammenhang zwischen unabhängigen (spektralen) Meßgrößen und abhängigen Eigenschaften (z.B. der Konzentration) einer Substanz in einer Probe formuliert. Allgemein kann das Gesetz für die Intensitätsänderung $\mathrm{d}i$ einer Strahlung an der Wellenlänge λ in Abhängigkeit von einer absorbierenden Spezies im Schichtdickensegment $\mathrm{d}x$ als

$$\frac{\mathrm{d}i}{i} = -\alpha_\lambda p \, \mathrm{d}x \tag{3.1}$$

formuliert werden. Der Wert von p entspricht dabei z.B. der Konzentration der betrachteten Spezies[1]. Der Proportionalitätsfaktor α_λ ist eine stoffspezifische Größe und von der Wellenlänge λ abhängig. Durch Integration über die Schichtdicke d

[1] Normalerweise wird die Konzentration der Standards mit c bezeichnet. In der Spektrochemometrie können jedoch auch andere Eigenschaftswerte durch das Kalibrationsmodell erfaßt werden, die keine direkte konzentrationsabhängige chemische Eigenschaft darstellen (wie z.B. die Temperatur, Viskosität, etc.). Daher ist die im Umfeld der Chemometrie übliche Bezeichnung p für *property* allgemeiner und zutreffender.

ergibt sich für die Intensitätsänderung von I_0 nach I

$$\int_{I_0}^{I} i^{-1}\,\mathrm{d}i = -\int_0^d \alpha_\lambda p\,\mathrm{d}x \quad . \tag{3.2}$$

In der Gleichung steht I_0 für die in die Probe einfallende und I für die aus der Probe austretende Intensität der Strahlung der Wellenlänge λ. Ist die absorbierende Spezies homogen in der Probe verteilt, so ist p unabhängig von x und durch Lösen des Integrales in Gleichung (3.2) ergibt sich das Lambert–Beersche Gesetz:

$$I = I_0 \cdot 10^{-\epsilon_\lambda p d} \quad \text{mit } \epsilon_\lambda = \frac{\alpha_\lambda}{\ln 10} \quad . \tag{3.3}$$

Dabei ist ϵ_λ der molare Absorptionskoeffizient (auch Extinktionskoeffizient genannt) der absorbierenden Spezies an der Wellenlänge λ. Durch Umformung der Gleichung (3.3) erhält man die dimensionslose Größe A

$$A = \lg \frac{I_0}{I} = \lg \frac{1}{T} = -\lg T = \epsilon_\lambda p d \quad , \tag{3.4}$$

die als Absorption (bzw. optische Dichte oder Extinktion) bezeichnet wird. Die Größe $T = \frac{I}{I_0}$ wird als Durchlässigkeit oder Transmission bezeichnet.

Betrachtet man die Konzentration p als abhängige und die gemessene Absorption als unabhängige Variable, so ergibt sich aus Gleichung (3.4) das inverse Lambert–Beersche Gesetz

$$p = b_\lambda \cdot A \quad \text{mit} \quad b_\lambda = (\epsilon_\lambda \cdot d)^{-1} \quad . \tag{3.5}$$

Der wellenlängenabhängige, stoffspezifische Koeffizient b_λ ist für eine feste Schichtdicke d konstant. Durch Erweiterung von Gleichung (3.5) lassen sich systematische Abweichungen (b_0) und zufällig verteilte Fehler (e) der Messung berücksichtigen. Für eine bestimmte Messung j gilt dann

$$p_j = b_0 + b_\lambda \cdot A_j + e_j \quad . \tag{3.6}$$

p_j ist hierbei der mit dem Regressionsmodell auf Grund der Messung j vorausgesagte Eigenschaftswert. In dieser univariaten Schreibweise hat das inverse Lambert–Beersche Gesetz die Form einer einfachen Geradengleichung.

Die Kalibrierempfindlichkeit ergibt sich dann gemäß IUPAC[2] als Inverse von b_λ [23, 24].

[2]Sie ist ein Maß für die Fähigkeit einer Methode zwischen kleinen Differenzen der untersuchten Eigenschaft zu unterscheiden. Von der IUPAC wird die Kalibrierempfindlichkeit (das einfachste quantitative Maß der Empfindlichkeit) als die 1. Ableitung der Kalibrationsfunktion im Meßbereich definiert. Für lineare Modelle entspricht dies der Steigung der Geradengleichung.

Der *systematische* Fehler (b_0) einer Kalibration ergibt sich als konstante, additive Abweichung zwischen der tatsächlichen Eigenschaft einer Probe und ihrem Erwartungswert. Er ist ein Maß für die gerichtete Abweichung einer bestimmten Analysenmethode und kann durch entsprechende Korrekturen erfaßt oder eliminiert werden. Der durch Rauschen bedingte *zufällig* verteilte Fehler (e_j) läßt sich bei Vorliegen einer ausreichenden Zahl von Standards statistisch bestimmen. Er wirkt sich auf die Reproduzierbarkeit der Methode aus und sollte so klein wie möglich sein.

3.2 Multivariate Analysenverfahren

Während mit univariaten Analysenverfahren nur ein Zusammenhang zwischen einer einzelnen Variablen und einer bestimmten Eigenschaft herstellbar ist, können mit Hilfe *multivariater* Analysenverfahren mehrere Variablen gleichzeitig zur Beschreibung einer Eigenschaftsänderung herangezogen werden. Multivariate Analysenverfahren kommen immer dann zum Einsatz, wenn aus sehr großen Datenmengen Information herausgefiltert werden soll, die mit einer bestimmten Eigenschaft korreliert. Diese Problemstellung tritt z.B. bei der Auswertung von Schwingungsspektren auf, wenn die Konzentration einer Komponente in einer komplexen Matrix bestimmt werden soll [21]. Durch den Einsatz multivariater spektrochemischer Analysenverfahren kann man heute in vielen Fällen den Einsatz teurer naßchemischer oder aufwendiger Verfahren wie z.B. der HPLC deutlich einschränken. Der Grund für diese hohe Leistungsfähigkeit ist, daß die ursprünglichen Meßdaten selbst nicht mehr direkt erkennbar selektiv sein müssen — die Kalibration selbst extrahiert die mit der Probeneigenschaft korrelierenden spektralen Features [20].

Zur Lösung analytischer Fragestellungen werden heute verschiedene multivariate Kalibrationsverfahren eingesetzt. Der naheliegendste Ansatz ist die Erweiterung univariater Verfahren in Form der Messung der Spektrenintensität an mehreren Wellenlängen. Interferierende Effekte werden hierbei auf Basis einer Least–Squares–Rechnung minimiert. In der Spektrochemometrie wird ein solches Least–Squares–Problem als *Multiple–Linear–Regression* (MLR) bezeichnet. Die Auswahl geeigneter Wellenlängen ist jedoch angesichts von Matrixeffekten und Interferenzen nicht trivial, sondern stellt ein erhebliches kombinatorisches Optimierungsproblem dar. Für dessen Lösung gewinnen Genetische Algorithmen (GAs) in zunehmendem Maße an Bedeutung [25–28] (*s.* 4.3).

Eine Alternative zur Wellenlängenselektion bieten faktoranalytische Verfahren wie die *Principal Component Regression* (PCR) und *Partial Least Squares* (PLS), die sich einer faktoriellen Zerlegung der spektralen Datenmatrix bedienen und wesentliche Vorzüge anderer multivariater Auswerteverfahren wie der *Classical Least Squares* (K-Matrix Methode) und der *Inverse Least Squares* (P-Matrix Methode) in sich vereinen [2, 29]. Dazu gehört die Invarianz des Modells in bezug auf Art und Anzahl der Begleitkomponenten. Das heißt, daß eine quantitative Analyse auch dann durch-

geführt werden kann, wenn nur die Konzentration der zu bestimmenden Komponente (jedoch nicht die von sonstigen Begleitkomponenten) in den Proben bekannt ist. Zum anderen handelt es sich bei PCR und PLS um Vollspektrenmethoden, d.h. eine Wellenlängenselektion entfällt [21, 30]. An deren Stelle tritt eine Faktor–Selektion.

PCR- und PLS-Verfahren liefern annähernd gleiche Ergebnisse, wie in einer Reihe von Anwendungen gezeigt werden konnte [2, 31, 32]. Diese beiden prinzipiell sehr ähnlichen Verfahren unterscheiden sich darin, daß im Falle der PLS neben den Spektren der Standards auch die Probeneigenschaft im Rahmen eines iterativen Verfahrens in die Varianzanalyse eingeht [19, 20, 29].

