4 Faktor–Selektion

Das wesentliche Problem bei der Kalibrationserstellung liegt darin, diejenige Kombination latenter Variablen, das heißt in diesem Fall von Faktoren, zu finden, die eine optimale Kalibration der Probeneigenschaft ermöglicht. Doch wie läßt sich *optimal* in diesem Kontext definieren? Bei der Modellbildung steht vor allem die zuverlässige Vorhersage der Probeneigenschaften von Standards, die nicht im Kalibrationsdatensatz enthalten sind, im Vordergrund. Ein Kalibrationsmodell, das diese Anforderung erfüllt, wird als robust bezeichnet. Die Hauptgefahr besteht darin, durch die Berücksichtigung von zuvielen Faktoren, zufällig auftretende Schwankungen in den spektralen Daten der Kalibrationsspektren mitzukalibrieren und dadurch ein Overfitting zu erzeugen.

Die Anwendung deterministischer Verfahren zur Selektion der Faktoren stand in der Vergangenheit im Vordergrund. Hierbei werden sukzessiv Faktoren einem Kalibrationsmodell hinzugefügt (oder aus diesem entfernt) bis sich keine weitere Verbesserung der Kalibrationsergebnisse mehr feststellen läßt. Diese klassischen, deterministischen Verfahren sind zwar gut automatisierbar, bringen aber durch ihren Aufbau zwei kritische Randbedingungen mit sich: Zum einen beginnen diese Methoden ihre Optimierung von einem bestimmten Punkt aus. Die Optima, die durch deterministische Methoden gefunden werden, sind daher möglicherweise nur lokale Optima in der Nachbarschaft dieses Startwertes. Zum anderen gehen deterministische Methoden von einer kontinuierlichen Verbesserung des zu optimierenden Kalibrationsmodells durch Hinzufügen oder Entfernen einzelner Faktoren aus. Für ein solches Verfahren ist es daher unmöglich, eine Verbesserung zu erkennen, die durch Hinzufügen eines bestimmten Faktors und gleichzeitiges Weglassen eines anderen erzielt wird. Dies führt unter Umständen dazu, daß einzelne möglicherweise gute Lösungen übersehen werden.

In bezug auf die Automatisierung des Selektionsprozesses ist ein weiterer Aspekt wichtig: Es muß ein Kriterium angewandt werden, das eine differenzierte Aussage darüber zuläßt, ob die Veränderung eines Kalibrationsmodells eine Verbesserung darstellt oder nicht. Diese Kriterien werden in der Regel auf Basis statistischer Tests oder empirisch ermittelter Größen festgelegt. Häufig ist der Vergleich der Vorhersagefehler (die Varianz oder Standardabweichung) von zwei Kalibrationsmodellen die Basis dieser Kriterien. *E. Malinowski* führte dazu eine Reihe von Untersuchungen durch und schlug die Verwendung von *Fischer*–Tests (F–Tests) zum Vergleich der

Varianzen vor [39]. Hierauf wird in 4.1.1 und 4.1.2 näher eingegangen.

4.1 Klassische Selektionsverfahren

Die klassischen Verfahren zur Faktor–Selektion stützen sich in erster Linie auf deterministische Methoden zur Differenzierung bestimmter Lösungen oder Faktorkombinationen. Erster Schritt ist in aller Regel eine Unterteilung der in der Principal Component Analysis (PCA) bestimmten Faktoren in *primäre* und *sekundäre* Faktoren. Diese Einteilung beruht darauf, daß die in der Reihenfolge ihrer abnehmenden Eigenwerte geordneten Faktoren, insbesondere bei großen Kalibrationsdatensätzen, ab einem bestimmten unteren Grenzwert ihrer Eigenwerte vernachlässigt werden können. Als sekundäre Faktoren werden diejenigen bezeichnet, deren Eigenwerte diesen Grenzwert unterschreiten und die lediglich Rauschen oder Minoritätseffekte darstellen. Alle übrigen Faktoren werden als primäre Faktoren bezeichnet. Sie repräsentieren die wesentliche spektrale Information der spektralen Daten. Die Anzahl primärer Faktoren entspricht dem Rang des Faktorraums (n_r).

Zur eindeutigen Differenzierung zwischen primären und sekundären Faktoren ist die Festlegung eines geeigneten Grenzwerts der Eigenwerte erforderlich. Als Kriterium mit dem ermittelt wird, ob ein Faktor noch einen signifikanten Beitrag zur Beschreibung des Datensatzes liefert oder nicht, haben sich eine Reihe empirischer Testverfahren durchgesetzt, die auf die Fehlertheorie von Malinowski [63] zurückgehen.

Malinowski hat die sogenannte Indikator–Funktion (IND) eingeführt, mittels derer die Zahl primärer Faktoren bestimmt werden kann. Letztere ergibt sich hierbei aus der Bestimmung des Minimums der Indikator–Funktion

$$\mathrm{IND} = \frac{\mathrm{RE}}{n_s - n_f} \quad . \tag{4.1}$$

RE ist der *Real–Error* und berechnet sich nach:

$$\mathrm{RE} = \sqrt{\frac{\sum_{i=n_f+1}^{n_s} \lambda_i}{n_\nu \left(n_s - n_f\right)}} \quad . \tag{4.2}$$

RE ist demnach eine Funktion der Anzahl der in die Kalibration eingehenden Spektren n_s, der berücksichtigten Faktoren n_f, der Eigenwerte λ_i der sekundären Faktoren $(n_s - n_f)$ und der Anzahl der Datenpunkte im Spektrum n_ν.

Neben der IND-Funktion ist es auch durch einen Varianztest möglich, die Bedeutung einzelner Faktoren für die Beschreibung eines Datensatzes zu bestimmen [39]. Malinowski konnte zeigen, daß die sogenannten reduzierten Eigenwerte λ_i^r in einem

direkt proportionalen Zusammenhang zur Varianz des Datensatzes stehen (*s.* 3.2.1 und 3.2.1). Die reduzierten Eigenwerte berechnen sich wie folgt:

$$\lambda_i^r = \frac{\lambda_i}{(n_\nu - i - 1)\,(n_s - i - 1)} \quad \text{mit } i = 1, \ldots, \min(n_f - 2, n_s - 2) \quad . \tag{4.3}$$

Die relativ kleinen reduzierten Eigenwerte, die auf Rauschen zurückzuführen sind, haben in etwa die gleiche Größenordnung und sind durch einen Varianztest von den wesentlich größeren reduzierten Eigenwerten, die einen signifikanten Beitrag spektraler Information beinhalten, unterscheidbar. Zur Differenzierung zwischen den Eigenwerten wird daher ein F–Test der reduzierten Eigenwerte durchgeführt.

Der kleinste Eigenwert wird dabei als zum Rauschen zugehörig definiert. Mit einem F–Test wird in einem paarweisen Vergleich ermittelt, ob sich der nächstgrößere reduzierte Eigenwert signifikant vom vorhergehenden unterscheidet. Ist dieser reduzierte Eigenwert in seiner Größenordnung dem zum Rauschen gezählten Eigenwert gleich, so werden diese beiden Eigenwerte zu einem gewichteten Durchschnitt eines Rausch-Eigenwertes zusammengefaßt und gegen den nächstgrößeren Eigenwert getestet. Dieser Vergleich wird solange wiederholt, bis ein Eigenwert sich signifikant von den Rausch-Eigenwerten unterscheidet [2, 39].

4.1.1 Deterministische Auswahlverfahren

Letztendlich gehen jedoch nicht alle primären Faktoren, die auf eine der oben beschriebenen Weisen bestimmt wurden, in eine Regressionsrechnung ein, die das Kalibrationsmodell beschreibt. Eine Reihe zusätzlicher deterministischer Verfahren wird hier zur weiteren Abstufung und Selektion von Faktoren herangezogen.

Das einfachste Verfahren zur Faktor–Selektion ist die *top–down variable selection* (TD). Die Faktoren werden nacheinander solange in ein Modell aufgenommen, bis anhand eines geeigneten Kriteriums keine weitere Verbesserung der Kalibration mehr festgestellt wird [43, 44, 47, 58].

Selektiver ist die *forward–stepwise variable selection* (FS). Das Verfahren beginnt mit der Bestimmung desjenigen Faktors, der den kleinsten Fehler für ein bestimmtes Kriterium[1] liefert. Dann wird unter den verbleibenden Faktoren derjenige ausgewählt, der zur größten Verbesserung der Beobachtungsfunktion führt. Das Verfahren wird solange fortgesetzt, bis etwa durch F–Tests keine signifikante Verbesserung des Kriteriums mehr beobachtet wird. Die *backward–stepwise variable selection* (BS) ist ein analoges Verfahren, welches aber in umgekehrter Weise verläuft. Es wird mit dem vollständigen Satz von Faktoren begonnen, der dann solange durch sukzessives Herausfallenlassen von Faktoren (in der Reihenfolge aufsteigender Eigenwerte)

[1] Beobachtungsfunktion — in der Regel die Standardabweichung der Vorhersage, z.B. der SEP–Wert.

reduziert wird, bis sich eine signifikante Verschlechterung des beobachteten Kriteriums ergibt. Aus Gründen der Stabilität wird ein Kalibrationsmodell angestrebt, das möglichst wenige Faktoren berücksichtigt. Deshalb wird mit Hilfe eines F–Tests beurteilt, inwieweit sich der SEP–Wert signifikant verändert, wenn weitere Faktoren aus dem Modell entfernt werden (siehe auch Kapitel 4.1.2). Als F–Wert für die jeweilige Faktorkombination ergibt sich:

$$\mathrm{F}_i = \left(\frac{\mathrm{SEP}_i}{\mathrm{SEP}_{min}} \right)^2 \quad . \tag{4.4}$$

Ergänzend kann zur Bestimmung der Relevanz einzelner Faktoren in der multiplen linearen Regression ein t-Test durchgeführt werden, der die Bedeutung der Regressionskoeffizienten in $\boldsymbol{\beta}$ ermittelt (vgl. Gleichung (3.38), S. 24). Der t-Test gibt im vorliegenden Fall an, ob sich die Regressionskoeffizienten der jeweiligen Faktoren signifikant voneinander unterscheiden oder nicht [64].

Die forward– und backward–Methoden können in einer *forward-backward–stepwise variable selection* zusammengefaßt werden. In diesem Fall wird das Modell mit dem signifikant niedrigsten Wert des beobachteten Kriteriums und der kleinsten Anzahl berücksichtigter Faktoren als endgültiges Kalibrationsmodell ausgewählt. In verschiedenen Literaturstellen findet man auch den Ausdruck *stepwise variable selection* (SV) als abkürzende Bezeichnung für dieses Verfahren.

In den beschriebenen Verfahren sind die Faktoren und Principal Components (PCs) nach fallenden Eigenwerten geordnet, und die Größe der Eigenwerte spielt in vielerlei Hinsicht eine wesentliche Rolle bei der Faktor–Selektion. Allerdings ist die Größe eines Eigenwerts nur ein Ausdruck für die durch den zugehörigen Faktor repräsentierte spektrale Varianz, nicht jedoch zwangsläufig auch für dessen Bedeutung im Rahmen einer quantitativen PCR–Kalibration (*s.* [45] und die darin enthaltenen Verweise). Bessere Kalibrationsergebnisse werden daher durch Anwendung des *correlation ranking* erhalten, das heißt einer Sortierung der latenten Variablen (Faktoren und PCs) nach fallender Korrelation mit der zu kalibrierenden Probeneigenschaft[2]. Im Anschluß an die Sortierung können die oben beschriebenen Methoden in analoger Weise angewandt werden. Bei Verwendung einer *top–down variable selection* wird das Verfahren dann als *correlated principal component regression* (CPCR) bezeichnet [45].

