4 Chemometrie

Die Auswertung spektroskopischer Daten erfolgt heute nahezu ausschließlich rechnerunterstützt. Die mathematisch-statischen Methoden, die dabei zum Einsatz kommen, werden unter dem Begriff der Chemometrie zusammengefaßt [47–49]. Chemometrie kann als die chemische Disziplin verstanden werden, die mathematische und statistische Methoden nutzt, um optimale Meßprozeduren und Experimente zu entwerfen oder auszuwählen [50, 51]. Weiter liefert sie bei der Analyse chemischer Daten das Maximum an relevanter chemischer Information [52, 53]. Da die Chemometrie eine allgemeine und weitgefächerte Disziplin mit vielen unterschiedlichen Einsatzmöglichkeiten ist, wird für die Analyse spektroskopischer Daten im Arbeitskreis der Begriff „Spektrochemometrie“ verwendet. Hierbei werden im allgemeinen multivariate Rechenverfahren eingesetzt [54]. Moderne chemometrische Verfahren sind sehr leistungsfähig und werden aufgrund der rasanten Entwicklung der Computertechnik seit einigen Jahren immer mehr eingesetzt. So werden neben der Faktoranalyse ebenfalls Mustererkennungstechniken wie die Clusteranalyse oder Neuronale Netze verwendet. Während mit univariaten Verfahren nur der Zusammenhang zwischen einer Variablen und einer Eigenschaft herstellbar ist, können mit Hilfe von multivariaten Verfahren Zusammenhänge zwischen mehreren Variablen gleichzeitig beschrieben werden. Multivariate Analysenverfahren kommen immer dann zum Einsatz, wenn aus sehr umfangreichen Datenmengen für eine bestimmte Meßgröße spezifische Informationen herausgefiltert werden sollen. Eine Anwendung ist die spektrometrisch quantitative Analyse (z.B. wenn die Konzentrationen einzelner Komponenten in einer komplexen Matrix oder einem Mehrkomponentengemisch bestimmt werden sollen) [55].

Zur Lösung eines solchen Analysenproblems werden heute verschiedene Verfahren eingesetzt, wobei die faktoranalytischen Verfahren Principal Component Regression (PCR) und Partial Least Squares (PLS) wesentliche Vorzüge anderer multivariater Auswerteverfahren wie der Classical Least Squares (K-Matrix Methode) und der Inverse Least Squares (P-Matrix Methode) in sich vereinen [56]. Dazu gehört die Invarianz in Bezug auf die Anzahl der zu berücksichtigenden Komponenten; d.h. eine Analyse kann auch dann durchgeführt werden, wenn nur die Konzentration der zu bestimmenden Komponente, nicht aber die restliche Zusammensetzung der Matrix in den Standards bekannt ist. Außerdem handelt es sich bei der PCR und der PLS um Vollspektrenmethoden [56, 57], d.h. es geht im allgemeinen das komplette Spektrum in die Rechnung ein. Beide Verfahren lie-

fern annähernd gleich gute Ergebnisse, wie in vielen Anwendungen gezeigt werden konnte [56, 58, 59].

Die auf dem Lambert-Beerschen Gesetz beruhende Annahme der linearen Superposition von Spektren in Gemischen ist ein physikalisches Modell, welches das Spektrum eines Multikomponentensystems allerdings nicht genau beschreibt, sondern eine gewisse Vereinfachung der physikalischen Realität darstellt. Die Vereinfachung besteht unter anderem darin, daß chemische und physikalische Wechselwirkungen zwischen den Komponenten einer Mischung (z.B. durch Assoziation, Wasserstoffbrückenbindungen etc.) vernachlässigt werden. Allerdings sollte bei der hier vorliegenden Situation verdünnter Gase davon auszugehen sein, daß solche Effekte nur eine geringe Rolle spielen. Die Anwendung des Lambert-Beerschen Gesetzes auf ein Multikomponentensystem führt zu einem entsprechenden mathematischen Modell. In der Chemometrie wird von einem Kalibrationsmodell gesprochen. Ein solches Modell wird mit einer Validation, bei der eine Voraussage der kalibrierten Eigenschaften für unbekannte Proben erfolgt, auf die Nützlichkeit oder Gültigkeit desselben hin überprüft.

Beide, das mathematische Modell, das im vorliegenden Fall durch ein lineares Gleichungssystem gegeben ist, und das physikalische Modell besitzen ihre eigenen objektiven Eigenschaften [60]. Auf jeden Fall muß vorab die Entscheidung getroffen werden, ob das mathematische Modell (Kalibrationsmodell) eine annehmbare Beschreibung des physikalischen Modells liefert. Da eine Kalibration eine Approximation eines mathematischen Modells an experimentell gewonnenen Daten darstellt, muß das Kalibrationsmodell auch gewisse Annahmen über die Art der zu erwartenden Fehler und eine Vorgehensweise für deren Minimierung enthalten. Auf diesbezüglich unterschiedlichen statistischen Voraussetzungen beruhen im wesentlichen die Unterschiede zwischen den im folgenden beschriebenen Kalibrationsmodellen.