Einen anderen Ansatz beinhalten Verfahren auf Basis von *Fourier–* oder *Wavelet–Transformationen*. Sie bieten die Möglichkeit zur Auswertung einzelner Frequenzanteile (im Sinne einer Signalanalyse) der spektralen Daten. Eine Kombination dieser Transformationen mit verschiedenen Regressionsverfahren ermöglicht eine deutlich vereinfachte und selektive Analyse der Meßdaten. Einführende Literatur und Anwendungsbeispiele zu diesem Themenkreis finden sich in den Veröffentlichungen von *X. Dai et al.* [33], *D. Massart* und *B. Walczak* [34], *A. Graps* [35], *B.K. Alsberg et al.* [36] und *U. Depczynski et al.* [37].

3.2.1 Principal Component Regression

Die sich im Zusammenhang mit der NIR–Spektrochemometrie ergebenden Problemstellungen der Principal Component Regression (PCR) bilden den thematischen Schwerpunkt der vorliegenden Arbeit. Die PCR stellt ein leistungsfähiges Verfahren zur quantitativen Analyse spektraler Daten dar und bietet gleichzeitig die Möglichkeit der Datenreduktion.

Im Rahmen einer PCR erfolgt eine Faktorisierung der Datenmatrix $\mathbf{X}$ und anschließend eine Regression der Faktoren auf die zu kalibrierende Probeneigenschaft mit Hilfe einer *Least–Squares* Rechnung.

Die Zerlegung in Faktoren

Für den Zerlegungsschritt existieren zwei mögliche Ansätze: Die Eigenwert- und die Singulärwertzerlegung. In den folgenden Abschnitten werden die Verfahren erläutert und ihre Unterschiede herausgearbeitet.

Singulärwertzerlegung

Bei der Singulärwertzerlegung (SVD) erfolgt eine direkte Zerlegung einer rechteckigen Spektrenmatrix $\mathbf{X}$ in ihre Singulärwerte σ_i und zwei zugehörige Singulärvektormatrizen $\mathbf{U}$ und $\mathbf{V}$. Für einen reellen Datensatz $\mathbf{X} \in \mathbb{R}^{l \times k}$ mit k Spaltenvektoren als Kalibrationsspektren und l diskreten Wellenlängen in jedem Spektrum lassen sich

nach Definition *orthogonale* Singulärvektormatrizen bestimmen, so daß gilt:

$$\mathbf{X} = \mathbf{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\mathbf{V}^T \quad \text{mit } n = \min\{l, k\} \quad \text{und } \sigma_1 \geq \sigma_{i+1} \geq 0 \tag{3.7}$$

Die Gleichung läßt sich mit $\mathbf{W} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ als

$$\mathbf{X} = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{3.8}$$

schreiben. Die Matrizen $\mathbf{U} \in \mathbb{R}^{l\times l}$ und $\mathbf{V} \in \mathbb{R}^{k\times k}$ werden als linke und rechte Singulärvektormatrizen von $\mathbf{X}$ bezeichnet und enthalten die Eigenvektoren der symmetrischen Matrizen $\mathbf{X}\mathbf{X}^T$ und $\mathbf{X}^T\mathbf{X}$. Die Vektoren der Matrizen $\mathbf{U}$ und $\mathbf{V}$ sind orthonormal[3]. $\mathbf{W}$ enthält die Singulärwerte, die ein Maß für die durch die zugehörigen Singulärvektoren repräsentierten spektralen Varianzen darstellen.

Beispiel 3.1
Am folgenden Beispiel wird der Zusammenhang deutlich: Der Datensatz in Ab-

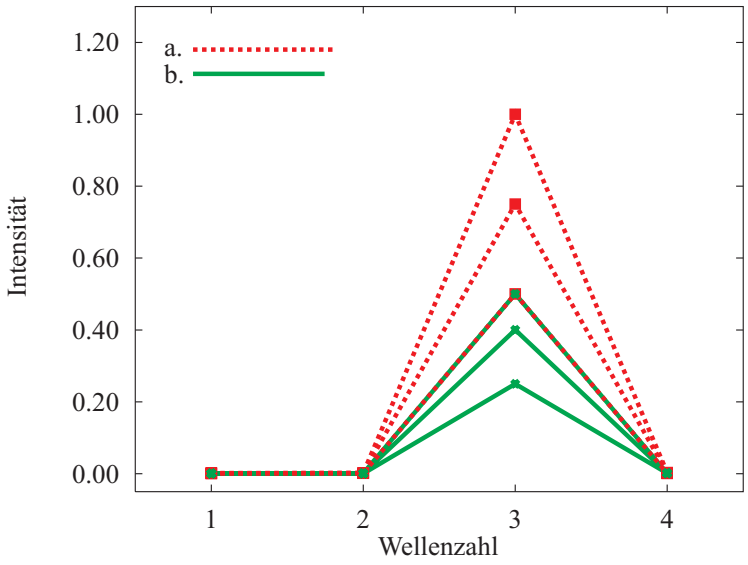


Abbildung 3.1: Original–Datensatz. Die mit *a.* bezeichneten Spektren bilden den Kalibrations-, die mit *b.* bezeichneten den Validationsdatensatz.

bildung 3.1, der sechs stark vereinfachte Spektren mit jeweils vier Intensitätswerten

[3] Eine Menge von Vektoren $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_m\} \subset \mathbb{R}^m$ einer Matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$ wird als *orthogonal* bezeichnet wenn gilt $\mathbf{m}_i^T\mathbf{m}_j = 0$ für alle $i \neq j$. Und sie ist *orthonormal* wenn $\mathbf{m}_i^T\mathbf{m}_j = \delta_{ij}$ bzw. wenn die Kovarianzmatrix von $\mathbf{M}$ gleich der Identitätsmatrix $\mathbf{I}$ ist: $\mathbf{M}^T\mathbf{M} = \mathbf{I}$. In diesem Fall wird von $\mathbf{M}$ als einer orthogonalen Matrix gesprochen. Für das Kronecker–Delta gilt $\delta_{ij} = \begin{cases} 1 : i = j \\ 0 : i \neq j \end{cases}$.

beinhaltet, soll analysiert werden. Drei Spektren hiervon dienen als Kalibrationsdaten.

$$\mathbf{A} = \begin{pmatrix} 0.0010 & 0.0020 & 0.0010 \\ 0.0025 & 0.0018 & 0.0013 \\ 1.0001 & 0.7505 & 0.5001 \\ 0.0030 & 0.0025 & 0.0009 \end{pmatrix} \tag{3.9}$$

Die übrigen Spektren werden zur Validation verwendet. Aus numerischen Gründen wird der Datensatz zentriert. Dazu wird das aus den drei Kalibrationsspektren

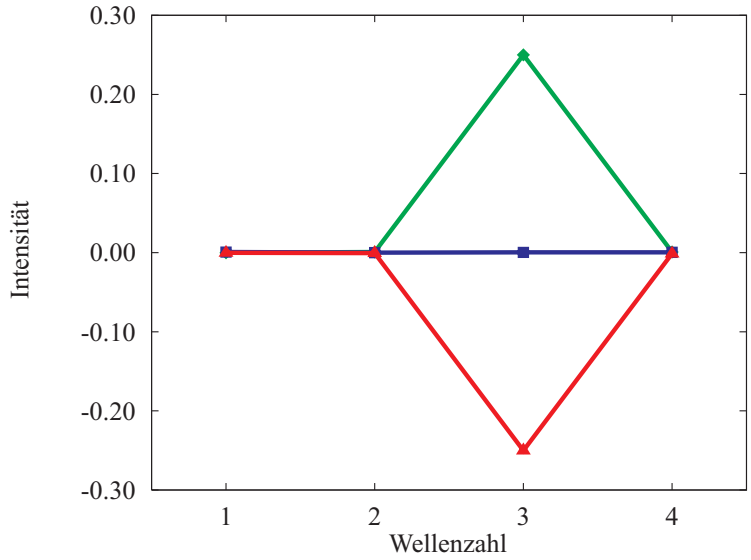


Abbildung 3.2: Zentrierter Kalibrationsdatensatz.

berechnete Mittelwertspektrum von allen Spektren subtrahiert. Durch Zentrierung von $\mathbf{A}$ *ergibt sich die Matrix* $\mathbf{X}$*, deren Zeilen den Mittelwert Null besitzen.*

$$\mathbf{X} = \begin{pmatrix} -0.0003 & 0.0007 & -0.0003 \\ 0.0006 & -0.0001 & -0.0006 \\ 0.2499 & 0.0003 & -0.2501 \\ 0.0009 & 0.0004 & -0.0012 \end{pmatrix} \tag{3.10}$$

