6 Anwendung der Faktor–Selektion mit Genetischen Algorithmen auf reale Datensätze und Disskussion der Ergebnisse

In diesem Kapitel werden die Ergebnisse der Untersuchungen zum Einsatz Genetischer Algorithmen bei der Faktor–Selektion in der Nah-Infrarot–Spektrochemometrie (NIR) beschrieben.

Zunächst wird detailliert auf die Problematik der Modellbildung eingegangen. Anhand einer Kalibration des Proteingehaltes in Weizen werden Overfitting- und Underfitting-Effekte verdeutlicht und die Herleitung geeigneter Fitnessfunktionen zur Behebung dieses Problems dargestellt.

2Das Hauptthema dieser Arbeit sind Untersuchungen zur automatisierten Erstellung quantitativer NIR-Kalibrationen im Kontext der Faktor–Selektion in der *Principal Component Regression* (PCR). Neben Vergleichen zwischen Ergebnissen klassischer Selektionsverfahren und Genetischer Algorithmen spielt die Bewertung der Güte von Kalibrationsmodellen eine wichtige Rolle. Der Abschnitt 6.2 befaßt sich mit dieser Thematik und faßt die auf Basis realer Datensätze erhaltenen Ergebnisse zusammen. Dem Aspekt der Automatisierung dieser komplexen Optimierungen kam besondere Bedeutung zu. Daher schließen sich in 6.3 die Untersuchungsergebnisse zu diesem Themenkreis an. Hier wird dargelegt, wie durch geeignete Fitnessfunktionen eine vollständige Automatisierung der Faktor–Selektion erreicht werden konnte.

6.1 Over- und Underfitting

Bereits in 3.3 wurde auf die Problematik von *over-* und *underfitting* im Zusammenhang mit der Faktor–Selektion hingewiesen. Die Auslegung der Fitnessfunktion ist daher gerade im Zusammenhang mit der Automatisierung der Faktor–Selektion auf Basis eines Genetischen Algorithmus von entscheidender Bedeutung. Die Fitnessfunktion muß so ausgelegt sein, daß sie sowohl Under- als auch Overfitting vermeidet, dabei jedoch ein adäquates Maß für die Güte des Kalibrationsmodells darstellt.

Die Güte einer Kalibration läßt sich durch die Präzision beschreiben, mit der bekannte Werte einer Probeneigenschaft einzelner Spektren durch das Modell vorhergesagt werden können. Dabei läßt sich grundsätzlich zwischen Standards, die in den Kalibrationsdatensatz eingehen und solchen, die nicht Teil dieses Datensatzes sind, differenzieren. Letztere sind unabhängig von der Kalibration und werden als Validationspektren bezeichnet.

Zur Bewertung der Präzision werden die Vorhersagefehler der Standards in Form von Standardabweichungen herangezogen. Für den Kalibrationsdatensatz ist dies der *Standard Error of Estimate* (SEE-Wert) und für die Validationsdaten der *Root Mean Square Error of Prediction* (RMSEP–Wert):

$$\mathrm{SEE} = \sqrt{\frac{\sum_{i=1}^{n_s} (\hat{p}_i - p_i)^2}{n_s - n_f - 1}} \quad , \qquad \mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_e} (\hat{p}_i - p_i)^2}{n_e}} \quad .$$

Hier wird die Abweichung zwischen den Referenzwerten p_i und den Vorhersagewerten $\hat{p}_i$ für die zu kalibrierende Probeneigenschaft der Kalibrations- bzw. Validationsstandards (n_s bzw. n_e Spektren) bestimmt. In die Berechnung des SEE-Werts geht zusätzlich die Zahl der Freiheitsgrade ein, die durch die Berücksichtigung von n_f Variablen im Kalibrationsmodell wegfallen und die für die Erhöhung der statistischen Unsicherheit verantwortlich sind (vgl. 3.3).

Die Beurteilung und Optimierung von Kalibrationsmodellen auf Basis der SEE- und RMSEP–Werte ist generell nicht unproblematisch. Unabhängig vom verwendeten Zerlegungs- und Regressionsverfahren (Eigenwert-, Waveletzerlegung, etc.) ist das Ziel, ein ausgewogenes Verhältnis zwischen der Berücksichtigung zu weniger Variablen (Underfitting) und zu vieler Variablen (Overfitting) zu finden.