Die vorgestellten Selektionsmethoden stellen klassische, deterministische Optimierungsverfahren dar. Sie alle sind so angelegt, daß ausgehend von einem bestimmten Punkt (Lösungsansatz) im Lösungsraum[3] des Problems ausgehend geprüft wird, in

[2] Für jede zu kalibrierende Eigenschaft im Datensatz gibt es eine separate Sortierung nach den Korrelationskoeffizienten. Die Eigenwertsortierung ist dagegen unabhängig von den Eigenschaften und daher für alle gleich.

[3] Die Menge aller möglichen Lösungen.

welcher „Richtung“ sich eine Verbesserung des Beobachtungskriteriums ergibt (*hill-climbing* oder *direct–search*: Finde ein lokales Optimum und erklimme das Maximum in Richtung der größten Steigung). Neben dem Beobachtungskriterium selbst (in der Regel die Standardabweichung des jeweiligen Kalibrationsmodells) wird oft noch ein zusätzliches Kriterium benötigt, das eine Aussage darüber zuläßt, ob sich potentielle Lösungen auch in statistischer Hinsicht voneinander unterscheiden oder nicht. Eine Möglichkeit dazu bieten die im nächsten Abschnitt besprochenen Signifikanz-Tests.

4.1.2 F–Tests

In der Spektrochemometrie werden häufig Standarabweichungen zur Beurteilung der Güte von Kalibrationsmodellen herangezogen. Um eine Entscheidung treffen zu können, ob zwei Modelle sich tatsächlich und nicht nur zufällig voneinander unterscheiden, verwendet man in diesem Zusammenhang F–Tests [65–67] (eine Ausführliche Beschreibung des Testverfahrens findet sich im Anhang A.2). Dazu werden die im folgenden beschriebenen Testgrößen herangezogen.

Signifikanz der vom Modell erfaßten Varianz der Probeneigenschaft

Als Testgröße bei der Beurteilung der Probenvarianz eines Kalibrationsmodells verwendet man:

$$\mathrm{F} = \frac{\sum_{i=1}^{n_s} \left(\hat{p}_i - \bar{p}\right)^2 * (n_s - n_f - 1)}{\sum_{i=1}^{n_s} \left(\hat{p}_i - p_i\right)^2 * (n_f - 1)} \quad . \tag{4.5}$$

In der Gleichung entspricht die Zahl n_f der Anzahl im Modell berücksichtigter Faktoren, n_s der Zahl der Kalibrationsspektren und p bzw. $\hat{p}$ dem tatsächlichen bzw. dem geschätzten Wert einer Probeneigenschaft. Der resultierende F–Wert kann als Maß für das auf die Probeneigenschaft bezogene Signal–Rausch–Verhältnis innerhalb des Kalibrationsmodells angesehen werden und sollte möglichst groß sein.

Signifikanz von Faktoren

Ausgehend von der Faktorkombination, die zum Minimum der SEP–Funktion gehört, ist es möglich, diejenigen (primären) Faktoren zu bestimmen, die letztendlich in das Kalibrationsmodell Eingang finden sollen. In erster Linie kommt es bei der Optimierung der Modelle darauf an, *robuste* Kalibrationen zu erstellen, die möglichst wenige Faktoren berücksichtigen. Iterativ werden daher bei einer Backward–Stepwise Variable Selection solange Faktoren entfernt, bis eine signifikante Verschlechterung des SEP–Wertes eintritt. Um zu entscheiden, ob die Zunahme des SEP–Wertes signifikant ist oder nicht, wird eine F–Testgröße berechnet und mit tabellierten Werten

der F–Verteilung verglichen. Die Berechnung der F–Testgröße erfolgt nach

$$\mathrm{F}_{n_f}^{min} = \frac{\mathrm{SEP}_{n_f}^2 * (n_s - n_{min} - 1)}{\mathrm{SEP}_{min}^2 * (n_s - n_f - 1)} \quad . \tag{4.6}$$

n_{min} gibt dabei die Zahl der (primären) Faktoren an, die zum Minimum in der SEP–Funktion führen (SEP_{min}^2) und n_f ist die Anzahl der im Test berücksichtigten Faktoren, die zum Varianzwert $\mathrm{SEP}_{n_f}^2$ gehören. n_s gibt die Anzahl der Spektren in der Kalibration wieder.

Signifikanz zusätzlicher Faktoren

Das alternative Verfahren — die Forward–Stepwise Variable Selection — bietet die Möglichkeit, ausgehend von einem bestimmten Modell zu eruieren, ob zusätzlich hinzugenommene Faktoren den SEP–Wert noch signifikant senken. Die Berechnung der Testgröße erfolgt dann nach

$$\mathrm{F}_{n_f-1}^{n_f} = \frac{\left(\mathrm{SEP}_{n_f-1}^2 - \mathrm{SEP}_{n_f}^2\right) * (n_s - n_f - 1)}{\mathrm{SEP}_{n_f}^2} \quad . \tag{4.7}$$

Durch inkrementelles Hinzufügen von Faktoren wird so nach möglichem Verbesserungspotential im Kalibrationsmodell gesucht.

4.2 Kritische Betrachtung der klassischen Methoden

Die zuvor beschriebenen klassischen Verfahren zur Faktor–Selektion gehen überwiegend auf die Arbeiten von *E.R. Malinowski* zurück [18, 39, 47, 63] und sind in der praktischen Anwendung weit verbreitet.

Eine Grundvoraussetzung für die Anwendbarkeit der Varianztests zur automatisierten Faktor–Selektion ist das Vorhandensein unabhängiger, normalverteilter Stichproben. Zur Überprüfung, ob eine Abweichung der Varianzen in diesen Stichproben als zufällig angesehen werden kann oder signifikant ist, wird der Quotient der Varianzen berechnet. Die Stichprobenverteilung dieses Quotienten wurde von *Fischer* für den Fall berechnet, daß zwei Stichproben Grundgesamtheiten mit derselben Varianz angehören [68, 69].

B.R. Kowalski und *K. Faber* untersuchten diese Testverfahren zur Faktor–Selektion [42] und veröffentlichten 1994 die Ergebnisse. Sie kamen darin zu dem Schluß, daß vom statistischen Standpunkt aus betrachtet, die F–Tests in bezug auf diese Fragestellung nicht korrekt angewandt werden. Hauptkritikpunkt an den von *E.R. Malinowski* vorgeschlagenen Verfahren ist der Umstand, daß die zu vergleichenden Varianzen der betrachteten Modelle nicht unabhängig voneinander sind. Der Grund ist vor allem die Verwendung desselben Kalibrationsdatensatzes für die zu

vergleichenden Modelle bzw. deren Varianzen. Damit sind die Varianzen der Modelle nicht unabhängig voneinander, und die Grundvoraussetzung für das Vorliegen einer F–Verteilung ist nicht gegeben. Die Anwendung eines F–Tests läßt also vom statistischen Standpunkt aus betrachtet streng genommen keine Aussage darüber zu, ob sich zwei Faktorkombinationen in Kalibrationen signifikant unterscheiden oder nicht.

Dennoch lassen sich auf Grund der Orthogonalität der Principal Components (PCs), die zumindest innerhalb eines Kalibrationsdatensatzes die Unabhängigkeit der Variablen garantiert, mit Hilfe der beschriebenen „klassischen“ Selektionsverfahren in aller Regel brauchbare Faktorkombinationen für die Modellierung der spektralen Daten finden. Es ist jedoch zu beachten, daß die PCs nur für den Kalibrationsdatensatz, nicht aber für Validationsdatensätze eine orthogonale Basis bilden. Dies ist besonders kritisch für Modelle, in denen PCs mit sehr niedrigen Eigenwerten berücksichtigt werden. In diesem Fall tritt häufig *overfitting* auf und die jeweils als „beste“ deklarierte Faktorkombination hängt nicht zuletzt vom verwendeten (deterministischen) Selektionsverfahren ab.

Ein einfaches Beispiel soll das Verhalten von klassischen, deterministischen Such–Algorithmen verdeutlichen:

Beispiel 4.1
Aus einer Menge von dreißig Zufallszahlen sollen Kombinationen der einzelnen Werte gebildet werden, deren kumulierte Summen die Zahl 123.45 möglichst gut approximieren. Die zur Verfügung stehenden Werte sind: 9.43, 9.21, 9.12, 8.95, 8.9, 8.72, 8.23, 8.03, 7.85, 7.72, 7.5, 7.2, 7.11, 6.92, 5.9, 5.42, 4.73, 4.7, 3.71, 3.3, 2.71, 2.51, 2.3, 2.13, 2.11, 2.1, 1.87, 1.3, 1.23, 1.09 mit einer Gesamtsumme von 162. Für eine Anwendung der klassischen, deterministischen Selektionsverfahren sind die Werte nach abnehmender Größe sortiert. In Abbildung 4.1 ist dargestellt, welche Ergebnisse sich aus einer Top–Down und der Forward–Stepwise Variable Selection ergeben. In Falle der Top–Down Methode erreicht man durch schrittweise Addition der einzelnen Zahlen die beste Annäherung an den gesuchten Wert mit 120.79. Wird in einem weiteren Schritt der nächste Wert aus der Zahlenreihe hinzugenommen, dann beträgt die kumulierte Summe bereits 126.21 und stellt eine schlechtere Approximation des gesuchten Wertes 123.45 dar. Die Forward–Stepwise Variable Selection findet zunächst dieselbe Lösung von 120.79 ist aber in der Lage dieses Ergebnis durch die Addition von 2.51 auf 123.3 zu verbessern. Dagegen liefert die Backward–Stepwise Variable Selection keine so gute Approximation (s. Abb. 4.2). Durch sukzessives Weglassen der niedrigen Werte aus der Berechnung der Summe ergibt sich hier ebenfalls die Lösung mit dem Wert 120.79. Eine Alternative besteht im Rahmen der Backward–Stepwise Variable Selection darin, daß nicht mit den kleinsten, sondern den größten Werten begonnen wird. Dadurch konvergiert die kumulierte Summe schneller gegen den gesuchten Wert von 123.45. Auf diesem Weg erreicht man mit 8.9 + 8.72 + 8.23

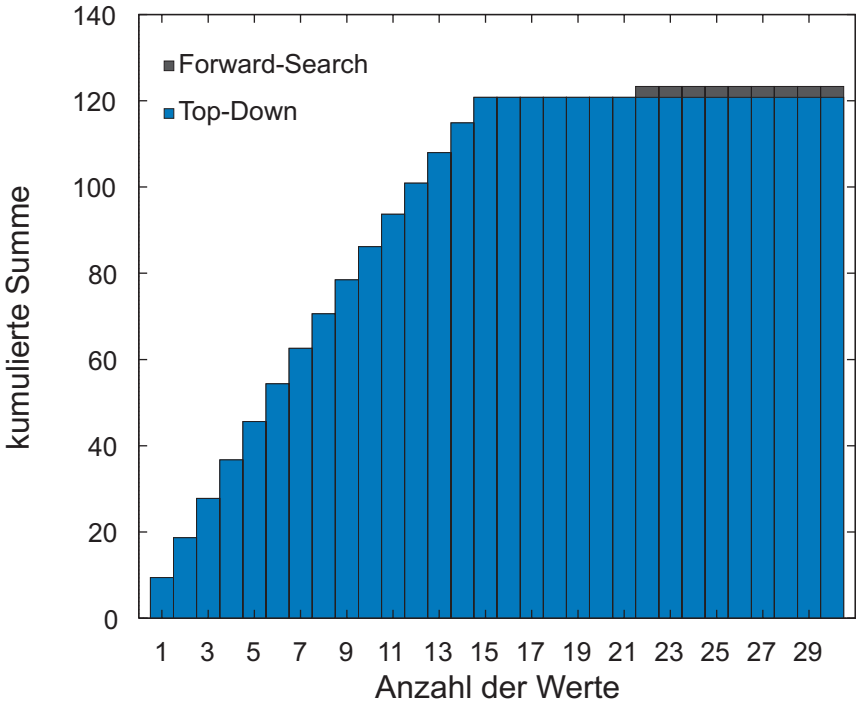

Abbildung 4.1: Lösungen des 123.45–Problems mit der Top–Down und der Forward–Stepwise Variable Selection.