Aus der Singulärwertzerlegung von $\mathbf{X}$ *gemäß Gleichung (3.8) ergeben sich die Sin-*

gulärvektormatrizen $\mathbf{U}$ *und* $\mathbf{V}$ *sowie die Singulärwerte* $\mathbf{W}$*:*

$$\mathbf{U} = \begin{pmatrix} 0.000002 & 0.873420 & -0.482890 & 0.062862 \\ 0.002400 & -0.088185 & -0.029892 & 0.995650 \\ 0.999990 & -0.001802 & -0.003604 & -0.002678 \\ 0.004201 & 0.478910 & 0.875160 & 0.068681 \end{pmatrix} \tag{3.11}$$

$$\mathbf{V} = \begin{pmatrix} 0.706730 & -0.408900 & 0.577350 \\ 0.000758 & 0.816500 & 0.577350 \\ -0.707490 & -0.407590 & 0.577350 \end{pmatrix} \tag{3.12}$$

$$\mathbf{W} = \begin{pmatrix} 0.353560 & 0 & 0 \\ 0 & 0.000935 & 0 \\ 0 & 0 & 0.000000 \\ 0 & 0 & 0 \end{pmatrix} \tag{3.13}$$

Die Singulärwertmatrix $\mathbf{W}$ *besitzt nur zwei von Null verschiedene Einträge. Das heißt, daß durch zwei Singulärvektoren die Varianz des aus drei Spektren bestehenden Kalibrationsdatensatzes beschrieben wird. Dieser Effekt ist auf die Zentrierung des Datensatzes zurückzuführen, die den Verlust eines Freiheitsgrades nach sich zieht.*

Die Singulärwerte $\sigma_1, \ldots, \sigma_n$ von $\mathbf{X}$ bilden die Diagonalelemente der Singulärwertmatrix $\mathbf{W} \in \mathbb{R}^{l \times k}$. Ist $\mathbf{X} \in \mathbb{R}^{l \times k}$ keine quadratische Matrix ($l > k$), dann wird die $l \times k$ Singulärwertmatrix $\mathbf{W}$ in den Zeilen $k+1$ bis l mit Nullen aufgefüllt (*s.* Gleichung (3.13)). In vielen Softwareprogrammen für die Spektrochemometrie bleiben Singulärwerte mit dem Wert Null bei der Angabe von $\mathbf{W}$ unberücksichtigt. Dies führt zu dem falschen Eindruck, es handele sich bei der Singulärwertmatrix um eine $k \times k$ Matrix [19, 20, 38].

Aus Gleichung (3.8) und dem Beispiel 3.1 ist offensichtlich, daß die Spektrenmatrix $\mathbf{X}$ mit Hilfe der Singulärvektormatritzen $\mathbf{U}$ und $\mathbf{V}$ sowie der Singulärwertmatrix $\mathbf{W}$ vollständig beschrieben werden kann. Die Singulärwerte repräsentieren den Anteil, der durch die zugehörigen Eigenvektoren erfaßten spektralen Information (spektrale Varianz). Das heißt, die gesamte in einem Datensatz enthaltene spektrale Varianz (Gesamtvarianz) spiegelt sich in der Summe über alle Singulärwerte wieder [39]. Dies läßt sich wie folgt für die Elemente $x_{a,b}$ der Spektrenmatrix $\mathbf{X}$ verdeutlichen, es gilt

$$\sum_{a=1}^{k} \sum_{b=1}^{l} x_{a,b}^2 = \sum_{i=1}^{s} \sigma_i^2 \quad \text{mit } s = min(k,l) \quad , \tag{3.14}$$

oder

$$\sum_{a=1}^{s} \text{diag}\left(X^T X\right)_a = \sum_{i=1}^{s} \sigma_i^2 \quad . \tag{3.15}$$

Bei der Zerlegung fallen die Singulärwerte geordnet nach abnehmenden Beiträgen zur Gesamtvarianz des ursprünglichen Datensatzes $\mathbf{X}$ an. Es ist anschaulich klar, daß je kleiner ein Singulärwert σ_i ist, desto geringer ist der durch den zugehörigen Singulärvektor ($\mathbf{u}_i$ bzw. $\mathbf{v}_i$) wiedergegebene Anteil an der Gesamtvarianz des zerlegten Datensatzes $\mathbf{X}$. Ein solcher Singulärvektor repräsentiert also überwiegend zufälliges Rauschen bzw. Minoritätskomponenten. Dieser Zusammenhang wird im Rahmen von multivariaten Kalibrationen ausgenutzt, um chemisch relevante Information von meßtechnischen Störungen und Rauschen zu separieren (*s.* 3.3 und 4).

Eigenwertzerlegung

Im Gegensatz zu einer Singulärwertzerlegung kann eine Eigenwertzerlegung nur mit quadratischen Matrizen durchgeführt werden. In aller Regel wird daher nicht die im allgemeinen rechteckige Datenmatrix $\mathbf{X}$, sondern deren Kovarianzmatrix zerlegt. Für die reelle Datenmatrix $\mathbf{X} \in \mathbb{R}^{l\times k}$ sei k wiederum die Zahl der Spektren im Kalibrationsdatensatz und l die Zahl der diskreten Wellenlängen in den Spektren. Die Kovarianzmatrix $\mathbf{Z} = \mathbf{X}^T\mathbf{X},\ \mathbf{Z} \in \mathbb{R}^{k\times k}$ ist dann reell und symmetrisch. Das Ziel der Eigenwertzerlegung ist die Faktorisierung der Kovarianzmatrix in Eigenwerte und eine orthonormale Basis von Eigenvektoren. Durch die Symmetrie von $\mathbf{Z}$ wird die Existenz einer reellen orthogonalen Eigenvektormatrix $\mathbf{C} \in \mathbb{R}^{k\times k}$, welche die Eigenvektoren zeilenweise enthält, und reeller Eigenwerte, die größer oder gleich Null sind, impliziert.

Durch die Lösung des Eigenwertproblems können die Matrizen $\mathbf{C}$ und $[\boldsymbol{\lambda}]$ so ermittelt werden, daß gilt

$$\mathbf{C}\mathbf{Z}\mathbf{C}^T = \mathrm{diag}(\lambda_1, \ldots, \lambda_k) \tag{3.16}$$

$$= [\boldsymbol{\lambda}] \quad . \tag{3.17}$$

Die Eigenwerte $\lambda_1, \ldots, \lambda_k$ bilden die Diagonalelemente der quadratischen Eigenwertmatrix $[\boldsymbol{\lambda}]$. Die Matrixelemente jenseits der Diagonalen besitzen den Wert 0. Im Vergleich zur Singulärwertzerlegung fallen die Eigenwerte mit ihren zugehörigen Eigenvektoren bei dieser Form der Zerlegung nicht zwangsläufig nach Größe geordnet an. Ob und wie sie geordnet sind, ist abhängig vom verwendeten Zerlegungsalgorithmus.

Beispiel 3.2
Aus der Eigenwertzerlegung des Datensatzes $\mathbf{X}$ *(s. Gleichung (3.10)) lassen sich die*

Matrizen $[\boldsymbol{\lambda}]$ *und* $\mathbf{C}$ *wie folgt berechnen:*

$$[\boldsymbol{\lambda}] = \begin{pmatrix} 0.000001 & 0 & 0 \\ 0 & 0.000000 & 0 \\ 0 & 0 & 0.125000 \end{pmatrix} \tag{3.18}$$

$$\mathbf{C} = \begin{pmatrix} -0.408900 & 0.577350 & -0.706730 \\ 0.816500 & 0.577350 & -0.000758 \\ -0.407590 & 0.577350 & 0.707490 \end{pmatrix} \tag{3.19}$$

Die Eigenwertmatrix $[\boldsymbol{\lambda}]$ *weist in der Diagonalen zwei von Null verschiedene Einträge aus. Der Eigenvektor, der zum zweiten Eigenwert mit* $\lambda = 0$ *gehört, beinhaltet keine spektrale Varianz. Dieser Effekt ist (wie bereits für die Matrix* $\mathbf{W}$ *diskutiert) auf die Zentrierung des Datensatzes zurückzuführen, die den Verlust eines Freiheitsgrades nach sich zieht. Dies ist auch der Grund dafür, daß die Einträge des zweiten Eigenvektors einen konstanten Wert besitzen.*

Sortiert man $[\boldsymbol{\lambda}]$ nach fallenden Eigenwerten λ_i, dann ist

$$\mathrm{diag}(\lambda_1, \ldots, \lambda_k) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2) \quad . \tag{3.20}$$

Der Grund für den quadratischen Zusammenhang ist in der Zerlegung der Kovarianzmatrix $\mathbf{Z} = \mathbf{X}^T\mathbf{X}$ bei der Eigenwertzerlegung im Gegensatz zur Datenmatrix $\mathbf{X}$ im Fall der Singulärwertzerlegung zu sehen.