4.1 Regressionsmodelle

Da es sich bei der NDIR-Spektroskopie um keine „Vollspektrenmethode“ handelt, scheiden die multivariaten faktoranalytischen Verfahren zur Kalibrierung des Verfahrens aus; als Meßgrößen stehen lediglich die Phasenlagen und Amplituden bzw. der auf den Phasenschieber des Lockin-Verstärkers (SQ) projizierten Anteile des Gesamtsignals zur Verfügung. Zu berücksichtigende Einflußgrößen bei diesem Meßverfahren sind zum einen die Konzentrationen der zu erwartenden Gaskomponenten und zum anderen die Gaszusammensetzungen innerhalb des Detektors. Hinzu kommt die Modulationsfrequenz des Choppers, so daß es sich je nach Analysensystem um Abhängigkeiten von bis zu fünf Parametern gleichzeitig handeln kann. Zur Vereinfachung erfolgt die Beschreibung bei einer festen Chopperfrequenz. Zunächst wird auch die Konzentration der Füllgase im Detektor konstant gehalten. Das zu bestimmende Kalibrationsmodell basiert somit auf

einem Datensatz, der für unterschiedliche Zusammensetzungen des Zweikomponentengemisches die Konzentrationen der beiden Komponenten, die Amplitude und die Phase des Meßsignals enthält. Ferner wird die Phasenlage für die Signale der Reinkomponenten benötigt. Sind diese konzentrationsunabhängig, so ist für jede Reinkomponente nur eine Messung erforderlich, d.h. im endgültigen Kalibrationsmodell werden die Reinkomponenten mittels geeigneter Regressionsverfahren an diesen Datensatz angepaßt. Erst in einem zweiten Schritt werden zusätzlich auch die Konzentrationen der Detektorfüllgase variiert, um dadurch eine Detektoralterung zu simulieren. Die üblichen Verfahren der univariaten bzw. bivariaten linearen Regression [47] führen hierbei nicht immer zum Erfolg. Daher wurden in der vorliegenden Arbeit spezielle nichtlineare Regressionsverfahren (z.B. Levenberg-Marquardt-Verfahren s.Anhang), Response Surface Methoden (s.Anhang) und Krigingmethoden (s.Anhang) eingesetzt.

4.2 Bewertungs- und Auswerteverfahren für Datenanalysen

4.2.1 Validation von Regressionsfunktionen

Bei gegebenen Wertepaaren (Konzentration/Signal) stellt sich im Rahmen einer Kalibration die Frage, ob der entsprechende Zusammenhang durch ein mathematisches Modell in Form von einfachen Gleichungen beschrieben werden kann und ob die Beschreibung durch das Modell gelungen ist oder nicht [61, 62]. Bei der Wahl des Modells sind dem Anwender viele Möglichkeiten gegeben. Von einfachen linearen oder polynomischen Modellen mit wenigen Parametern bis hin zu extrem komplexen Modellen mit vielen Parametern ist alles möglich. Unabhängig davon, ob es sich bei dem dann ausgesuchten Modell um ein lineares oder nichtlineares handelt, eine erste Aussage wie gut dieses Modell ist, liefert das Bestimmtheitsmaß, das den Anteil der Varianz der Probeneigenschaft angibt, welche durch das berechnete Kalibrationsmodell beschrieben wird. Dieser Wert ist definiert als:

$$R^2 = \frac{QS_{Korr} - QS_R}{QS_{Korr}} \tag{4.1}$$

$$R^2 = 1 - \frac{QS_R}{QS_{Korr}} \tag{4.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4.3}$$

Hierbei ist QS_R oder RSS die *residual sum of squares*, welche die durch das Modell nicht beschriebene Varianz angibt und QS_{Korr} die Gesamtvarianz mit der die Streuung der Probeneigenschaft um den Mittelwert angegeben wird.

$$\mathrm{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{4.4}$$

$\hat{y}_i$ und y_i bezeichnen den vorhergesagten bzw. den vorgegebenen Wert für die Probeneigenschaft des i-ten Standards und $\bar{y}$ deren Mittelwert. Die Zahl n gibt die Anzahl der Kalibrationsstandards wieder. Je kleiner der Quotient aus RSS und QS_{Korr} ist, desto besser wird die zu kalibrierende Probeneigenschaft durch das Modell erfaßt und wiedergegeben. Die Quadratwurzel des Bestimmtheitsmaßes (R^2) ist der Korrelationskoeffizient (r). Ein „perfekter Fit“ zeichnet sich durch den Wert Null für die *residual sum of squares* und den Wert Eins für den Korrelationskoeffizienten aus. Je besser die Anpassung der abhängigen und der unabhängigen Werte durch das angenommene Modell gelungen ist, desto näher liegt der Wert bei eins.