+ 8.03 + 7.85 + 7.72 + 7.5 + 7.2 + 7.11 + 6.92 + 5.9 + 5.42 + 4.73 + 4.7 + 3.71 + 3.3 + 2.71 + 2.51 + 2.3 + 2.13 + 2.11 + 2.1 + 1.3 + 1.23 + 1.09 = 123.42 die beste Näherung.

An diesem Beispiel wird deutlich, wie sehr das Ergebnis einer Optimierung von der eingesetzten Such–Methode und ihrer Auslegung abhängig ist. Es ist aber auch festzuhalten, daß alle beschriebenen deterministischen Verfahren in der Lage sind, eine Lösung zu finden, die dem gesuchten Wert sehr nahe kommt.

In bezug auf die Faktor–Selektion besteht eine Alternative darin, Modelle für alle Kombinationen zu berechnen und zu vergleichen. Dieses Vorgehen ist jedoch realistisch nicht umsetzbar, da die Zahl möglicher Kombinationen von Faktoren mit 2^n zunimmt, wobei n die Anzahl von Faktoren ist, aus der ausgewählt werden kann. Für nur 10 Faktoren ergeben sich theoretisch

$$K = \sum_{k=1}^{10} \binom{10}{k} = 2^{10} = 1024 \tag{4.8}$$

Kombinationsmöglichkeiten und bei 20 Faktoren schon ca. eine Million. Der Umfang der durchzuführenden Berechnungen erschöpft daher bei einer hohen Anzahl von

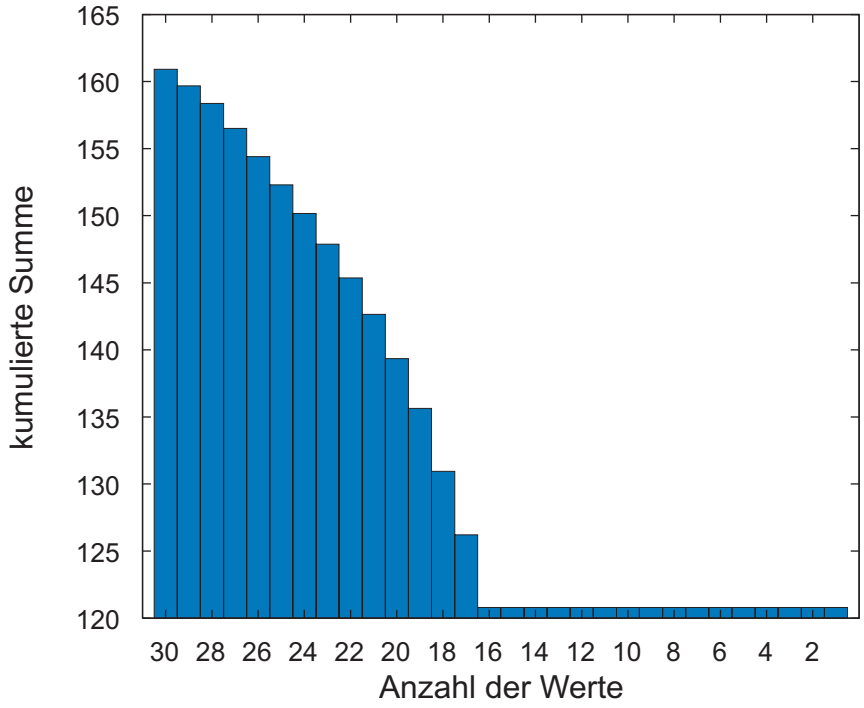

Abbildung 4.2: Lösungen des 123.45–Problems mit der Backward–Stepwise Variable Selection.

Faktoren sehr schnell die Leistungsfähigkeit moderner Rechnersysteme. Die Zahl der möglichen Kombinationen durch die Einführung von Regeln einzuschränken, scheint eine Alternative zu sein. Allerdings ist nicht sichergestellt, daß solche Regeln so allgemein aufgestellt werden können, daß sie im Einzelfall tatsächlich Gültigkeit besitzen.

Insgesamt wird deutlich, daß es sich bei der Fragestellung nach der bestmöglichen Faktorkombination für ein Kalibrationsmodell um ein schwieriges Optimierungsproblem handelt. Für seine Lösung gilt es, eine möglichst gute Näherung zum tatsächlichen, globalen Optimum zu finden. Eine Möglichkeit besteht in der Anwendung von modernen Suchalgorithmen auf Basis von *generalized simulated annealing* oder *Genetischen Algorithmen* [41, 43, 46, 70, 71].

4.3 Genetische Algorithmen als Alternative in der Faktorauswahl

Die Schwierigkeiten, die mit dem Einsatz der deterministischen Verfahren zur Faktor–Selektion in bezug auf ihre Robustheit verbunden sind, lassen sich durch Berechnung aller Faktorkombinationen beheben. Dieses Vorgehen ist zwar effektiv,

weil das Optimum sicher gefunden wird, aber es ist nicht effizient, denn die Anzahl durchzuführender Berechnungen ist impraktikabel groß. Eine Lösung dieses Dilemmas bietet der Einsatz von Genetischen Algorithmen, die nicht alle möglichen Kombinationen errechnen, sondern eine gewählte Zielfunktion systematisch unter Verwendung zufallsgesteuerter Verfahren optimieren.

Genetische Algorithmen werden schon seit einiger Zeit erfolgreich zur Lösung kombinatorischer Optimierungsprobleme in der Chemometrie eingesetzt [4, 72]. Ein Schwerpunkt ihrer Anwendung liegt in der Auswahl von Variablen bei der Erstellung multivariater Kalibrationsmodelle der quantitativen NIR–Spektrochemometrie:

- Selektion von Wellenzahlen oder Spektralbereichen in Verbindung mit der Multiplen Linearen Regression (MLR) oder Partial Least Squares Regression (PLS) [40, 71, 73–77];
- Bestimmung der optimalen Faktorkombination in der Principal Component Regression (PCR) [41, 70, 78];
- Auswahl von Fourier– [79] und Wavelet–Koeffizienten [37, 80–82].

Konkrete Anwendungsbeispiele für die Bereiche Chemie und Chemometrie finden sich in Übersichtsartikeln von *C.B. Lucasius et al.* [28, 72]. Eine gute Einführung in die allgemeine Thematik Genetischer Algorithmen stellt darüber hinaus das „*Handbook of Genetic Algorithms*“ [5] von *L. Davis* und das Buch „*Genetic Algorithms*“ [3] von *D.E. Goldberg* dar.

Der Genetische Algorithmus, der im Rahmen dieser Arbeit zur Faktor–Selektion und im Kontext eines parallel laufenden DFG–Projekts „*NIR/Wavelets*“ zur Waveletkoeffizienten-Selektion eingesetzt wurde, basiert im wesentlichen auf den von Davis vorgestellten Grundprinzipien. Er wurde auf die speziellen Anforderungen der jeweiligen Auswahlverfahren angepaßt und erweitert.

4.3.1 Aufbau und Grundfunktionen Genetischer Algorithmen: Ursprung und wichtige Begriffe

In den folgenden Abschnitten wird im einzelnen beschrieben, welche Ideen und Vorstellungen hinter Genetischen Algorithmen stecken. Die grundlegenden Funktionen werden an kurzen Beispielen erklärt. Die grundlegenden Arbeiten gehen auf *Davis* [5], *Goldberg* [3] und *Holland* [83] zurück.

Natürliche Evolution als Vorbild

Der Haupteinsatzbereich Genetischer Algorithmen liegt in der Lösung kombinatorischer Probleme, die wirksam und in einem vertretbaren Zeitrahmen gelöst werden müssen.

Als Vorbild zu Genetischen Algorithmen dienen Prozesse, welche die natürliche Evolution entscheidend beeinflussen. Der Mechanismus, der die Evolution voran treibt, ist bis heute nicht vollständig aufgeklärt, aber einige wesentliche Prozesse sind bekannt. Nach der von *C. Darwin* 1859 veröffentlichten Evolutionstheorie beruht die Entstehung und Veränderung von Arten auf der Abfolge von Reproduktion, Mutation und Selektion. In einem „Kampf ums Dasein" (*struggle for life*) überleben dabei nur die am besten an ihre Umwelt angepaßten Individuen einer Art (*survival of the fittest*).

Der Mechanismus der Evolution setzt in diesem Kampf ums Überleben an den Chromosomen an. Diese „kodieren" die Struktur jedes Lebewesens und damit die Anpassungsfähigkeit an seine Umwelt. Ein Lebewesen entsteht durch die Dekodierung seiner Chromosomen. Damit verbunden ist die Ausprägung seiner individuellen Erscheinungsform (biologisch: Phänotyp). Die wesentlichen und weitestgehend akzeptierten Punkte in der Evolutionstheorie sind folgende:

- Evolution ist ein Prozeß, der mehr an Chromosomen denn an Lebewesen stattfindet.

- Natürliche Selektion ist das Bindeglied zwischen Chromosomen und Leistungsfähigkeit (Effizienz) eines Individuums (besser: seines Phänotyps). Der Selektionsprozeß ermöglicht es Individuen mit Chromosomen, die erfolgreiche Strukturen kodieren, sich häufiger als andere zu reproduzieren (*survival of the fittest*). Durch natürliche Selektionsfaktoren kann der Erfolg einer Struktur quantifiziert werden.

- Der Prozeß der Reproduktion ist der Augenblick in dem Evolution stattfindet. Durch Rekombination und Mutation können die Chromosomen biologischer Kinder von den Chromosomen der Eltern abweichen.

- Biologische Evolution hat kein Gedächtnis. Alles was sie über erfolgreiche Individuen weiß, ist im Gen–Pool enthalten. Der Gen–Pool enthält alle Chromosomen der gegenwärtig vorhandenen Individuen.

In der Natur ist häufig der entscheidenste Selektionsfaktor die Fähigkeit eines Individuums, in seiner Umgebung Nahrung zu finden.

Das bekannteste Beispiel für die Anpassung an diese Anforderung fand *C. Darwin* auf den Galapagos-Inseln, wo er siebzehn verschiedene Finkenarten entdeckte, deren Ähnlichkeit er auf eine gemeinsame Ahnenreihe zurückführte. Die Finken, die nach der Entstehung der Inseln von Südamerika her einwanderten, spezialisierten sich unter dem Druck gegenseitiger Konkurrenz auf bestimmte Nahrungsquellen, wobei sie Schnabelformen ausbildeten, die genau auf die jeweiligen Anforderungen angepaßt sind [84]. Die Finken entstammen ursprünglich einer *heterogenen Population*, aus der sich im Laufe der Zeit einzelne, spezialisierte Gruppen bildeten. Zum Zeitpunkt

ihrer Entdeckung hatte sich eine *konvergente Population* gebildet, deren Mitglieder unterschiedliche Nahrungsquellen für sich erschlossen hatten.