Für die Spaltenvektoren der quadratischen Matrix $\mathbf{C}^T$, welche die orthonormale Basis von Eigenvektoren der Kovarianzmatrix $\mathbf{Z}$ bilden, wird häufig auch der Begriff *Principal Components* (PCs) verwendet. Insbesondere für große Datensätze gilt, daß die Lösung der Gleichungen (3.7) und (3.16) kein triviales Problem der numerischen Linearen Algebra darstellt [38]. Für die im weiteren beschriebenen Berechnungen wurde das als sehr stabil geltende QR–Verfahren [40] zur Eigenwertzerlegung eingesetzt, das auch in numerisch ungünstigen Fällen die Orthogonalität der Eigenvektoren gewährleistet.

Der Vorteil der Eigenwertberechnung gegenüber der Singulärwertzerlegung besteht in der Zerlegung der im Vergleich zur Datenmatrix $\mathbf{X}$ $(l \times k)$ in der Regel deutlich kleineren und symmetrischen Kovarianzmatrix $\mathbf{Z}$ $(k \times k)$. Dies spart vor allem bei der Berücksichtigung vieler Variablen Rechenspeicher. Der Nachteil besteht in der notwendigen Berechnung der Kovarianzmatrix, was zusätzliche Rechenzeit beansprucht. Eine Singulärwertzerlegung bietet die Möglichkeit, die Datenmatrix $\mathbf{X}$ nach Gleichung (3.7) zu rekonstruieren, während Gleichung (3.16) nur die Berechnung der Kovarianzmatrix $\mathbf{Z}$ zuläßt — eine Rekonstruktion der Datenmatrix ist auf diesem Wege nicht möglich. Die Entscheidung darüber, welches der beiden Zerlegungsverfahren eingesetzt wird, ist daher auf ein Abwägen dieser Argumente zu stützen.

Faktorisierung der Daten

An die Eigenwert- bzw. die Singulärwertzerlegung schließt sich bei der PCA eine Faktorisierung der Datenmatrix $\mathbf{X}$ in eine Spaltenmatrix $\mathbf{C}$ (**C**olumns) und eine Zeilenmatrix $\mathbf{R}$ (**R**ows) an: An die Eigenwert- bzw. die Singulärwertzerlegung der Datenmatrix $\mathbf{X}$ bzw. ihrer Kovarianzmatrix $\mathbf{Z}$ schließt sich die Faktorisierung der Spektrenmatrix gemäß Gleichung (3.21) an:

$$\mathbf{X} = \mathbf{RC} \quad . \tag{3.21}$$

$\mathbf{C}$ entspricht hierbei der oben beschriebenen Eigenvektormatrix. Die Gleichung läßt sich wie folgt so umformen, daß ein Ausdruck für die Berechnung der Zeilenmatrix $\mathbf{R} \in \mathbb{R}^{l \times k}$ resultiert:

$$\mathbf{R} = \mathbf{XC}^T \quad . \tag{3.22}$$

Die Spalten von $\mathbf{R}$ und die Zeilen von $\mathbf{C}$ sind die sogenannten Faktoren. In der Spektrochemometrie beschränkt sich der Begriff *Faktor* jedoch im allgemeinen auf die Spalten von $\mathbf{R}$, die auch Eigenspektren genannt werden.

Die Einträge der Spaltenvektoren von $\mathbf{C}$ können gemäß Gleichung (3.21) als die Wichtungskoeffizienten verstanden werden, mit denen die (abstrakten) Eigenspektren in das jeweilige (gemessene) Spektrum eingehen. Sie werden daher auch als Faktorgewichte bezeichnet. Ganz analog zur Singulärwertzerlegung ergibt sich auch hier, daß die Eigenwerte $[\boldsymbol{\lambda}]$ den jeweiligen Anteil spektraler Varianz beschreiben, der durch einen Faktor repräsentiert wird. Die Summe aller Eigenwerte stellt also die Gesamtvarianz im Datensatz dar.

In der angelsächsischen Literatur werden für die Zeilen von $\mathbf{R}$ und die Spalten von $\mathbf{C}$ häufig die Begriffe *loadings* und *scores* verwendet. Die Definitionen hierzu sind jedoch nicht eindeutig, da sie von der Interpretation der Problemstellung abhängig sind. *E.R. Malinowski* [18] beschrieb diese Situation wie folgt:

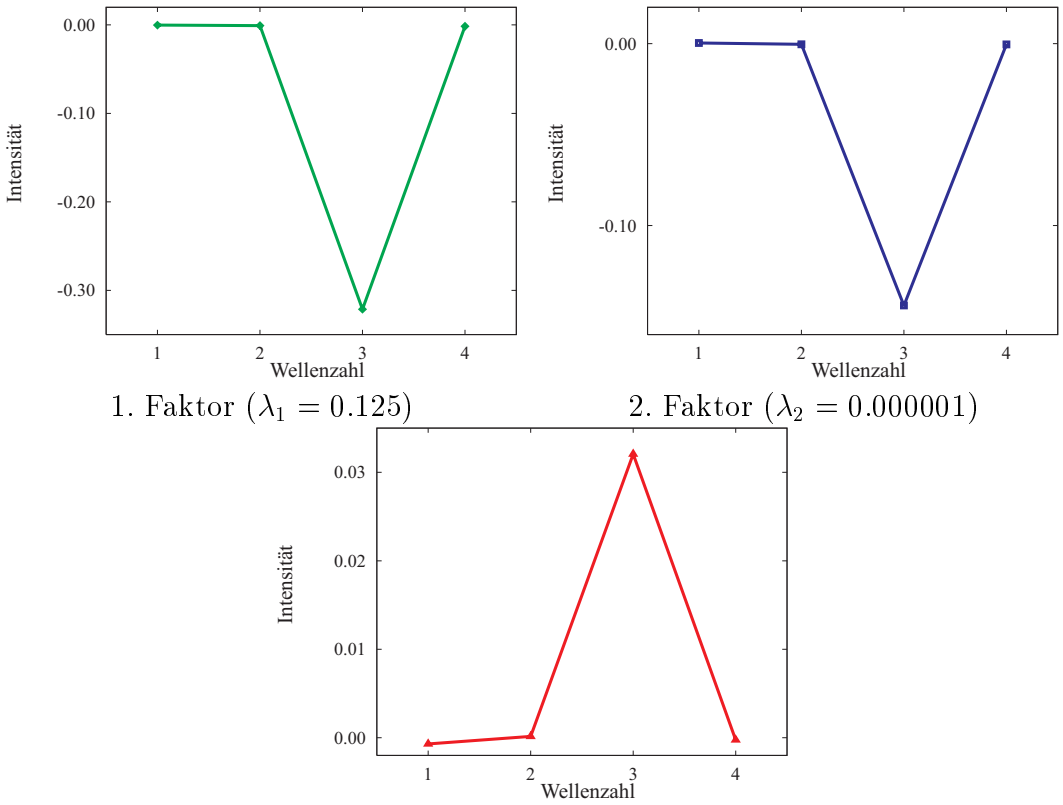
> ≪Often, we focus attention on either the row design or the column design[4]. Whichever we focus attention on we call the *score* and its counterpart we call the loading. ≫

Die Bezeichnung ist also abhängig von dem Stellenwert, welcher der einen oder anderen Matrix zugeordnet wird. In der Spektrochemometrie wird den Faktorgewichten die höhere Bedeutung zugemessen, denn ausschließlich diese werden für die Bestimmung des Regressionsvektors herangezogen. Im folgenden wird daher in Übereinstimmung mit den Arbeiten von *D.M. Haaland* [2] die Matrix $\mathbf{C}$ als Score-Matrix und die Matrix $\mathbf{R}$ als Loading-Matrix bezeichnet.

[4] Im Falle von Gleichung (3.22) $\mathbf{R}$ und $\mathbf{C}$.

Beispiel 3.3
Für den in Beispiel 3.1 beschriebenen Datensatz ergibt sich als Matrix der Faktoren

$$\mathbf{R} = \begin{pmatrix} -0.000228 & 0.000352 & -0.000701 \\ -0.000748 & -0.000381 & 0.000148 \\ -0.321300 & -0.144010 & 0.032064 \\ -0.001475 & -0.000412 & -0.000250 \end{pmatrix} . \tag{3.23}$$



1. Faktor ($\lambda_1 = 0.125$) 2. Faktor ($\lambda_2 = 0.000001$)

3. Faktor ($\lambda_3 = 0$)

Abbildung 3.3: Faktoren des Beispiel–Datensatzes, sortiert nach fallenden Eigenwerten.