$$r = \sqrt{R^2} = \sqrt{1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{4.5}$$

Eine Aussage über die Güte einer Kalibration [63, 64] ist mit den folgenden drei Werten möglich. Der **SEE–Wert** (*standard error of estimate*) bezeichnet den absoluten Fehler bei der Vorhersage der zu bestimmenden Eigenschaft y_i. Er ergibt sich aus der Quadratwurzel des Quotienten der *residual sum of squares* (RSS) und der Anzahl der Freiheitsgrade.

$$\mathrm{SEE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n - k - 1}} \tag{4.6}$$

Der Wert n steht für die Anzahl der Kalibartionsstandards und k für die Anzahl ausgewählter Faktoren oder Parameter. Eine andere wichtige Größe, welche die Güte eines Fits beschreibt, ist der **SEP–Wert** (*standard error of prediction*). Der SEP–Wert, ermöglicht eine Abschätzung des Fehlers bei der Vorhersage von unbekannten Proben unter Verwendung eines Kalibrationsdatensatzes bei dem vor jeder Kalibration jeweils ein Standard herausgenommen wird und mit den restlichen Standards eine neue Kalibration berechnet wird. Für den herausgenommenen Standard wird dann über das ermittelte Modell die Eigenschaft

vorhergesagt.

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1}} \tag{4.7}$$

Zur Validierung der Kalibration ist es erforderlich eine Anzahl unabhängiger kalibrationsexterner Standards n_{ext} herzustellen und zu analysieren. Der aus der Validation mit diesen Standards resultierende Fehler des Kalibrationsmodells wird als **SEA–Wert** oder *standard error of analysis* bezeichnet.

$$\mathrm{SEA} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{n_{ext}}} \tag{4.8}$$

In dieser Arbeit wird das Hauptaugenmerk bei der Bewertung der Regressionsfunktionen auf die Werte für SEE- und SEA-Wert gelegt.

4.2.2 Varianz

Die Varianzanalyse (*analysis of variance*, ANOVA) untersucht die Auswirkung von unabhängigen Variablen (Einflußgrößen) auf die Mittelwerte der beobachteten (abhängigen) Variable. Im Vordergrund dieser Untersuchung steht die Frage, ob die unabhängigen Variablen einen signifikanten Einfluß auf die abhängige Variable ausüben, oder ob sich die beobachteten Mittelwerte nur rein zufällig unterscheiden. Die Varianzanalyse basiert dabei auf der Zerlegung der Gesamt-

Tabelle 4.1: Ergebnistabelle für die einfache Varianzanalyse.

Streuung	Freiheitsgrad f	Quadratsumme QS	mittlere Quadratsumme MQS	F-Test F
Regression (Effekt)	k	QS_{Fakt}	$\frac{QS_{Fakt}}{k}$	$\frac{QS_{Fakt}(n-k-1)}{QS_R(k)}$
Fehler (Residuen)	n-k-1	QS_R	$\frac{QS_R}{n-k-1}$	
Total	n-1	QS_{Korr}		

varianz in verschiedene Varianzteile. Bei der einfachen Varianzanalyse wird nur der Einfluß einer unabhängigen Variable betrachtet. Die ANOVA-Tabelle bei dieser Varianzanalyse ist in Tabelle 4.1 wiedergegeben. Dabei steht QS_{Fakt} für den

Effekt, den die unabhängige Einflußgröße an der Gesamtvarianz verursacht. Ein weiterer Teil der Gesamtvarianz entsteht durch Versuchsfehler und wird durch die Größe QS_R beschrieben. Die Prüfung auf signifikante Unterschiede erfolgt über die F-Verteilung mit der in Tabelle 4.1 beschriebenen Prüfgröße. Die beteiligten Größen sind dabei wie folgt definiert:

- Gesamtvarianz

$$QS_G = \sum_{i=1}^{n} y_i^2 \tag{4.9}$$

- Mittlere Gesamtvarianz

$$QS_M = n\bar{y}^2 \tag{4.10}$$

- Um den Mittelwert korrigierte Gesamtvarianz

$$QS_{Korr} = QS_G - QS_M = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.11}$$

- Modellvarianz oder durch die Regression erklärte Varianz

$$QS_{Fakt} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{4.12}$$

- Residuumsvarianz oder durch die Regression nicht erklärte Varianz

$$QS_R = QS_{Korr} - QS_{Fakt} = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} \epsilon_i^2 \tag{4.13}$$