Die Begriffe 'heterogen' und 'konvergent' sollen im folgenden erklärt werden. Die Güte der Anpassung wird in der natürlichen Umgebung der Finken über den Selektionsfaktor ‚Nahrungsbeschaffung' evaluiert. Der Begriff *heterogen* bezieht sich nicht auf die Verteilung des genetischen Potentials auf die ganze Population, sondern vielmehr auf die Verteilung der potentiell guten Lösungen in bezug auf das Anpassungsproblem in der Population. Die guten Lösungsansätze sind zu Beginn des Evolutionsprozesses nur bei einigen Individuen zu finden – die Population ist also in bezug auf die Verteilung möglicher, guter Lösungen des Anpassungsproblems heterogen.

Die Darwin–Finken-Population ist genau genommen nicht konvergent, sondern divergent, da sich verschiedene, auf bestimmte Nahrungsquellen spezialisierte, Gruppen gebildet haben. Diese Ausbildung ist jedoch keine Eigenschaft der Population an sich, sondern hängt von der Art des Selektionsfaktors ab, der hier mehr als eine Lösung für das Anpassungsproblem zuläßt. Jede Finken-Spezies für sich bildet dagegen eine konvergente Population, die sich an die Bedingungen ihrer jeweiligen ökologischen Nische angepaßt hat.

Hollands Idee der „künstlichen Intelligenz“

Diese Eigenschaften der natürlichen Selektion wurden seit 1960 von *J. Holland* und seinen Mitarbeitern aufgegriffen und in Form von Computer-Algorithmen abstrahiert (*s.* Tab.4.1). *J. Holland* konnte nachweisen, daß diese Algorithmen in der Lage sind, komplizierte Fragestellungen zu lösen, ohne Vorwissen über die speziellen Problemstellungen selbst zu haben. In Anlehnung an ihre natürlichen Vorbilder bezeichnete *J. Holland* diese Algorithmen als Genetische Algorithmen (GAs). Für seine Untersuchungen verwendete er Reihen von Nullen und Einsen (Bit–Strings), welche mögliche Lösungen beschreiben und als Chromosomen bezeichnet werden.

Ein Genetischer Algorithmus funktioniert so, daß aus einer zufällig zusammengesetzten Population (die sogenannte Start–Population) von Lösungen über eine geeignete Zielfunktion (Selektionsfaktor), das heißt in Abhängigkeit von der Güte der Anpassung, eine Rangfolge der Chromosomen (Individuen) in der Population bestimmt wird. Die einzelnen Chromosomen entsprechen Lösungen für das Anpassungsproblem, die durch die Zielfunktion evaluiert werden. Individuen mit überlegenen Eigenschaften können sich — ganz analog zum natürlichen Vorbild des Selektionsprozesses — häufiger reproduzieren als andere.

Es gibt eine Reihe von Möglichkeiten, natürliche Phänomene in Form eines Algorithmus zu abstrahieren. Voraussetzung ist, daß die folgenden zwei Mechanismen implementierbar sind:

Tabelle 4.1: Begriffe der Evolutions-Theorie und ihre Bedeutung im Zusammenhang mit Genetischen Algorithmen

	Evolution	Genetische Algorithmen
Population von Individuen	Chromosomen beinhalten die Erbanlagen	Bit-Strings ('Chromosomen') sind potentielle Lösungen
Informationsträger	Gen, Erbanlage (chem. Information)	'Gen', ein bestimmtes Bit im 'Chromosom'
Selektions-kriterium	Anpassung an die Umwelt (*survival of the fittest*)	Ziel- oder Fitnessfunktion
Rekombination	Fortpflanzung der „starken" Individuen	Rekombination von Lösungen mit hoher Fitness
Mutation	Zufällige Veränderungen der Erbanlagen	Zufällige Veränderungen von 'Genen'
Gendrift	Zu-/Abwanderung von Individuen	Zufälliges Generieren neuer 'Chromosomen'

1. Möglichkeit zur Kodierung des Problems (z.B. in Form von Chromosomen).

2. Definierbarkeit einer Zielfunktion, welche die Güte einer individuellen Lösung des Problems eindeutig ermittelt und quantifiziert.

Die Art der Kodierung ist von der Art des zu untersuchenden Problems abhängig und variiert daher sehr. In vielen Fällen wird eine Bit–String-Kodierung verwendet. Diese stellt für den überwiegenden Teil kombinatorischer Optimierungsprobleme, bei denen die Fragestellung „*Ist ein bestimmtes Element Teil einer Selektion oder nicht ?*" lautet, eine gute Problembeschreibung dar. Versteht man den Bit–Sting als Chromosom, so entspricht das einzelne Bit einem Gen (*s.* Tab.4.1). Jedes innerhalb des Bit–Strings „gesetzte" Gen oder Bit erhält den boolschen Wert 1. Es steht für ein bestimmtes Element, das Teil der Auswahl ist. Jedes „nicht gesetzte" Bit (boolscher Wert 0) symbolisiert ein Element, das in der Auswahl nicht berücksichtigt wird. Für Fragestellungen, die einer solchen einfachen Kodierung nicht zugänglich sind, müssen andere sinnvolle und praktikable Kodier–Schemata gesucht werden. In die Suche eines geeigneten Kodier–Schemas muß in aller Regel viel Arbeit investiert werden, damit die folgenden beiden Randbedingungen erfüllt werden: Das Hauptproblem liegt zum einen in der genauen Wiedergabe des Opimierungsproblems an sich, da eine ungenügende Problemerfassung keine wirklich guten Lösungen liefern kann, zum anderen in der von Holland formulierten Forderung, daß eine kleine Veränderung an einem Chromosom auch nur eine kleine Veränderung seiner Fitness nach sich zieht. Diese Forderung wird als Schemata–Theorem bezeichnet [3]. Die Einhaltung

des Theorems ist wichtig, da nur so eine Konvergenz im Verlauf einer GA basierten Optimierung erreicht werden kann.

Die Zielfunktion stellt die Schnittstelle zwischen dem Genetischen Algorithmus und dem zu lösenden Problem dar. Sie spielt für Genetische Algorithmen dieselbe Rolle wie die Selektionsfaktoren in der natürlichen Evolution. Es handelt sich um eine Funktion, welche die Anpassung eines Individuums an die Problemstellung eindeutig und quantitativ bewertet. Diese Anpassung gilt es zu optimieren. In der Literatur zu Genetischen Algorithmen werden Zielfunktionen daher auch als *Fitnessfunktionen* bezeichnet.

Aus den beschriebenen Mechanismen läßt sich die grundlegende Struktur eines Genetischen Algorithmus in Form eines Ablaufschemas ableiten (*s.* Abb. 4.3). Neben der Art der Kodierung und der Wahl einer geeigneten Fitnessfunktion sind die für die Generationswechsel eingesetzten *Techniken*[4] und *Operatoren* entscheidend für die Effizienz Genetischer Algorithmen. Als Techniken, die den Generationswechsel steuern, werden in dieser Arbeit die *roulette wheel parent selection* und die *steady state reproduction technique with no duplicates* eingesetzt.

Die Operatoren, die ebenfalls während der Generationswechsel Anwendung finden, wirken dagegen direkt auf die Chromosomen ein. Sie sind für die „Evolution“ im Genetischen Algorithmus verantwortlich. Als Operatoren werden *uniform crossover, mutation* und *invader* eingesetzt.

Roulette Wheel Parent Selection

Sinn der Roulette Wheel Parent Selection ist es, eine Auswahl von Chromosomen in einer Generation eines Genetischen Algorithmus zu treffen, und zwar so, daß den Chromosomen mit höherer Fitness eine größere Chance zur Reproduktion eingeräumt wird. Dabei erfolgt die Auswahl wie in Abbildung 4.4 dargestellt. Im Verlauf mehrerer Generationswechsel werden die Individuen mit höherer Fitness deutlich häufiger ausgewählt als dies für Individuen mit niedriger Fitness der Fall ist. Ein Beispiel zeigt Tabelle 4.2.

In der Tabelle ist eine Reihe von zehn Chromosomen mit ihren jeweiligen Fitness–Werten sowie den resultierenden Zwischensummen dargestellt. Die Auswahl von sieben Chromosomen auf Basis des Roulette-Wheel-Algorithmus wird im unteren Teil der Tabelle wiedergegeben. Es wird eine Zufallszahl zwischen Null und der Summe aller Fitnesswerte generiert und dann dasjenige Chromosom bestimmt, für das die Zwischensumme der Fitnesswerte dieser Zufallszahl am nächsten kommt, wobei $Q_i \geq n$ gelten soll.

Dem so ausgewählten Individuum wird dann die Möglichkeit zur Reproduktion gegeben. Auf diese Weise ist die Fitness der einzelnen Chromosomen direkt mit der

[4] Unter *'techniques'* werden im Zusammenhang mit Genetischen Algorithmen alle Mechanismen verstanden, die ausschließlich der Selektion dienen [3, 5].

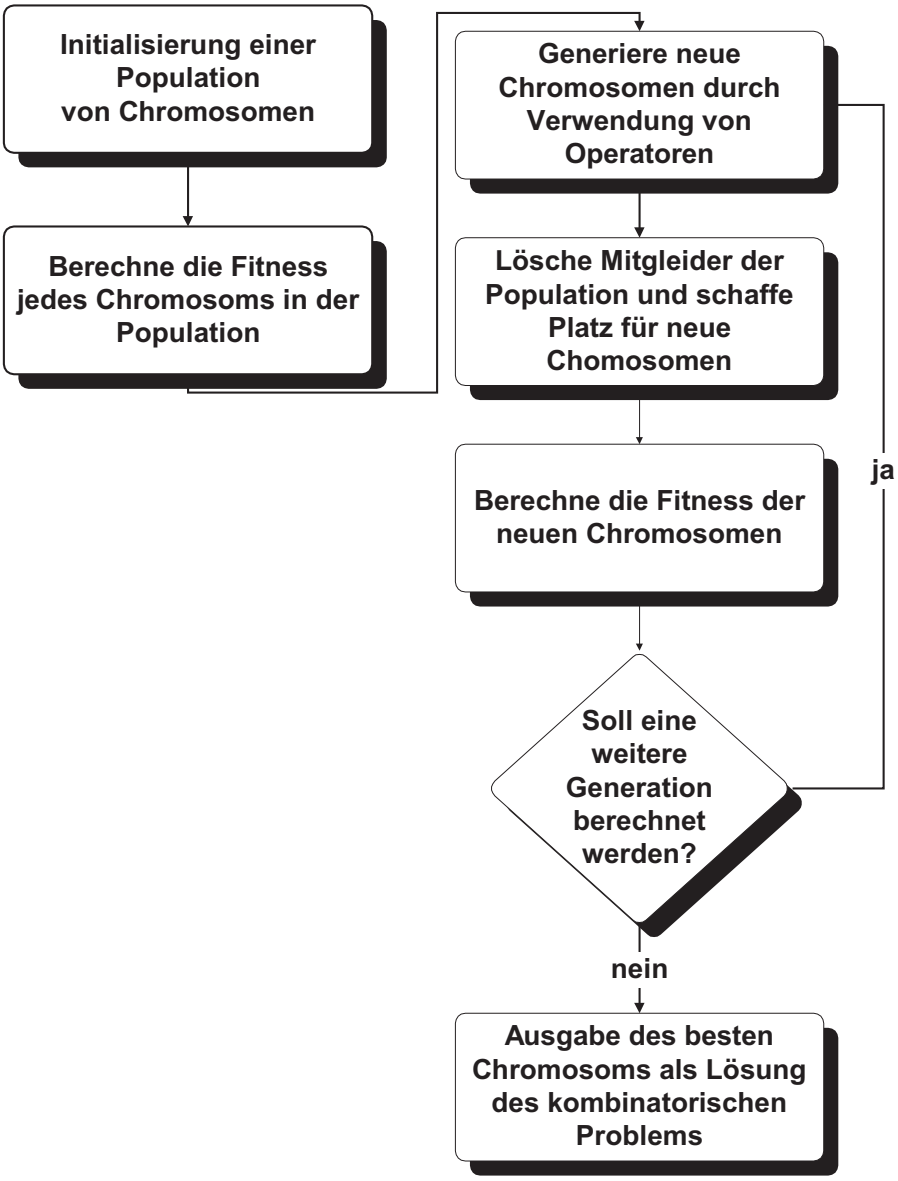

Abbildung 4.3: Ablaufschema eines Genetischen Algorithmus.