Werden zu wenige Variablen im Kalibrationsmodell berücksichtigt, so wird die verfügbare spektrochemische Information nicht voll erfaßt — daraus resultiert ein erhöhter Vorhersagefehler (Underfitting). Im Fall von Overfitting werden so viele latente Variablen (Principal Components, Faktoren) berücksichtigt, daß ein zu komplexes Kalibrationsmodell entsteht. Dieses ist derart stark an die Kalibrationsdaten angepaßt, daß nicht nur relevante Information in das Modell eingeht, sondern auch spektrale Minoritätseffekte und sogar Rauschen Berücksichtigung finden. Im statistischen Sinne stellen diese Effekte vornehmlich zufällig verteilte Fehler dar, die sich nicht in jedem Spektrum in gleicher Größenordnung wiederfinden. Daher wirkt sich Overfitting insbesondere störend auf die Vorhersage unabhängiger Standards aus und führt zu erhöhten RMSEP–Werten. In bezug auf die Validation und zukünftige Vorhersagen sind solche Kalibrationsmodelle daher nicht stabil.

Ein Kalibrationsmodell, das eine gute Balance zwischen Over- und Underfitting darstellt, wird als *robust* bezeichnet (*s.* 3.3.1). Es soll bei Verwendung möglichst

weniger latenter Variablen eine realistische Einschätzung der zukünftig zu erwartenden Standardabweichung des Vorhersagefehlers geben.

Im Hinblick auf diese Problematik ist die Auslegung der Zielfunktion des Genetischen Algorithmus nicht trivial: Die Fitnessfunktion darf aus Gründen der Eindeutigkeit nur einen einzigen Wert zur Beurteilung einer potentiellen Lösung liefern (*s.* 4.3.1). Mit zunehmender Anzahl der Variablen, aus denen der Genetische Algorithmus Lösungen auswählen kann, steigt bei der ausschließlichen Verwendung des SEE- bzw. des RMSEP–Werts als Optimierungskriterium die Gefahr der Überanpassung des Kalibrations- bzw. des Validationsdatensatzes. Instabile, das heißt nicht robuste Kalibrationen sind daher oft das Ergebnis.

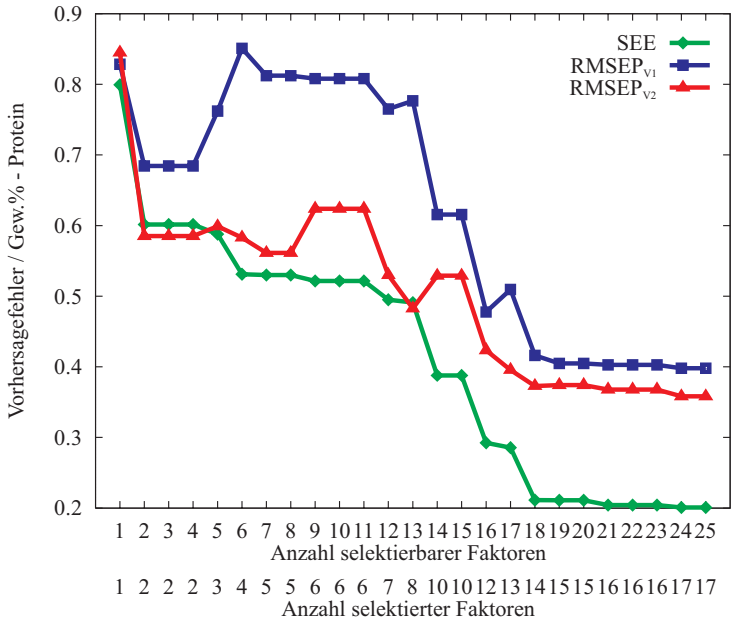

Abbildung 6.1: Optimierung der Protein–Kalibration des Weizen–Datensatzes auf Basis des SEE-Werts.