Die Faktoren geben die spektralen Eigenschaften des Datensatzes in der Reihenfolge abnehmender Beiträge zur Gesamtvarianz wieder. Dabei weist neben den Eigenwerten $[\boldsymbol{\lambda}]$ *bereits auch die Skalierung der Faktoren in Abb. 3.3 auf die Höhe der von ihnen repräsentierten spektralen Varianz hin.*

Unter Berücksichtigung von Gleichung (3.16) gilt für die Matrix $\mathbf{R}$:

$$\mathbf{R}^T\mathbf{R} = [\boldsymbol{\lambda}] \tag{3.24}$$

Im Rahmen der Faktoranalyse (PCA) werden die Faktoren und ihre zugehörigen Eigenvektoren üblicherweise in der Reihenfolge abnehmender Eigenwerte, d.h. abnehmender Beiträge zur Gesamtvarianz des ursprünglichen Datensatzes geordnet (*s.* Gleichung (3.23)). Das heißt, der erste Faktor repräsentiert den Hauptanteil spektraler Varianz, der zweite Faktor den nächst kleineren Anteil und so weiter.

Regression: Least–Squares–Verfahren

Im weiteren wird darauf eingegangen, wie nach der Faktorisierung aus den Probeneigenschaften und der Matrix der Eigenvektoren / Eigenspektren der Regressionsvektor mit Hilfe einer Least–Squares–Rechnung bestimmt werden kann.

Zur Verdeutlichung wird das in Gleichung (3.6) beschriebene univariate Regressionsmodell aus dem *Lambert–Beerschen* Gesetz herangezogen, in dem eine Variable p (z.B. eine Analytkonzentration) mit einer Variablen A (z.B. der Meßwert eines Instruments) in einen linearen Bezug gesetzt wird:

$$p_j = b_0 + b_\lambda \cdot A_j + e_j \quad \text{mit } j = 1, \ldots, k \quad . \tag{3.25}$$

Die Modellparameter b_0 und b_λ sowie die normalverteilten Residuen e_j sind unbekannt.

Ziel einer Kalibration ist es, für nachfolgende Messungen den Wert von p auf Basis eines geeigneten Regressionsmodells direkt durch die Messung von A zu ermitteln. Dazu müssen die Koeffizienten b_0 und b_λ in der Regression bestimmt werden. Darüber hinaus ist es notwendig, Aussagen über die Qualität des Modells zu treffen.

Ein einfacher und effektiver Weg besteht darin, die Werte der Parameter b_0 und b_λ so zu wählen, daß sich der kleinst mögliche Wert für die Summe der Fehlerquadrate[5] von e ergibt:

$$\sum_{j=1}^{k} (p_j - (b_0 + b_\lambda A_j))^2 \quad . \tag{3.26}$$

Die Lösungen werden mit $\hat{b}_0$ und $\hat{b}_\lambda$ bezeichnet und Schätzer von b_0 und b_λ genannt. Der Ausdruck (3.25) gilt allgemein für univariate Regressionen. Durch Verwendung von Matrizen an Stelle von Skalaren läßt sich die Berechnung auf multivariate Regressionen übertragen.

So läßt sich die Regression in Gleichung (3.6) und Gleichung (3.25) in Matrix–Notation als

$$\mathbf{p} = \mathbf{b}_0 + \mathbf{bX} + \mathbf{e} \tag{3.27}$$

[5] engl.: *residual sum of squares* (RSS) *s.* 3.5.1

schreiben, wobei die Meßgrößen als Datenmatrix $\mathbf{X}$ bezeichnet werden. Für eine zentrierte Matrix vereinfacht sich die Gleichung zu

$$\mathbf{p} = \mathbf{bX} + \mathbf{e} \quad . \tag{3.28}$$

Die gemessenen Spektren bilden dann die Einträge der Matrix $\mathbf{X}$ und der Vektor $\mathbf{p}$ enthält die Werte p_i für die zu kalibrierende Eigenschaft der Proben. Das Least–Squares–Problem kann dann als Minimierung der Länge von $\mathbf{e} = \mathbf{p} - \mathbf{bX}$, d.h. des Skalarprodukts $\mathbf{ee}^T = (\mathbf{p} - \mathbf{bX})(\mathbf{p} - \mathbf{bX})^T$ formuliert werden. Der gesuchte Regressionsvektor $\hat{\mathbf{b}} \in \mathbb{R}^{1\times l}$ kann dann durch

$$\hat{\mathbf{b}} = \mathbf{pX}^+ \tag{3.29}$$

bestimmt werden. Der Ausduck $\mathbf{X}^+$ wird als Pseudo–Inverse von $\mathbf{X}$ bezeichnet und berechnet sich ebenfalls durch die Lösung eines Least–Squares–Problems, für das die Matrix–Norm

$$\|\mathbf{X}^+\mathbf{X} - \mathbf{I}\| \tag{3.30}$$

zu minimieren ist. $\mathbf{I}$ steht dabei für die Einheitsmatrix. Durch Multiplikation des Vektors $\hat{\mathbf{b}}$ mit dem Spaltenvektor eines Spektrums $\mathbf{x}_j \in \mathbb{R}^l$ läßt sich dann umgekehrt der Wert der zum Spektrum gehörenden Probeneigenschaft ermitteln

$$\hat{p}_j = \hat{\mathbf{b}}\mathbf{x}_j \quad . \tag{3.31}$$

Der Regressionsvektor $\hat{\mathbf{b}}$ läßt sich jedoch nicht nur durch Gleichung (3.29), sondern auch unter Berücksichtigung von Gleichung (3.22) auf Basis der Eigenvektoren und Eigenspektren bestimmen. Dies läßt sich einfach durch Umformen von Gleichung (3.28) zeigen. Für die Berechnung des Regressionsvektors gilt

$$\mathbf{p} = \hat{\mathbf{b}}\mathbf{X} + \mathbf{e} \quad , \tag{3.32}$$

wobei $\hat{\mathbf{b}}$ in einem Least–Squares Schritt so zu bestimmen ist, daß die Norm

$$\|\mathbf{p} - \hat{\mathbf{b}}\mathbf{X}\|$$

minimiert wird. Durch Erweitern von Gleichung (3.32) mit $\mathbf{X}^T$ erhält man

$$\mathbf{pX}^T = \hat{\mathbf{b}}\mathbf{XX}^T \tag{3.33}$$

und mit $\mathbf{X} = \mathbf{RC}$ (*s.* Gl. (3.22)) ergibt sich daraus

$$\mathbf{pC}^T\mathbf{R}^T = \hat{\mathbf{b}}\mathbf{R}\underbrace{\mathbf{CC}^T}_{=\mathbf{I}}\mathbf{R}^T \tag{3.34}$$

$$= \hat{\mathbf{b}}\mathbf{RR}^T \quad . \tag{3.35}$$

Wenn die Matrix $\mathbf{R}\mathbf{R}^T$ vollen Rang hat, dann gilt

$$\mathbf{p}\mathbf{C}^T\mathbf{R}^T\left(\mathbf{R}\mathbf{R}^T\right)^{-1} = \hat{\mathbf{b}} \quad . \tag{3.36}$$

Hier kann $\mathbf{R}^T\left(\mathbf{R}\mathbf{R}^T\right)^{-1}$ als Pseudo–Inverse von $\mathbf{R}$ aufgefaßt werden, und es gilt:

$$\mathbf{p}\mathbf{C}^T\left(\mathbf{R}\right)^{+} = \hat{\mathbf{b}} \quad . \tag{3.37}$$

Durch Substitution von $\mathbf{C}\mathbf{p}^T$ mit $\boldsymbol{\beta}$ läßt sich die Gleichung (3.36) auch als

$$\hat{\mathbf{b}} = \boldsymbol{\beta}^T\mathbf{R}^T\left(\mathbf{R}\mathbf{R}^T\right)^{-1} \tag{3.38}$$

$$= \boldsymbol{\beta}^T\left(\mathbf{R}\right)^{+} \tag{3.39}$$

schreiben. Die Einträge des Vektors $\boldsymbol{\beta} \in \mathbb{R}^k$ können als Gewichtungskoeffizienten der Faktoren ($\mathbf{R}$) in bezug auf die zu kalibrierende Probeneigenschaft angesehen werden. Wie in 6.1.1 noch gezeigt wird, lassen sich auf Basis von $\boldsymbol{\beta}$ die Korrelationskoeffizienten zwischen den Scores und der Probeneigenschaft ermitteln.

Die Anwendung auf spektrale Daten

Die Gleichungen lassen sich einfach auf spektrale Daten im Zusammenhang mit der Principal Component Regression (PCR) anwenden. Wird Gleichung (3.38) im Sinne von Gleichung (3.28) erweitert, so läßt sich der Eigenschaftswert $\hat{p}$ eines Spektrums $\mathbf{x}$ direkt berechnen. Dieser Schritt wird auch als *„Vorhersage“* der Probeneigenschaft bezeichnet. Durch Einsetzen erhält man

$$\hat{\mathbf{b}}\mathbf{x} = \boldsymbol{\beta}^T\mathbf{R}^T\left(\mathbf{R}\mathbf{R}^T\right)^{-1}\mathbf{x} \tag{3.40}$$

$$= \hat{p}$$

als Ausdruck für das Regressionsmodell. Der Vektor $\hat{\mathbf{b}}^T$ wird als Property–Weighting–Spektrum bezeichnet und besitzt die gleiche Dimension wie ein Spektrum ($\hat{\mathbf{b}}^T \in \mathbb{R}^l$).