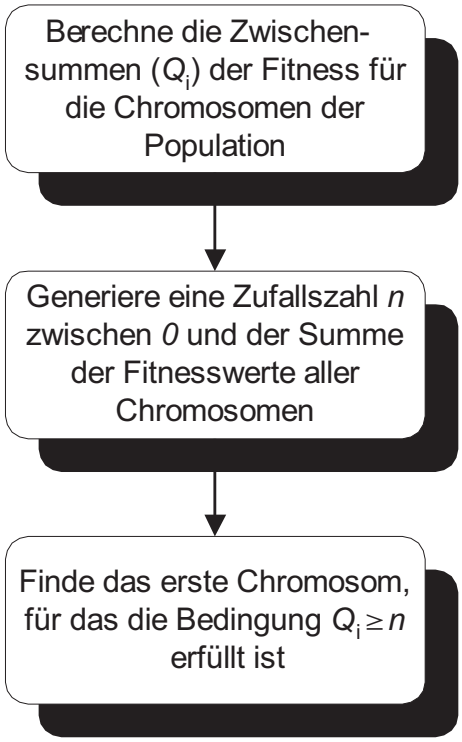

Abbildung 4.4: Der Roulette Wheel Parent Selection Algorithmus.

Auswahlwahrscheinlichkeit verknüpft. Die Wahrscheinlichkeit, daß ein bestimmtes Chromosom ausgewählt wird, steigt mit der Fitness des Chromosoms. Eine graphische Darstellung zeigt Abbildung 4.5. Bildlich kann man sich diesen Mechanismus wie ein Roulette mit unterschiedlich großen Segmenten vorstellen. Die Wahrseinlichkeit, daß eine bestimmte Zahl häufiger fällt, steigt mit der Größe ihres Segments.

Prinzipiell ist es zwar möglich, daß beim Einsatz des Roulette Wheel Parent Selection Algorithmus auch jedesmal das Mitglied der Population mit der niedrigsten Fitness (im Beispiel Chromosom 2 oder 5) ausgewählt wird, jedoch ist die Wahrscheinlichkeit hierfür in großen Populationen und über mehrere Generationen hinweg vernachlässigbar gering.

Steady State Reproduction Technique

Der Generationswechsel in einem Genetischen Algorithmus (GA) wird durch den Ersatz der Eltern–Generation durch die Folgegeneration vollzogen. Man spricht von einem GA ohne „Gedächtnis“, wenn alle Eltern durch neue Kinder ersetzt werden (*generational replacement*). In diesem Fall kommt es häufig zu unerwünschten Nebeneffekten wie dem Verlust guter Chromosomen, die nicht in irgendeiner Form reproduziert wurden. Dieses Problem kann leicht durch den Einsatz der *steady state*

Tabelle 4.2: Beispiel für eine *Roulette Wheel Parent Selection* [5]

Chromosom i	1	2	3	4	5	6	7	8	9	10
Fitness $F(i)$	8	2	17	7	2	12	11	7	3	7
Zwischensumme $Q_i = \sum_{j=1}^{i} F(i)$	8	10	27	34	36	48	59	66	69	76

Zufallszahl $n \in [0; 76]$	23	49	76	13	1	27	57
ausgew. Chromosom	3	7	10	3	1	3	7

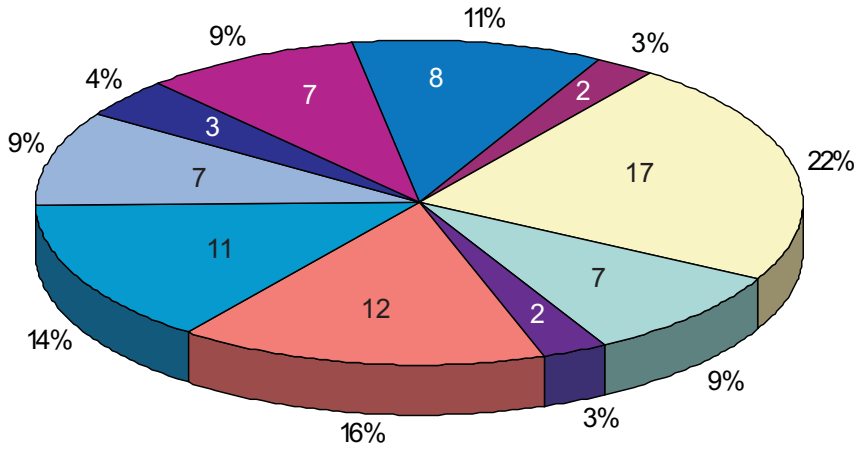

Abbildung 4.5: Beispiel einer *Roulette Wheel Parent Selection*: Die Segmente charakterisieren die einzelnen Chromosomen, wobei jeweils deren Fitness–Werte aus Tab. 4.2 angegeben sind. Die Größe der Segmente entspricht der prozentualen Wahrscheinlichkeit, mit der die einzelnen Segmente ausgewählt werden.

reproduction technique umgangen werden. Sie hat als Parameter die Zahl neu zu generierender Chromosomen, welche die Eltern ersetzen sollen. Es werden also nicht alle Chromosomen bei einem Generationswechsel ersetzt, sondern ein Teil der Eltern mit guten Fitnesswerten wird (z.B. mit der oben beschriebenen Roulette Wheel Parent Selection) in die Folgegeneration übernommen. Gute Lösungen des Optimierungsproblems bleiben so über die Generationswechsel hinweg erhalten.

Die in der vorliegenden Arbeit eingesetzte *steady state without duplicates technique* stellt eine Erweiterung dar. Diese Reproduktionstechnik verwirft neue Chromosomen, die Duplikate bestehender Individuen in der Elterngeneration sind. Der Einsatz dieser Technik führt schnell dazu, daß alle Chromosomen einer Generation voneinander verschieden sind, die Population insgesamt aber nach wie vor konvergiert. Dies ist der größte Vorteil dieser Technik gegenüber dem einfachen Generational Replacement, das eine Vielzahl von Duplikaten erzeugt und so unter Umständen

viel mehr Rechen- und Optimierungszeit benötigt, um das globale Optimum der Fitnessfunktion zu finden.

Ein Nachteil besteht darin, daß sich die Häufigkeiten von Mutation, Rekombination etc. nicht mehr ohne weiteres zuverlässig angeben lassen. Dies liegt daran, daß die Überprüfung, ob ein neues Chromosom in der Elterngeneration vorhanden ist oder nicht, erst nach dem Einsatz des jeweiligen Operators erfolgen kann, so daß die Zahl der Operatoraufrufe in der Regel nicht mit der Zahl neu in die Population eingefügter Chromosomen übereinstimmt.

Operatoren

Um in einem Genetischen Algorithmus Generationswechsel durchführen zu können, bedarf es des Einsatzes unterschiedlicher Operatoren, die neue Individuen generieren. In dem hier eingesetzten Algorithmus sind drei Typen von Operatoren implementiert:

- Uniform–Crossover,
- Mutation und
- Invader.

Uniform–Crossover

Der Uniform–Crossover–Operator steuert die Reproduktion von Chromosomen im Genetischen Algorithmus. Durch Auswahl von zwei Eltern mit Hilfe der Roulette Wheel Parent Selection Technik und der zufälligen Wahl einer Kombinationsschablone werden zwei neue Chromosomen („Kinder“) generiert. Diese der Natur nachempfundene Form der Kombination von Eigenschaften unterscheidet Genetische Algorithmen wesentlich von anderen Optimierungsverfahren (*s.* 4.1.1).

Tabelle 4.3: Ein Beispiel für die Funktion des *Uniform–Crossover*–Operators

Chromosom 1:	1	0	0	1	0	1	1
Chromosom 2:	**0**	**1**	**0**	**1**	**1**	**0**	**1**
Schablone:	1	1	0	1	0	0	1
Neues Chromosom 1:	1	0	**0**	1	**1**	**0**	1
Neues Chromosom 2:	**0**	**1**	0	**1**	0	1	**1**

Ein Beispiel für die Arbeitsweise des Operators ist in Tabelle 4.3 dargestellt. Für jede Bitposition der Kinder wird bestimmt, welcher Elternteil sein Bit an welches

Kind weitergibt. Ist der Wert in der Schablone gleich 1, dann werden die Bits der neuen Chromosomen entsprechend den der Ausgangschromosomen besetzt. Ist der Schablonenwert 0, werden die Bits vertauscht.

Der Vorteil dieses Operators gegenüber einer Vertauschung der Bits an nur einer Stelle des Chromosoms (*one point crossover*) liegt darin, daß die Position, an der eine gute Eigenschaft innerhalb eines Bitstrings kodiert ist, unerheblich ist. Andererseits birgt dies aber auch die Gefahr, daß durch die gewählte Schablone positive Eigenschaften eines Chromosoms nicht in einem Stück an die Kinder weitergegeben werden. In Kombination mit der Steady State Without Duplicats Technique, die für einen Erhalt guter Chromosomen sorgt, sowie der angepaßten Crossover-Rate im Verlauf der Generationswechsel des Genetischen Algorithmus tritt dieser Nachteil jedoch in den Hintergrund.

Mutation

Der *mutation*–Operator ändert einzelne Bits eines Strings durch Ersatz eines zufällig neugewählten Bits. Durch Mutation werden die ausgewählten Chromosomen in der Regel deutlich weniger stark verändert als durch den Crossover–Operator, und gute Lösungen bleiben eher erhalten. In Tabelle 4.4 sind drei Beispiele für die Wirkungs-

Tabelle 4.4: Funktionsweise des *Mutations*–Operators in einem Genetischen Algorithmus. Eine Mutation erfolgt nur dann, wenn die Zufallszahl $0 \geq z \geq 1$ an einem bestimmten Gen kleiner als P ist

$P = 0.8\,\%$

Chromosom				Zufallszahl				Bit	Neues Chromosom			
1	0	1	0	.801	.102	.266	.373	-	1	0	1	0
1	1	0	0	.120	.096	**.005**	.840	0	1	1	0	0
0	0	1	0	.760	.473	.894	**.001**	1	0	0	1	**1**

$P_{Ende} = 75.0\,\%$

Chromosom				Zufallszahl				Bit	Neues Chromosom			
1	0	1	0	.801	**.102**	**.266**	**.373**	0	1	0	**0**	0
1	1	0	0	**.120**	**.096**	**.005**	.840	0	**0**	**0**	0	0
0	0	1	0	.760	**.473**	.894	**.001**	1	0	**1**	1	**1**

weise des Mutations-Operators an je drei Bitstrings aufgezeigt. Bei einer Mutationswahrscheinlichkeit von 0.8 % bedeutet dies, daß von 1000 Bits gerade acht dem Mutations-Operator unterworfen werden. Im ersten Teil der Tabelle ist ein mögliches Ergebnis dieser Operation dargestellt.