Am Beispiel der Protein–Kalibration des Weizen–Datensatzes sollen diese Effekte verdeutlicht werden. Es wurden unterschiedliche Kalibrationen mit verschiedenen Fitnessfunktionen durchgeführt, wobei die Faktorzerlegung und die Regressionsrechnung stets mit dem gleichen Kalibrationsdatensatz erfolgte. In Abbildung 6.1 ist die GA–Optimierung der Protein–Kalibration auf Basis des SEE-Werts als Fitnessfunktion dargestellt. Sukzessive wurde dazu die Anzahl der latenten Variablen (Faktoren) $\mathcal{F}$, aus denen der Genetische Algorithmus während der Optimierung auswählt

und das Kalibrationsmodell bildet, von 1 auf 25 erhöht[1]. Die Anzahl der vom GA tatsächlich ausgewählten Faktoren ist unmittelbar darunter notiert. Die wachsende Zahl selektierter Faktoren entspricht einer zunehmenden Komplexität des Kalibrationsmodells, dessen Vorhersagefehler anhand des Kalibrationsdatensatzes WHT_K und der Validationsdatensätze WHT_{V1} und WHT_{V2} ermittelt wurde. Die Faktoren sind nach fallenden Eigenwerten sortiert (Eigenwert–Ranking). Wie erwartet, sinkt der Vorhersagefehler der einzelnen Datensätze zunächst ab, da die ersten selektierten Faktoren wesentliche spektrale Informationen enthalten. Doch schon ab dem dritten selektierten Faktor treten deutliche Overfitting–Effekte auf: Der Vorhersagefehler im ersten Validationsdatensatz RMSEP_{V1} steigt sprunghaft an, während der Vorhersagefehler des Kalibrationsdatensatzes (SEE) weiterhin stetig abnimmt. Auch der zweite Validationsdatensatz RMSEP_{V2} zeigt mit zunehmendem Freiheitsgrad ein Maximum. Bei sehr hohen Freiheitsgraden tendiert der Vorhersagefehler für beide Validationsdatensätze gegen einen Wert von 0.4 Gew.% Protein, während für die Kalibration ein Wert von 0.2 Gew.% erreicht wird. Nach dem bisher Gesagten dürfte jedoch klar sein, daß diese Werte nicht als realistisch anzusehen sind, sondern ein erhebliches Maß an Overfitting beinhalten, worauf auch die hohe Zahl der selektierten Faktoren (17 bei $\mathcal{F} = 25$) hinweist.

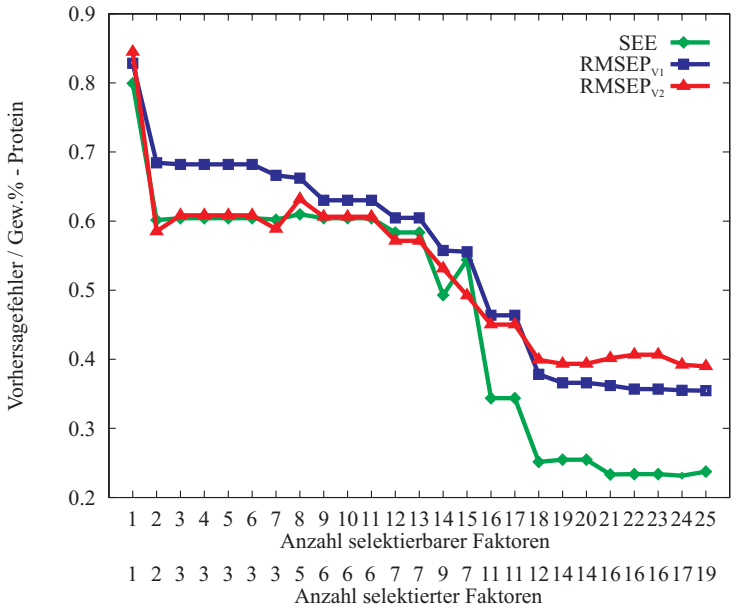

Abbildung 6.2: Optimierung der Protein–Kalibration des Weizen–Datensatzes auf Basis des RMSEP_{V1}-Werts.