Beispiel 3.4

Für den in den vorangegangenen Abschnitten diskutierten Datensatz ergibt sich aus Gleichung (3.29) mit

$$\mathbf{p}^T = \begin{pmatrix} 0.2500 \\ 0.0000 \\ -0.2500 \end{pmatrix} \textit{ der Regressionsvektor } \hat{\mathbf{b}}^T = \begin{pmatrix} -0.306720 \\ 0.033368 \\ 1.000600 \\ -0.163980 \end{pmatrix} \quad . \tag{3.41}$$

Er zeigt eine herausgehobene Gewichtung der dritten Wellenzahl. Ein Vergleich

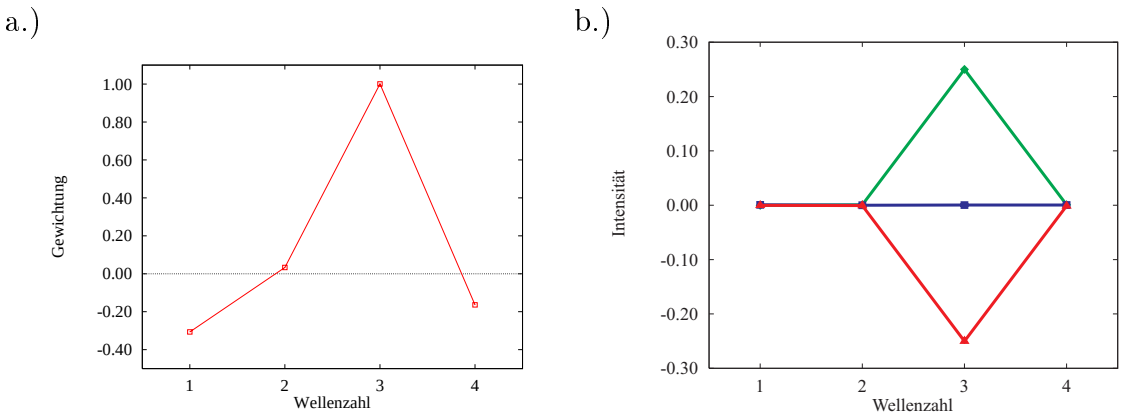


Abbildung 3.4: Regressionsvektor a.) und Kalibrationsdaten b.) der Beispieldaten.

mit den Spektren des Original–Datensatzes macht sofort deutlich, daß an dieser Wellenzahl die Bande und damit die für die Kalibration relevante Information lokalisiert ist (s. Abb. 3.4). Der Regressionsvektor gibt also die Bedeutung einzelner Spektralbereiche in bezug auf die Kalibration wieder und ist damit auch in chemischer Hinsicht interpretierbar. Spektralbereiche mit hohen Regressionskoeffizienten besitzen eine hohe Relevanz für die Bestimmung der Probeneigenschaft. Für die Kalibration unbedeutende Spektralbereiche, die keine chemische Information über diese Probeneigenschaft beinhalten, weisen im Vergleich dazu nur kleine Regressionskoeffizienten auf. Die für die Bestimmung einer Probeneigenschaft wesentlichen Spektralbereiche eines Spektums sind also direkt über das Property–Weighting–Spektrum zugänglich.

Neben dem beschriebenen Verfahren zur Berechnung des Regressionsvektors $\hat{\mathbf{b}}$ existieren noch einige andere Least–Squares–Verfahren zu seiner Bestimmung. Insbesondere dann, wenn unterschiedliche Spektralbereiche unterschiedlich gewichtet werden sollen, sollten Verfahren wie *Weighted–Least–Squares* oder *Generalized–Least–Squares* zur Bestimmung von $\hat{\mathbf{b}}$ herangezogen werden [20].

Systematische Abweichungen

Der in Gleichung (3.25) ausgewiesene systematische Fehler (Achsenabschnitt) b_0 läßt sich am einfachsten im Anschluß an die Regression durch

$$b_0 = \bar{p} - \hat{\bar{p}} \tag{3.42}$$

$$= \frac{\sum_{i=1}^{n_s} p_i}{n_s} - \frac{\sum_{i=1}^{n_s} \hat{p}_i}{n_s} \tag{3.43}$$

bestimmen. Dabei ist $\bar{p}$ der Mittelwert der Referenzwerte und $\hat{\bar{p}}$ der aus der Kalibration berechnete Mittelwert der Probeneigenschaft. Gleichung (3.40) wird dann erweitert zu

$$\mathbf{p} = \mathbf{b}_0 + \hat{\mathbf{p}} = \mathbf{b}_0 + \hat{\mathbf{b}}\mathbf{X} \quad . \tag{3.44}$$

Der Vektor $\hat{\mathbf{p}}$ ist der Schätzer der Probeneigenschaften $\mathbf{p}$, und $\mathbf{b}_0$ ist ein konstanter Vektor, dessen Einträge alle denselben Wert besitzen.

Eine Alternative bietet sich durch die Berechnung des Mittelwertes der Probeneigenschaft aus der Kalibration $\hat{\bar{p}}$ auf Basis der Faktorgewichte $\mathbf{C}$

$$\hat{\bar{p}} = \boldsymbol{\beta}^T \bar{\mathbf{c}} \quad , \tag{3.45}$$

wobei $\bar{\mathbf{c}}$ die über alle Standard-Spektren gemittelten Faktorgewichte enthält

$$\bar{\mathbf{c}} = \frac{1}{n_s} \left(\sum_{i=1}^{n_s} \mathbf{C}_{1,i} \ , \quad \sum_{i=1}^{n_s} \mathbf{C}_{2,i} \ , \cdots \ , \sum_{i=1}^{n_s} \mathbf{C}_{n_s,i} \right)^T \quad . \tag{3.46}$$

Für Datensätze, die über die Mittelwertspektren skaliert sind, ist der Mittelwert aller Faktorgewichte gleich Null und damit $\hat{\bar{p}} = 0$.

3.3 Reduktion der Daten

In einem Interview [17] von *Kim Esbensen* und *Paul Geladi* antwortete *Harald Martens* auf die Frage, was seiner Meinung nach dazu geführt hat, daß sich Wissenschaftler mit Chemometrie beschäftigen, kurz: «Too much data!»

Dieses Statement weist auf ein grundsätzliches Problem neuer Analysenmethoden hin. In immer kürzerer Zeit können immer mehr Daten erhoben werden, die dann „nur noch“ ausgewertet werden müssen. Die in der Spektrochemometrie anfallenden großen Datenmengen erforden ein hohes Maße an Verwaltungsaufwand, der Personal und technische Ressourcen bindet. Die Frage nach geeigneten Verfahren zur Datenkomprimierung und schnelleren Verarbeitung ist von hohem anwendungstechnischem Interesse. Die Mathematik bietet hier eine Reihe von Möglichkeiten, z.B. auf Basis von Wavelettransformationen oder der Faktoranalyse. Letztere ist weit verbreitet und beinhaltet die im folgenden beschriebenen statistischen Ansätze:

Die bei der Faktoranalyse ermittelten Eigenwerte entsprechen unmittelbar den durch die zugehörigen Eigenspektren repräsentierten Anteil an der Gesamtvarianz der Daten. Da die Eigenwerte schnell abnehmen, ist es daher möglich, mit wenigen Faktoren den Hauptanteil der in den Daten enthaltenen Information zu erfassen. Hierdurch bietet sich die Möglichkeit zu einer deutlichen Reduktion der Daten. Wird

nur eine bestimmte Auswahl von Faktoren berücksichtigt, so läßt sich der prozentuale Anteil erfaßter spektraler Varianz ($\mathfrak{I}$) durch folgendes Verhältnis beschreiben:

$$\text{spektrale Varianz } (\mathfrak{I})\ [\%] = \frac{\sum\limits_{\text{Auswahl}} \lambda_i}{\sum\limits_{\text{ges.}} \lambda_i} \cdot 100\% \quad . \tag{3.47}$$

λ_i sind hierbei die den Faktoren zugeordneten Eigenwerte.