Im ersten Bitstring wird der Wahrscheinlichkeitstest von keinem der Bits erfüllt, so daß kein Bit mutiert wird. Der zweite String wird am dritten Bit dem Mutations-Operator unterworfen. Das für den Ersatz dieses Bits zufällig ausgewählte neue Bit ist Null, was dazu führt, daß der ursprüngliche String erhalten bleibt. Es hat effektiv keine Mutation stattgefunden. Erst im dritten String findet eine Mutation am vierten Bit statt, die zu einem neuen String führt.

Mit der Änderung der Mutationswahrscheinlichkeit ändert sich die Wirksamkeit der Mutation. Eine 75.0 %ige Mutationswahrscheinlichkeit, wie sie gegen Ende eines GA-Durchlaufs[5] vorliegt, würde in diesem Beispiel dazu führen, daß jeder String an durchschnittlich einer Stelle mutiert wird.

Invader

Der *invader*–Operator generiert neue Chromosomen durch wahlloses Setzen von Bits. Seine Aufgabe besteht darin, einen unbestimmten Grad an Varianz in den Generationen aufrechtzuerhalten wenn die Populationen bereits konvergent sind.

Die Operatoren werden mit unterschiedlichen Wahrscheinlichkeiten aufgerufen. Die Tabelle 4.5 zeigt die Aufrufwahrscheinlichkeiten der einzelnen Operatoren für

Tabelle 4.5: Wahrscheinlichkeitswerte der Operatoren

Operator	P_{Start}	P_{Ende}
Uniform–Crossover	89.1 %	9.8 %
Mutation	9.9 %	88.2 %
Invader	1.0 %	2.0 %

die erste (P_{Start}) und die letzte (P_{Ende}) Generation einer kompletten GA Optimierung. Die Wahrscheinlichkeiten unterliegen im Verlauf der Optimierung einer von der Anzahl berechneter Generationen abhängigen linearen Veränderung (*s.* Abb.4.6). Der Grund, diese Form der Anpassung zu wählen, liegt in der besseren Konvergenz der Population. Zu Beginn sind die Generationen sehr heterogen, so daß weder der Mutations- noch der Invader–Operator Aussicht hat, eine entscheidende Änderung des Chromosomensatzes zu erzielen. Je konvergenter eine Population jedoch ist, desto ähnlicher werden sich deren Individuen und die Crossover–Technik bewirkt kaum noch Veränderungen. Es liegt daher nahe, die Wahrscheinlichkeitswerte der Operatoren in geeigneter Weise anzupassen.

[5] Bei den im Rahmen dieser Arbeit durchgeführten GA-Berechnungen wird mit einer Mutationswahrscheinlichkeit von 10 % begonnen, die dann bis zum Ende linear auf 88 % gesteigert wird.

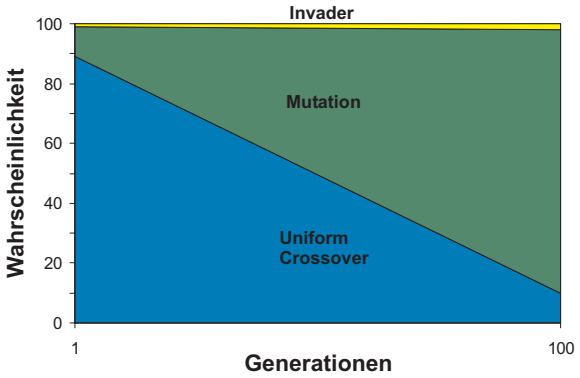

Abbildung 4.6: Veränderung der Aufrufwahrscheinlichkeiten einzelner Operatoren im GA über 100 Generationen.

4.3.2 Validation Genetischer Algorithmen

Wesentliche Größen in bezug auf die Beurteilung der Leistungsfähigkeit Genetischer Algorithmen sind deren Konvergenz und Robustheit. Sowohl die Bildung der Startpopulation als auch die Anwendung der einzelnen Operatoren beim Ablauf eines Genetischen Algorithmus werden durch Zufallsfunktionen gesteuert. Es ist daher nicht von vornherein garantiert, daß diese Systeme sich unter gleichen Startbedingungen auch gleich entwickeln, das heißt es ist nahezu ausgeschlossen, daß ein mehrfach aufgerufener Genetischer Algorithmus für die Lösung ein und desselben Problems identische Lösungswege findet [5]. Für den praktischen Einsatz muß sichergestellt sein, daß der entsprechende Algorithmus zuverlässig in Richtung des globalen Optimums seiner Fitnessfunktion tendiert.

Konvergenz Genetischer Algorithmen

Ein wichtiger Parameter für die Beurteilung der Leistungsfähigkeit eines GA ist seine Konvergenz. Je schneller der Algorithmus auf das gesuchte Optimum zukonvergiert, desto kleiner kann die Anzahl der zu berechnenden Generationen gehalten werden. Bei gleicher Populationsgröße verkürzt sich dadurch die notwendige Rechenzeit erheblich und nur so wird aus einem Genetischen Algorithmus ein wirklich schnelles Hilfsmittel zur automatisierten Selektion von Faktoren.

In der Anfangsphase der Optimierung ist der Uniform–Crossover–Operator derjenige, der sich am deutlichsten auf die Konvergenz auswirkt. Für die Beurteilung der Leistungsfähigkeit des Genetischen Algorithmus ist es daher wichtig abzuschätzen, welche Crossover Rate in Verbindung mit der Steady State Reproduction Technique zu einer hohen Konvergenz des Systems führt, bzw. ob es Crossover Raten mit besonders ungünstigen Auswirkungen auf das Konvergenzverhalten gibt.

Robustheit Genetischer Algorithmen

Wie in den vorangegangenen Abschnitten bereits erwähnt, ist neben der Konvergenz des Genetischen Algorithmus die Zuverlässigkeit mit der er eine Fitnessfunktion optimiert ein wichtiger Parameter. Je seltener ein GA in einem lokalen Optimum *„hängenbleibt“*, desto robuster ist er. Gerade im Hinblick auf den Einsatz in automatisierten Verfahren fällt der Robustheit großes Gewicht zu.
Ausführliche Untersuchungen zu diesen beiden Themenkreisen finden sich in Kapitel 7. Der im Kontext dieses Projektes entwickelte Genetische Algorithmus wird anhand verschiedener Datensätze validiert. Dabei wurden eine Reihe von Parametern systematisch geändert und ihr Einfluß auf die Konvergenz der Populationen und die Robustheit des Genetischen Algorithmus im Zusammenhang mit der Faktor–Selektion untersucht.

4.4 Alternative Kriterien zur Bewertung der Güte von Kalibrationsmodellen

Die Notwendigkeit die Güte zweier Kalibrationen vergleichen und bewerten zu können, ist in den vorhergehenden Abschnitten bereits angesprochen worden (*s.* 4). Eine häufig praktizierte Methode ist der Vergleich von Standardabweichungen auf Basis von F–Tests (*s.* 4.1.2). Eine der Grundannahmen, die bei der Berechnung von F–Tests erfüllt sein muß, ist die zufällige Zusammensetzung der Stichproben, auf deren Basis die zu vergleichenden Standardabweichungen bestimmt wurden [65–67]. Bei der Anwendung auf Faktor–Selektionen oder dem Vergleich von Daten-Pretreatments ist diese Voraussetzung streng genommen nicht gewährleistet. Zähler und Nenner im F–Test sind nicht unabhängig voneinander und unterliegen daher keiner F-Verteilung. Man kann infolgedessen nicht mehr von einem F–Test im üblichen Sinne sprechen. Es ist jedoch falsch, kategorisch zu behaupten, daß durch diese Vorgehensweise ausnahmslos falsche Entscheidungen herbeigeführt werden [42]. Vielmehr ist es ein zu berücksichtigender Aspekt, der zur Suche nach geeigneteren Alternativen anhalten sollte.

Zwei Möglichkeiten sollen in den folgenden Abschnitten 4.4.1 und 4.4.2 vorgestellt werden. Zum einen wird eine einfache und schnelle Methode beschrieben, wie die Varianz der Standardabweichung einer Stichprobe bestimmt und wie sie zur Beurteilung von Kalibrationen eingesetzt werden kann. Zum anderen wird eine Berechnungsmethode zur Bestimmung der Empfindlichkeit einer multivariaten Kalibration gezeigt [85].

4.4.1 Varianz von Standardabweichungen

Allgemein bekannt ist, daß sich für eine ausreichend große Stichprobe vom Umfang n einer Grundgesamtheit und deren Eigenschaft p und Mittelwert $\bar{p}$ die Werte für Varianz und Standardabweichung wie folgt schätzen lassen:

$$\mathrm{Var}(p) = \frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2 \quad , \tag{4.9}$$

$$s(p) = \sqrt{\mathrm{Var}(p)} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}} \text{ mit } X_i = p_i - \bar{p} \quad . \tag{4.10}$$

Es läßt sich zeigen, daß für $n \to \infty$ $s(p)$ gegen die tatsächliche Standardabweichung $\sigma(p)$ der Grundgesamtheit konvergiert [68]. Mit zunehmendem n gibt $s(p)$ die Standardabweichung der Grundgesamtheit $\sigma(p)$ genauer wieder. Wir nehmen an, X_i sei normalverteilt, besitze die Varianz 1 und habe den Erwartungswert 0. Für die relative Varianz der Bestimmung von $\mathrm{Var}(p)$ läßt sich nun folgende Überlegung anstellen: Setzt man mit den normierten X_i

$$Y := \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \quad , \tag{4.11}$$

dann läßt sich zeigen, daß

$$\mathrm{Var}(Y) \approx \frac{1}{2n} \tag{4.12}$$

und damit

$$S(Y) \approx \sqrt{\mathrm{Var}(Y)} = (\sqrt{2n})^{-1} \tag{4.13}$$

ist.

Definiert man Z nach folgender Formel

$$Z = \sum_{i=1}^{n} X_i^2 \quad , \tag{4.14}$$

so ist Z χ^2 verteilt mit n Freiheitsgraden. Die relative Varianz in der Bestimmung der Werte von Z läßt sich anhand der Dichtefunktion[6] bestimmen. Sie ist allgemein für χ_n^2–verteilte Funktionen [67, 68, 86] gegeben durch:

$$f_n(a) = \frac{1}{2^{\frac{n}{2}},\ \left(\frac{n}{2}\right)} a^{\frac{n}{2}-1} \cdot e^{-\frac{a}{2}} \quad , \quad a > 0. \tag{4.15}$$

[6] Die Funktion , (a) ist eine Verallgemeinerung der Fakultäten auf positive reelle Zahlen und wird als *Gammafunktion* bezeichnet.