Wird der Vorhersagefehler des ersten Validationsdatensatzes (RMSEP_{V1})

[1]Die Zahl wird als Freiheitsgrad $\mathcal{F}$ bezeichnet. $\mathcal{F}$ kann maximal den Wert der Anzahl der Kalibrationsstandards (hier 50) annehmen.

als Fitnessfunktion eingesetzt, so nimmt nun naturgemäß dieser mit zunehmendem Freiheitsgrad stetig ab (*s.* Abb. 6.2). Bei den beiden anderen Fehlergrößen (SEE und RMSEP_{V2}) sind wiederum einzelne Punkte im Optimierungsverlauf zu erkennen, die auf Overfitting hinweisen: So nimmt der Vorhersagefehler des Kalibrationsdatensatzes (SEE-Wert) bei 15 Freiheitsgraden und der des zweiten Validationsdatensatzes (RMSEP_{V2}-Wert) bei 3 und 8 Freiheitsgraden zu. Im Gesamtverlauf ist die Diskrepanz zwischen den einzelnen Fehlerwerten etwas kleiner als im Falle der SEE-Wert basierten Optimierung (*s.* Abb. 6.1). Aber auch hier kommt durch die hohe Zahl selektierter Faktoren (19 bei 25 Freiheitsgrade) die Überanpassung an den ersten Validationsdatensatz zum Ausdruck. Welche Faktoren im einzelnen in den SEE- und RMSEP–Wert basierten Optimierungen selektiert wurden, kann der Tabelle A.2 im Anhang entnommen werden.

Beide bisher verwendeten Fitnessfunktionen können das Auftreten von Overfitting und damit die Berücksichtigung irrelevanter spektraler Effekte nicht hinreichend unterbinden. Um Overfitting zu vermeiden oder zumindest deutlich zu reduzieren, wurde daher in Zusammenarbeit mit *A. Niemöller* [82] versuchsweise eine neue Fitnessfunktion, der sog. *Standard Error of Estimate and Prediction* (SEEP) definiert. Für die Anwendbarkeit des SEEPs ist ein ausreichend großer Spektrensatz notwendig, der, wie im Falle des Weizen–Datensatzes, in einen Kalibrations- und zwei Validationsdatensätze aufzuteilen ist. Der SEEP–Wert berechnet sich dann wie folgt:

$$\mathrm{SEEP} = \sqrt{\frac{(n_s - n_f - 1) \cdot \mathrm{SEE}^2 + n_e \cdot \mathrm{RMSEP}^2}{n_s - n_f - 1 + n_e}} \quad . \tag{6.1}$$

Durch Einsetzen der Gleichungen (3.53) und (3.56) ergibt sich

$$\mathrm{SEEP} = \sqrt{\frac{\sum_{i=1}^{n_s} (\hat{p}_i - p_i)^2 + \sum_{i=1}^{n_e} (\hat{p}_i - p_i)^2}{n_s - n_f - 1 + n_e}} \quad . \tag{6.2}$$

In dieser Fitnessfunktion wird neben dem Vorhersagefehler des Kalibrationsdatensatzes auch der Vorhersagefehler des ersten Validationsdatensatzes berücksichtigt. Dieser Datensatz wird im folgenden als (kalibrations)interner Validationsdatensatz bezeichnet. Der zweite (kalibrations)externe Validationsdatensatz dient der unabhängigen Validation.

Das Ergebnis der Faktor–Selektion auf Basis des SEEP–Werts als Funktion wird in Abbildung 6.3 wiedergegeben. Was den Verlauf der Vorhersagefehler betrifft, so ändert sich das Bild gegenüber Abb. 6.3 und 6.2 nicht grundlegend, und auch die Zahl der selektierten Variablen ist nur unwesentlich niedriger.

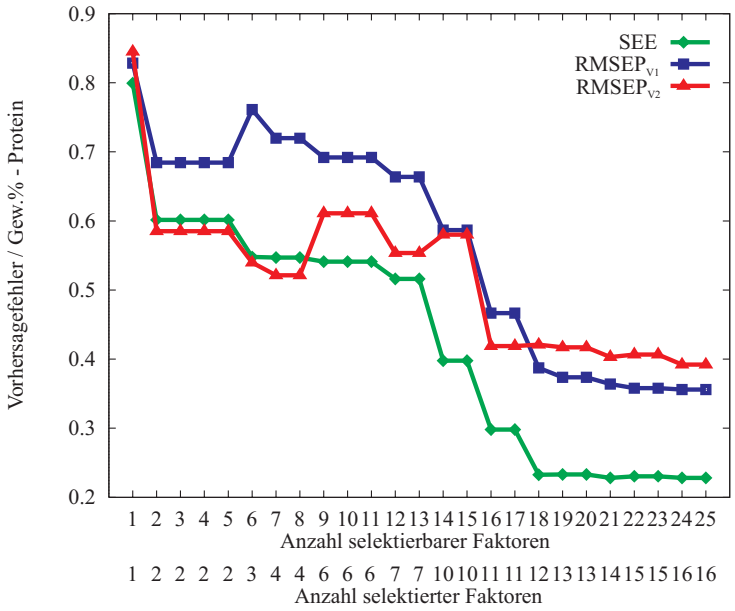

Abbildung 6.3: Optimierung der Protein–Kalibration des Weizen–Datensatzes auf Basis des SEEP-Werts.