Die Datenkomprimierungsrate ($\mathfrak{D}$) resultiert aus der Darstellung der gesamten Datenmatrix auf Basis einiger weniger Faktoren. Der prozentuale Anteil der Komprimierung (die sog. Datenkomprimierungsrate $\mathfrak{D}$) ergibt sich aus dem Verhältnis der Anzahl der Faktoren in der Auswahl (n_f) und der Zahl der Spaltenvektoren in der Datenmatrix (k):

$$\text{Datenkomprimierungsrate } (\mathfrak{D})\ [\%] = \left(1 - \frac{n_f}{k}\right) \cdot 100\% \quad . \tag{3.48}$$

Das in 3.2.1 beschriebene Verfahren zur Berechnung des Regressionsvektors $\hat{\mathbf{b}}$ besitzt zunächst keine Vorteile im Vergleich zur Berechnung auf Basis der ursprünglichen Spektren (3.29). Ein Vorteil ergibt sich erst dann, wenn nicht mehr alle Principal Components (PCs) aus $\mathbf{C}$ bzw. Faktoren aus $\mathbf{R}$ im Regressionsschritt berücksichtigt werden, sondern nur die Auswahl, die zu einem optimalen Kalibrationsmodell führt. Die Regressionsgleichung (3.28) läßt sich dann als

$$\hat{\mathbf{p}} = \mathbf{p}\mathbf{C}_r^T \left(\mathbf{R}_r\right)^+ \mathbf{X} \tag{3.49}$$

schreiben. Der Index r bezeichnet die reduzierten Matrizen, d.h. es werden nicht alle Spalten von $\mathbf{R}$ und Zeilen von $\mathbf{C}$ berücksichtigt.

3.3.1 Underfitting und Overfitting

Die Auswahl geeigneter Eigenvektor–Kombinationen ist im Zusammenhang mit PCR und PLS ein in der chemometrischen Literatur vielbeachtetes Problem [20, 41–58]. Die Entwicklung geeigneter Methoden zur Erstellung dieser Kombinationen ist Hauptgegenstand der vorliegenden Arbeit. Folgender Umstand stellt in diesem Zusammenhang ein Dilemma dar: Einerseits wird ein möglichst optimales Kalibrationsmodell zur Beschreibung der zu kalibrierenden Eigenschaft gesucht[6]. Andererseits sollen zur Kalibration so wenig latente Variablen wie nötig herangezogen werden, um den Anteil redundanter Information so klein wie möglich zu halten und eine gute Datenreduktion zu erreichen. Abbildung 3.5 verdeutlicht die sich daraus ableitende Problematik. Während die Selektion von zu wenigen Eigenvektoren zu einer

[6] Der Vorhersagefehler wird hier als Indikator eingesetzt (vgl. 3.5.2).

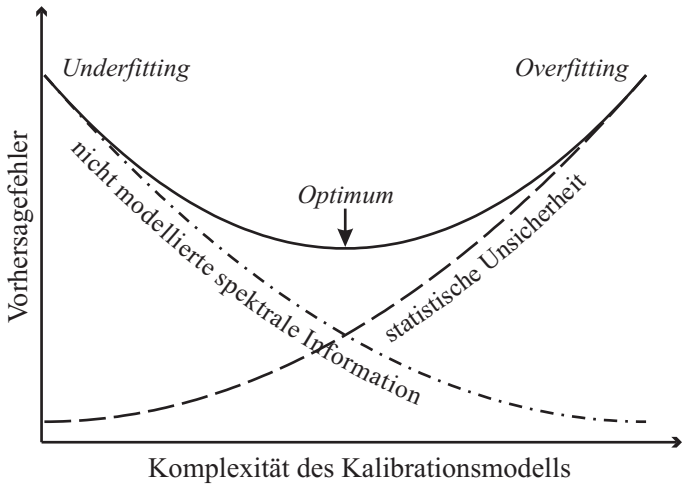


Abbildung 3.5: Die Abhängigkeit des Fehlers von der Komplexität des Kalibrationsmodells.

niedrigen Modellkomplexität führt, werden spektrale Effekte, die zur vollständigen Beschreibung der Probeneigenschaft notwendig sind, nur ungenügend erfaßt. Auf der anderen Seite führt eine zu hohe Modellkomplexität mit einer Vielzahl von Faktoren zum Anstieg der statistischen Unsicherheit im Kalibrationsmodell. Der Hintergrund für diesen Effekt wird aus der Diskussion der statistischen Zusammenhänge deutlich: Der Vorhersagefehler aus der Kalibration, der als Indikator zur Auswahl eines geeigneten Modells herangezogen wird, setzt sich hauptsächlich aus zwei Anteilen zusammen. Einerseits aus einem systematischen Fehler, der durch nicht modellierte spektrale Anteile hervorgerufen wird und andererseits durch zufällig verteiltes Rauschen in den Messungen. Unter der Vorannahme, daß die vorherzusagenden Probeneigenschaften ausreichend genau durch die latenten Variablen (PCs und Faktoren) vorhergesagt werden, nimmt der durch unberücksichtigte spektrale Anteile bedingte Fehler mit zunehmender Anzahl der im Modell berücksichtigten Faktoren ab. Andererseits steigt im gleichen Maße die statistische Unsicherheit der Vorhersage. Dieser Effekt ist auf die zunehmende Zahl unabhängiger Modellparameter, welche von den verfügbaren Daten geschätzt werden, zurückzuführen. Anders ausgedrückt heißt dies, daß die Zahl der Freiheitgerade im Modell sinkt und dadurch der Fehler durch die statistische Unsicherheit wächst (*s.* Gl. (3.53)).

Dieser Zusammenhang ist von entscheidender Bedeutung, wenn es um die Selektion latenter Variablen zur Modell-Erstellung geht. Eine optimale Kalibrationsgüte resultiert nur dann, wenn sich beide Fehler die Waage halten. Werden zuwenige Faktoren in einem Kalibrationsmodell berücksichtigt, dann spricht man von *underfitting*. Der umgekehrte Fall wird als *overfitting* bezeichnet.

In erster Linie kommt es bei der Modellbildung darauf an, möglichst *robuste* Kali-

brationen zu erstellen. Robust[7] bedeutet hier, daß die, mit Hilfe der zum Zeitpunkt der Kalibration zur Verfügung stehenden Kalibrations- und Validationsstandards, ermittelten Analysenfehler bei späteren Durchführungen von Analysen im praktischen Betrieb nicht deutlich zunehmen. Da davon auszugehen ist, daß eine Kalibration um so störanfälliger wird, je mehr spektrale Information sie beinhaltet, erscheint es im Sinne der oben definierten Robustheit zweckmäßig, auf redundante, das heißt für die Erreichung einer zufriedenstellenden Analysengenauigkeit nicht notwendige, spektrale Information zu verzichten. Es ist daher sinnvoll, nur so wenige Faktoren wie möglich im Kalibrationsmodell zu berücksichtigen.

3.4 Problematische Aspekte von Kalibrationsmodellen

In den vorangegangenen Abschnitten wurde dargelegt, wie mit chemometrischen Verfahren univariate oder multivariate Kalibrationsmodelle erstellt werden. Modelle werden benötigt, um den Zusammenhang zwischen der Meßgröße $\mathbf{X}$ und der gesuchten Information $\mathbf{p}$ herzustellen. Gerade multivariate Kalibrationen bedürfen aber einer vorsichtigen Modellierung der spekralen Daten [20, 59].

Multivariate Kalibrationen berücksichtigen drei verschiedene *„Ebenen“* von Modellen. Alle drei Modell-Ebenen sind miteinander verknüpft und können gegenseitig circulativ auf Plausibilität geprüft werden. Um eine gute Kalibration zu gewährleisten, ist es notwendig,

1. im Sinne der Anforderungen an das Modell ein klares Ziel zu formulieren,
2. ein hinreichend gutes Validationsverfahren zu definieren und anzuwenden,
3. eine präzise und interpretierbare Vorhersage aus dem Kalibrationsmodell abzuleiten.

   Für die Übertragung dieses Vorgehensmodells auf multivariate Kalibrationen ergeben sich folgende Schritte:

a. Zunächst ist unsere Vorstellung vom Zusammenhang der Daten $\mathbf{X}$ und $\mathbf{p}$ ausschlaggebend für die Wahl eines Ansatzes zur Modellbildung.

b. Nach unserer Vorstellung wird ein abstraktes mathematisches Modell formuliert, daß die erwarteten Zusammenhänge wiedergeben soll. Das Regressionsmodell in Gleichung (3.28) ist ein Beispiel dafür. Dieses mathematische Modell

[7] Der Begriff der Robustheit von Kalibrationen darf nicht mit der im Kontext von Genetischen Algorithmen gebräuchlichen Bezeichnung verwechselt werden. Während es im ersten Fall um die Beschreibung eines ausgewogenen Kalibrationsmodells geht, steht im zweiten das zuverlässige Erreichen des globalen Optimums im Vordergrund.

muß der Vorstellung vom Zusammenhang der Daten genügen, beinhaltet aber eine gewisse *„Unschärfe“* und liefert keine *„vollständige“* Beschreibung.

c. Dieses abstrakte Modell beinhaltet unbekannte Koeffizienten (wie z.B. **b**). Die Koeffizienten werden während der Kalibration bestimmt und gehen in das abschließende Kalibrationsmodell ein. Dieses abschließende Modell soll den gesuchten Zusammenhang zwischen den Daten **X** und **p** beschreiben und kann mit der anfänglichen Vorstellung vom Zusammenhang zwischen den Daten verglichen werden.