Gesucht ist aber nicht die Dichtefunktion von Z, sondern die der Zufallsvariablen $g(Z)$ mit $g(Z) := \sqrt{Z} = \sqrt{\sum_{i=1}^{n} X_i^2}$. Diese läßt sich mit Hilfe des Tranformations–Theorems für Lebesque–Dichten ermitteln. Hier gilt:

$$h(y) = \frac{f\left(g^{-1}(y)\right)}{\left|g'(g^{-1}(y))\right|} \quad . \tag{4.16}$$

Durch Substitution von $g^{-1}(y) = y^2$ und $g'(z) = \frac{1}{2\sqrt{z}}$ und Einsetzen in Gleichung (4.16) erhält man

$$h(y) = f_n(y^2) \cdot 2y = \frac{2}{2^{\frac{n}{2}},\ \left(\frac{n}{2}\right)} y^{n-1} e^{-\frac{y^2}{2}}, y > 0 \quad . \tag{4.17}$$

Aus den Gleichungen (4.18) und (4.19) können dann der Erwartungswert und die relative Varianz von $g(Z)$ bestimmt werden

$$\mu_n = \int_0^\infty y h(y)\, dy = \sqrt{2} \cdot \frac{,\ \left(\frac{n+1}{2}\right)}{,\ \left(\frac{n}{2}\right)} \quad , \tag{4.18}$$

$$\mathrm{Var}(\sqrt{Z}) = \int_0^\infty (y - \mu_n)^2 h(y)\, dy = n - 2 \cdot \left(\frac{,\ \left(\frac{n+1}{2}\right)}{,\ \left(\frac{n}{2}\right)} \right)^2 \quad . \tag{4.19}$$

Es soll aber nicht die Varianz von $g(Z)$, sondern von Y mit $Y = \frac{1}{\sqrt{n}} g(Z) = \frac{1}{\sqrt{n}} \sqrt{Z}$ bestimmt werden. Deshalb der letzte Schritt:

$$\mathrm{Var}(Y) = \left(\frac{1}{\sqrt{n}}\right)^2 \cdot \left(n - 2 \cdot \left(\frac{,\ \left(\frac{n+1}{2}\right)}{,\ \left(\frac{n}{2}\right)} \right)^2 \right) = \frac{1}{n} \cdot \left(n - \mu_n^2\right). \tag{4.20}$$

Was gibt nun die Differenz aus n und dem Quadrat des Erwartungswerts $(n - \mu_n^2)$ wieder? Setzt man für n Werte zwischen 1 und 200 ein, so ergibt sich der in Abbildung 4.7 dargestellte Plot. Wie sich aus der Abbildung 4.7 schon abzeichnet, gilt für die Differenz $\lim_{n\to\infty}(n - \mu_n^2) = 0,5$, und Gleichung (4.20) kann daher gut durch

$$\mathrm{Var}(Y) \approx \frac{1}{2n} \ \text{ und } \ \mathrm{S}(Y) \approx \frac{1}{\sqrt{2n}} \tag{4.21}$$

angenähert werden. Dies bestätigt die in Gleichung (4.12) und (4.13) angeführte Annahme. Die beiden Werte geben die relative Varianz der Varianz ($\mathrm{Var}(Y)$) und relative Varianz der Standardabweichung ($\mathrm{S}(Y)$) wieder.

Was läßt sich nun mit dieser Gleichung anfangen? Mit Gleichung (4.21) hat man ein einfaches Mittel an der Hand, das eine Aussage darüber zuläßt, ob eine Änderung der Standardabweichung über ihre Varianz hinaus vorliegt oder nicht. Man

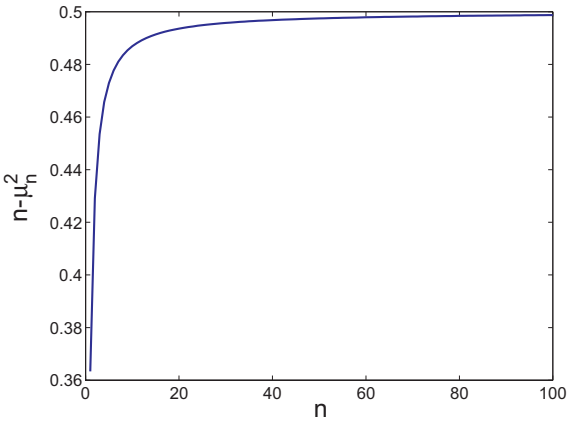

Abbildung 4.7: Plot von $(n - \mu_n^2)$ für $n = 1$ bis 200.

kann so auf Basis der Stichprobengröße ein geeignetes Kriterium festlegen, das eine qualifizierte Differenzierung von Kalibrationen zuläßt.

Durch Berücksichtigung der Varianzen lassen sich Intervallgrenzen um die Werte der Standardabweichungen definieren, so daß zwei Kalibrationsmodelle dann als gleichwertig definiert werden, wenn sich ihre Intervallgrenzen überschneiden. Diese Definition eines Vergleichskriteriums soll im folgenden als *Varianz–Kriterium* bezeichnet werden. Formal läßt es sich wie folgt formulieren:

Angenommen S_1 und S_2 sind Standardabweichungen zweier unterschiedlicher Kalibrationsmodelle und d_1, d_2 deren korrespondierende relative Varianzen aus dem Varianz–Kriterium. Um festzustellen, ob sich S_1 und S_2 tatsächlich unterscheiden, dürfen sich die beiden Intervalle

$$[S_i(1 - d_i), S_i(1 + d_i)], \quad i = 1, 2$$

nicht überlappen. Für die relativen Varianzen gilt die asymptotische Näherung aus Gleichung (4.21):

$$d_i \approx \frac{1}{\sqrt{2n}} \quad . \tag{4.22}$$

Hier steht n für die Anzahl der Freiheitsgrade, die bei der Berechnung der Standardabweichungen in den Nenner eingehen.

Beispiel 4.2
Es sollen drei Kalibrationen (auf Basis unterschiedlicher Faktor–Selektionen) mit jeweils 52 Kalibrationsstandards verglichen werden. Die zugehörigen Standardabweichungen ($s(x)$) sind in Tabelle 4.6 zusammengestellt und zeigen, daß sich diese nur leicht voneinander unterscheiden. Welche Selektionen sind gleichwertig und

Tabelle 4.6: Standardabweichungen in mol L^{-1} verschiedener $AlCl_3$ Kalibrationsmodelle mit und ohne Datenvorbehandlung

Kalibration	$s(x)$	$s(x)_u$	$s(x)_o$	Faktorkombination
A	2.66E-03	2.40E-03	2.92E-03	1–16,18
B	2.67E-03	2.40E-03	2.93E-03	1–6,8–18,24
C	2.36E-03	2.13E-03	2.59E-03	1–9,12–15,17,21–23,25

welche verschieden? Für die obere und untere Grenze der Standardabweichungen ($s(x)_u$ und $s(x)_o$) ergibt sich gemäß Gleichung (4.21) eine relative Varianz von $(\sqrt{2 \cdot 52})^{-1}$ =9.81E-02, was einer Varianz der Standardabweichung von ca. 10% entspricht. Die Kalibrationen A und B weisen zwar einen Unterschied in der Standardabweichung auf, sind aber auf Grund des oben definierten Varianz–Kriteriums nicht signifikant unterschiedlich[7]*. Nicht so die Kalibration C. Hier ist die Standardabweichung kleiner und sie liegt auch unter der unteren Grenze beider Vergleichsmodelle. Kalibration C läßt sich auf Basis des Varianz–Kriteriums von den Kalibrationen A und B differenzieren.*

Das betrachtete Beispiel bezieht sich auf die Kalibration von Aluminiumchlorid in einer wässrigen Lösung von Kalium- und Aluminiumchlorid, die einer Faktor–Selektion auf Basis von Genetischen Algorithmen unterworfen wurde. Bei den ersten beiden Kalibrationen wurden die Spektren keinem Pretreatment unterzogen. Bei der dritten Kalibration wurde eine Spektrenvorbehandlung mit Meancentering und Offset-Korrektur durchgeführt, was, wie gezeigt, zu einer signifikanten Verbesserung führt.

4.4.2 Net Analyte Signal: ein Qualitätswert für die Kalibrationsbewertung

In der Analytischen Chemie werden zur Bewertung von Kalibrationen neben der Varianz und der Sandardabweichung weitere Maßzahlen herangezogen, wie zum Beispiel die Empfindlichkeit oder das Signal zu Rauschverhältnis. Diese Begriffe entstammen den klassischen, univariaten Kalibrationsmethoden in der Analytik (*s.* 3.1), lassen sich jedoch, wie im folgenden gezeigt werden soll, auch auf multivariate Verfahren übertragen.

Für multivariate Verfahren gilt unter der Voraussetzung linearer Abhängigkeit zwischen Analytkonzentration und Signal

$$\mathbf{p} = \mathbf{bX} + \mathbf{e} \quad . \tag{4.23}$$

[7]Der Begriff „signifikant“ wird im Zusammenhang mit dem Varianz–Kriterium nicht in der im Rahmen von F- und t–Test üblichen Definition angewandt.

Hierbei ist $\mathbf{p} \in \mathbb{R}^{1\times k}$ der Vektor der tatsächlichen Analytkonzentration[8] in den Standards, $\mathbf{X} \in \mathbb{R}^{l\times k}$ ist die Matrix der unverfälschten spektralen Information, $\mathbf{b} \in \mathbb{R}^{1\times l}$ ist der unbekannte Regressionsvektor, $\mathbf{e} \in \mathbb{R}^{1\times k}$ ist der Fehler, k die Anzahl der Kalibrationsstandards und l die Anzahl der Variablen (z.B. Extinktionen an bestimmten Wellenlängen). Ein Vergleich von Gleichung (3.5) und Gleichung (4.23) zeigt, daß sich inverse univariate und inverse multivariate Kalibrationsmethoden wie PCR und PLS nur in ihrer Dimensionalität, nicht aber in der Art des zugrundliegenden linearen Modells unterscheiden. Dieser Umstand ermöglicht es, eine multivariate Kalibration als *pseudo-univariate* Kalibration darzustellen und die Fehlerdiskussion auf dieser Basis zu führen. Hierbei ist es wichtig zu berücksichtigen, daß nicht nur die Analytkonzentration, sondern auch das Meßsignal fehlerbehaftet ist. Üblicherweise werden Regressionsvektoren auf Basis der mit einem referenzanalytischen Verfahren ermittelten Eigenschaften, $\tilde{\mathbf{p}} = \mathbf{p} + \Delta\mathbf{p}$, und der mit einem Spektrometer gemessenen spektralen Daten, $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{G} + \Delta\mathbf{X}$ der Standards, bestimmt. Hier steht '~' für gemessene Werte, Δ für den Meßfehler der jeweiligen Größe und $\mathbf{G}$ für spektrales Rauschen. Betrachtet man ein Regressionsproblem in dem die Zahl zu berücksichtigender Wellenzahlen A ist, dann ist der Regressionsvektor für diese A-dimensionale multivariate Regression gegeben durch:

$$\hat{\mathbf{b}}_A = \tilde{\mathbf{p}}\tilde{\mathbf{X}}^T \left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\right)^+_A \quad . \tag{4.24}$$

Analog zu dem in 3.2.1 beschriebenen Verfahren kann der Least Squares Schritt zur Berechnung des Regressionsvektors statt im Wellenlängen- auch im Faktorraum durchgeführt werden, wobei Gleichung (4.24) ersetzt wird durch

$$\hat{\mathbf{b}}_r = \tilde{\mathbf{p}}\mathbf{C}^T (\mathbf{R})^+_r \quad \text{mit} \quad \mathbf{R} = \tilde{\mathbf{X}}\mathbf{C}^T \quad . \tag{4.25}$$

$\mathbf{C}$ enthält spaltenweise die aus der Eigenwertzerlegung ermittelten Eigenvektoren. Die Spalten von $\mathbf{R}$ entsprechen den Faktoren (Eigenspektren). Wie bereits in 3.3 beschrieben, wird der abschließende Regressionsschritt nur mit einer eingeschränkten Zahl selektierter Faktoren r durchgeführt. $\hat{\mathbf{b}}_r^T$ stellt dann die Gewichtung der Wellenlängen bezüglich der kalibrierten Property auf Basis dieser reduzierten Anzahl von Faktoren dar und wird als *Property–Weighting–Spektrum* bezeichnet.