Grundsätzlich muß berücksichtigt werden, daß Modelle immer nur Arbeitshypothesen darstellen, und ihre Leistungsfähigkeit beschränkt ist. Unerwartete Phänomene in der Kalibration müssen daher untersucht werden und in eine Korrektur der einzelnen Modell-Ebenen Eingang finden.

3.5 Validierung von Kalibrationsmodellen

Aus der Regressionsrechnung ergeben sich eine Reihe statistischer Schätzer, die Kalibrationsmodelle charakterisieren. Sie bieten einen Überblick über das Maß der Datenanpassung durch das Modell und den zu erwartenden Vorhersagefehler der Kalibration.

3.5.1 Restvarianzwert

Das Bestimmtheitsmaß (R^2) gibt den Anteil der Varianz der Probeneigenschaft an, die durch das berechnete Kalibrationsmodell beschrieben wird:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TPV}} = 1 - \frac{\sum_{i=1}^{n_s} (\hat{p}_i - p_i)^2}{\sum_{i=1}^{n_s} (p_i - \bar{p})^2} \quad . \tag{3.50}$$

Der Restvarianzwert (RSS — *residual sum of squares*) gibt hier die durch das Modell nicht beschriebene Varianz und TPV (*total property variance*) die Gesamtvarianz der Probeneigenschaft wieder. $\hat{p}_i$ und p_i bezeichnen den vorhergesagten bzw. den vorgegebenen Wert für die Probeneigenschaft des i-ten Standards und $\bar{p}$ den entsprechenden Mittelwert. Die Zahl n_s gibt die Anzahl der Kalibrationsstandards wieder. Je kleiner der Quotient aus RSS und TPV ist, desto besser wird die zu kalibrierende Probeneigenschaft durch das Modell erfaßt und wiedergegeben.

Im allgemeinen wird in der Literatur neben dem RSS–Wert der sogenannte PRESS–Wert (*predicted residual error-sum of squares*) zur Bewertung von Kali-

brationsmodellen herangezogen. Er wird analog zum RSS–Wert durch

$$\mathrm{PRESS} = \sum_{i=1}^{n_e} (\hat{p}_i - p_i)^2 \tag{3.51}$$

berechnet, beschreibt aber im Gegensatz zu diesem den Vorhersagefehler für Proben, die nicht Teil des Kalibrationsdatensatzes sind (Validationsspektren).

3.5.2 Standard Error of Estimate

Der *standard error of estimate* (SEE), bezeichnet den absoluten Vorhersagefehler in der Einheit der Kenngröße und ergibt sich aus der Quadratwurzel des Quotienten der *residual sum of squares* (RSS) und der Anzahl der Freiheitsgrade:

$$\mathrm{SEE} = \sqrt{\frac{\mathrm{RSS}}{n_s - n_f - 1}} \tag{3.52}$$

$$= \sqrt{\frac{\sum_{i=1}^{n_s} (\hat{p}_i - p_i)^2}{n_s - n_f - 1}} \quad . \tag{3.53}$$

Die Zahl der Freiheitsgrade in der Kalibration ergibt sich aus der Anzahl der Kalibrationsstandards n_s und der Anzahl berücksichtigter Faktoren (oder Principal Components) n_f. Der SEE-Wert ist für die Beurteilung der Vorhersagequalität eines Kalibrationsmodells nur bedingt aussagekräftig, da er im allgemeinen einen zu kleinen Analysefehler widerspiegelt und anfällig für *Overfitting–Effekte* (*s.* 3.3) ist.

3.5.3 Standard Error of Prediction

Unter der Bezeichnung *standard error of prediction* (SEP) werden zwei unterschiedliche Arten von Standardabweichungen zusammengefaßt: der *standard error of prediction estimate* und der *standard error of prediction* einer *full–cross–validation*. Beide sind Schätzer des zukünftig zu erwartenden Analysefehlers und werden auf Basis der Kalibrationsdaten bestimmt.

Standard Error of Prediction Estimate

Der SEP_{est}-Wert, der *standard error of prediction estimate*, ermöglicht eine Abschätzung des Fehlers bei der Vorhersage von unbekannten Proben unter Verwendung des Kalibrationsdatensatzes:

$$\mathrm{SEP}_{est} = \sqrt{\frac{\sum_{i=1}^{n_s} (\hat{p}_i - p_i)^2}{n_s - 1}} \quad . \tag{3.54}$$

Der SEP_{est}-Wert wird ermittelt, indem sukzessiv jeder Standard aus dem Datensatz herausgenommen wird und mit den übrigen $n_s - 1$ Standards das Regressionsmodell neu berechnet wird. Der Eigenschaftswert des herausgenommenen Standards wird dann auf Basis des neuen Kalibrationsmodells berechnet. Die Berechnung erfolgt ohne eine Neuberechnung der Faktormatrix. Diese Abfolge wird für alle Standards im Kalibrationssatz durchgeführt und als *leave one out* oder *cross validation* bezeichnet. Aus den Abweichungen zwischen den errechneten Probeneigenschaften und den mittels eines Referenzverfahrens bestimmten Werten wird dann die Standardabweichung berechnet.

Diese Art der Kreuzvalidation eröffnet einen schnellen Weg zur Abschätzung des zu erwartenden Analysefehlers. Sie liefert Ergebnisse, die denen einer Full-Cross-Validation sehr ähnlich sind.

Standard Error of Prediction — Full-Cross-Validation

Bei einer *full-cross-validation* wird jeweils ein Standard aus dem Kalibrationsdatensatz herausgenommen und mit den restlichen Standards eine neue Kalibration erstellt, wobei im Gegensatz zur Cross–Validation jeweils eine neue Faktorzerlegung berechnet wird. Mit dem so erhaltenen Modell wird dann der Eigenschaftswert des herausgenommenen Standards berechnet. Diese Prozedur wird sukzessiv mit allen Standards durchgeführt. Die Berechnung der Standardabweichung erfolgt analog zu Gleichung (3.54).

Der SEP-Wert auf Basis einer Full–Cross–Validation ist diejenige Größe, die dem zu erwartenden Analysefehler unabhängiger Validationsdaten am nächsten kommt, diesen aber nicht ersetzt.

Generalized Cross–Validation

Das umständliche und sehr zeitaufwendige Verfahren der Full–Cross–Validation zur Berechnung aussagekräftiger Parameter auf Basis des Kalibrationsdatensatzes kann durch die Anwendung der *generalized cross–validation* [60] wesentlich vereinfacht werden.

$$\mathrm{SEP}_{GCV} = \sqrt{\frac{n_s \mathrm{RSS}}{n_s - n_f}} \quad . \tag{3.55}$$

Der SEP_{GCV}-Wert ist bei bekannter Restvarianz (RSS), Anzahl der Kalibrationsspektren (n_s) und der Anzahl im Modell berücksichtigter Faktoren (n_f) einfach zu berechnen. Der mit dem GCV–Verfahren ermittelte Vorhersagefehler ist aus theoretischer Sicht mindestens gleichwertig mit dem aus der konventionellen Full–Cross–Validation ermittelten SEP–Wert.

3.5.4 Standard Error of Analysis

Zur experimentellen Validierung der Kalibration ist es erforderlich, regelmäßig in gewissen Zeitabständen nach Erstellung des Kalibrationsdatensatzes unabhängige Standards herzustellen und zu analysieren. Der aus der Validation mit diesen Standards resultierende Fehler des Kalibrationsmodells wird als *standard error of analysis* oder *root mean square error of prediction* (RMSEP) bezeichnet

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_e} (\hat{p}_i - p_i)^2}{n_e}} \quad . \tag{3.56}$$

Er zeigt, wie gut die Kalibration in der Praxis tatsächlich die zu kalibrierende Probeneigenschaft der n_e Validationsstandards erfaßt und liegt erfahrungsgemäß um den Faktor zwei bis drei höher als der SEP-Wert der Full-Cross-Validation [61, 62].

Die Aussagekraft ($\mathfrak{A}$) der einzelnen Standardabweichungen in bezug auf die Größenordnung der zukünftig zu erwartenden Analysefehler staffelt sich daher wie folgt:

$$\mathfrak{A}(\mathrm{SEE}) \leq \mathfrak{A}(\mathrm{SEP}_{est.}) \leq \mathfrak{A}(\mathrm{SEP}) \leq \mathfrak{A}(\mathrm{SEP}_{GCV}) \leq \mathfrak{A}(\mathrm{RMSEP}) \quad . \tag{3.57}$$