Die Analogie zu univariaten Kalibrationen läßt sich herstellen, wenn man entsprechend zu Gleichung (4.23) folgenden Zusammenhang formuliert:

$$p_u = b\, r_u^* + e_u \;\text{ mit }\; b = \|\hat{\mathbf{b}}_r\| \; . \tag{4.26}$$

Darin ist $\|\hat{\mathbf{b}}_r\|$ die Euklidische Norm des Regressionsvektors, e_u der Fehler und r_u^* das sogenannte skalare Net Analyte Signal (NAS). Das NAS ist also die skalare

[8] 'tatsächliche' Analytkonzentration steht hier im Gegensatz zu einer fehlerbehafteten Größe (z.B. der Fehler der Referenzmethode). Dadurch wird die Berücksichtigung eines zusätzlichen Fehlerterms in den folgenden Ableitungen notwendig.

Entsprechung eines Spektrums und errechnet sich durch

$$r_u^* = \frac{1}{b} p_u \tag{4.27}$$

$$= \gamma p_u \quad , \tag{4.28}$$

wenn e_u vernachlässigbar ist. Der Wert von $\gamma = \frac{1}{b}$ kann dann entsprechend der bei univariaten Kalibrationen verwendeten Terminologie als Empfindlichkeit und b entsprechend als 'inverse' Empfindlichkeit bezeichnet werden. Eine graphische Darstellung des Zusammenhangs von NAS und Property zeigt Abbildung 4.8. Aus ihr

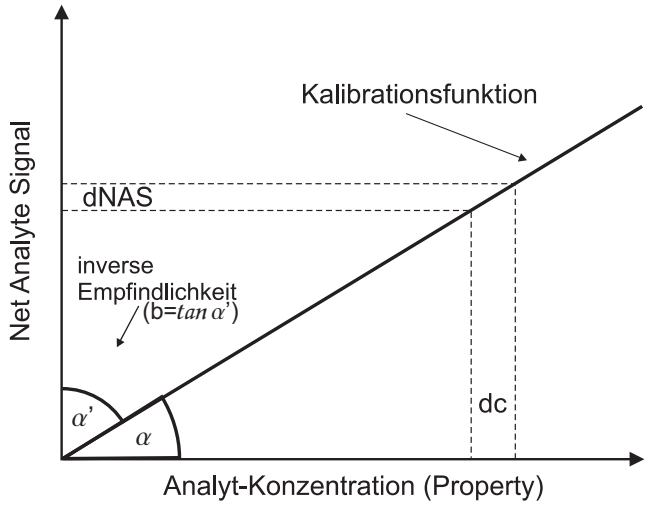

Abbildung 4.8: Darstellung des Zusammenhangs zwischen skalarem NAS und Property. Aus der Unsicherheit (dNAS) des Net Analyte Signals (NAS) resultiert nach Projektion auf die Abszisse die Unsicherheit der Property (dc).

wird ersichtlich, daß gilt:

$$\tan \alpha = \frac{1}{b} = \gamma \tag{4.29}$$

und

$$\tan \alpha' = \cot \alpha = b \quad . \tag{4.30}$$

Abbildungen wie diese ermöglichen eine Visualisierung des Zusammenhangs zwischen spektralem Fehler (Meßfehler) und Property–Fehler. Die bei multivariaten

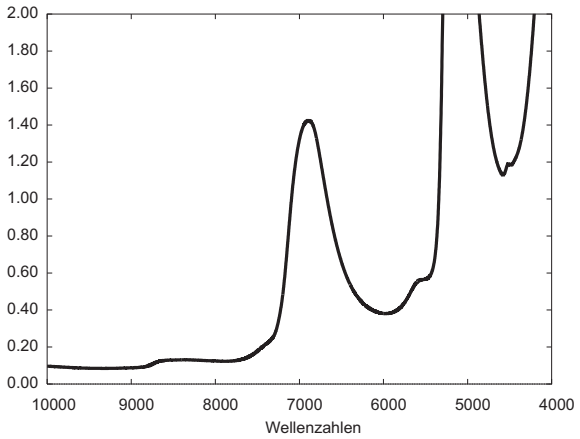

Abbildung 4.9: Mittelwertspektrum der Kalibrationsspektren.

Verfahren übliche Gegenüberstellung von 'wahrer' und vorhergesagter Property erlaubt dies nicht [85, 87–92].

Am Beispiel eines Zweikomponentensystems von Kalium- und Aluminiumchlorid in Wasser soll dieser Zusammenhang kurz erläutert werden. Kaliumchlorid ist hierbei etwas schlechter zu detektieren, da die durch die Kalium-Ionen hervorgerufenen spektralen Änderungen des wässrigen Systems im Vergleich zu denen der Aluminium-Ionen klein und gegenläufig zu diesen sind [61, 93]. Es sollen die Auswirkungen eines Daten-Pretreatments auf die Kalibration und den resultierenden Regressionsvektor gezeigt werden. Ein Vergleich des Mittelwertspektrums (*s.* Abb. 4.9) mit den Property–Weighting–Spektren in Abb. 4.10 zeigt deutlich, daß zur Kalibration die Flanken der Wasserbanden bei 7000 cm^{-1} und 5000 cm^{-1} in der Kalibration am höchsten gewichtet werden. Berechnet man nach Gleichung (4.26) die zu den beiden Kalibrationen gehörenden inversen Empfindlichkeiten, so ergeben sich die in Tabelle 4.7 aufgelisteten Werte. Hieraus ergibt sich wie in Abbildung 4.11 ver-

Tabelle 4.7: Inverse Empfindlichkeit (b) und Empfindlichkeit (γ) der KCl-Kalibration mit und ohne Datenvorbehandlung

Kalibration	b	γ
Originaldaten	7.477	0.134
vorbehandelte Daten	3.487	0.287

deutlicht, daß sich die Vorbehandlung der Spektren positiv auf die Empfindlichkeit auswirkt und zu mehr als einer Verdoppelung ihres Wertes führt. Ferner kann aus der univariaten Darstellung gemäß Abbildung 4.8 auch abgeleitet werden, wie groß die zu erwartende Ungenauigkeit der Konzentrationsbestimmung (dc) für bestimmte

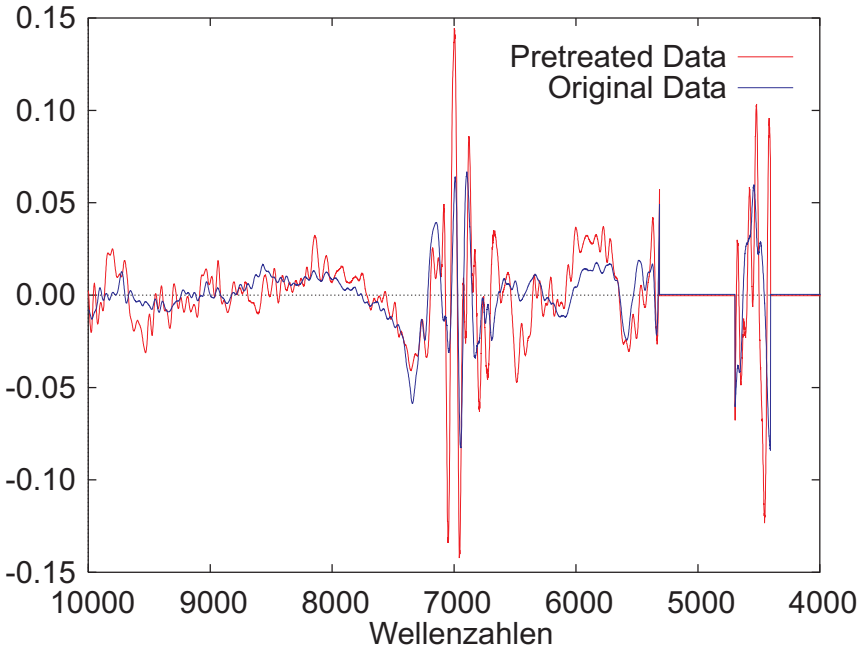

Abbildung 4.10: Property–Weighting–Spektren der KCl-Kalibrationen. Unbehandelte Daten und Daten nach Anwendung von Meancentering und Offset-Korrektur (Pretreated Data).

NAS–Fehler (dNAS) ist.

Auch die weniger stark strukturierte Form des Property–Weighting–Spektrums im Fall der vorbehandelten Daten in Abbildung 4.10 bestätigt die positive Auswirkung des Pretreatments, denn es ist sehr häufig festzustellen, daß Property–Weighting–Spektren mit scharfen und intensiven Amplituden ein Anzeichen für störanfällige Kalibrationen sind.

Mit dem Konzept des Net Analyte Signal wird eine quantitative Bewertung der 'Kalibrationsunsicherheit' möglich. Darüber hinaus bringt es den bei univariaten Kalibrationsmethoden etablierten Begriff der Empfindlichkeit auch für multivariate Techniken wieder in die Diskussion.

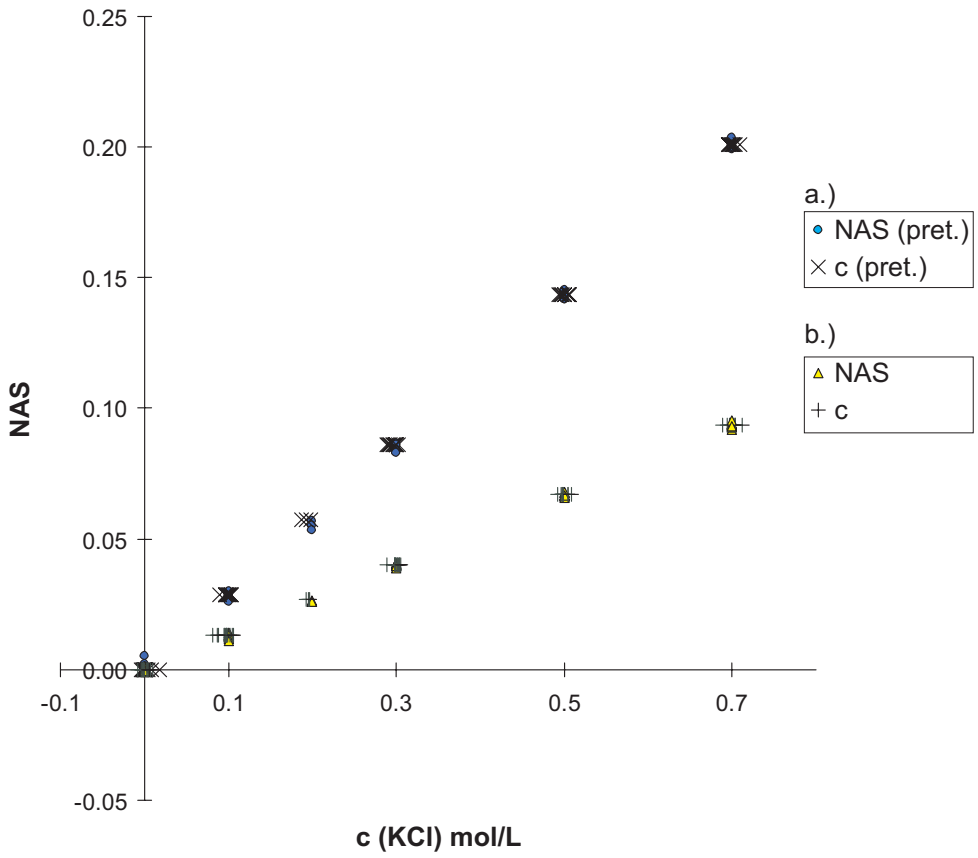

Abbildung 4.11: Pseudo-univariate Darstellung des PCR Modells a.) für die mit Meancentering und Offset-Korrektur vorbehandelten Daten (pret.) und b.) für die nicht vorbehandelten Daten. Mehrere Punkte an einer Konzentration beziehen sich auf Reproduzierbarkeitsmessungen.