Als weitere Alternative wurde der Kalibrationsdatensatz durch den ersten Validationsdatensatz erweitert und wiederum wie in Abbildung 6.2 eine Optimierung auf Basis des SEE–Werts durchgeführt (*s.* Abb. 6.4).

Der Vorhersagefehler des erweiterten Kalibrationsdatensatzes dient hier also als Fitnessfunktion und nimmt mit steigender Zahl von $\mathcal{F}$ stetig ab. Aber auch hier sind deutliche Effekte von Overfitting am Verlauf der externen Validation ablesbar. Die Standardabweichung des Vorhersagefehlers für den Datensatz WHT_{V2} wird mit zunehmender Modell–Komplexität größer, und im Vergleich zu den bisherigen Kalibrationen zeigt sich bei $\mathcal{F} = 25$ ein deutlich höherer Fehler von 0.45 Gew.% Proteingehalt.

Auch die hohe Anzahl selektierter Faktoren ist ein Indiz für die Überanpassung des Kalibrationsdatensatzes. Die Abbildung 6.5 zeigt die Zunahme der Modellkomplexität im Vergleich zur Kalibrationsoptimierung auf Basis der SEEP-Fitnessfunktion wie sie oben beschrieben wurde (*s.* Abb. 6.3). Wie aus Tabelle 6.1 hervorgeht, werden ab $\mathcal{F} = 14$ Freiheitsgrade zunehmend mehr Faktoren in der Kalibration mit dem erweiterten Datensatz berücksichtigt, als dies bei Verwendung der SEEP-Fitnessfunktion mit 50 Kalibrationsstandards der Fall ist. Auch dies ist ein deutlicher Hinwies auf Overfitting. Wenn der erste Validationsdatensatz in die Kalibration mit einfließen soll, ist es daher eindeutig vorteilhafter, nicht den Kalibrationsdatensatz entsprechend zu erweitern, sondern den Validationsdatensatz bei der Berechnung der Fitnessfunktion (z.B. in Form des beschriebenen SEEP–Werts) zu berücksichtigen.

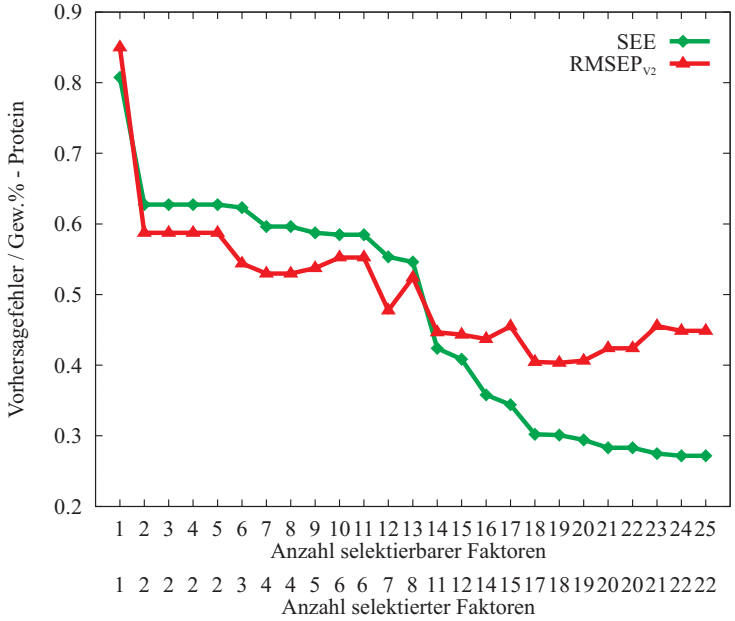

Abbildung 6.4: Optimierung der Protein–Kalibration des Weizen–Datensatzes auf Basis des SEE-Werts mit 70 Kalibrationsspektren.

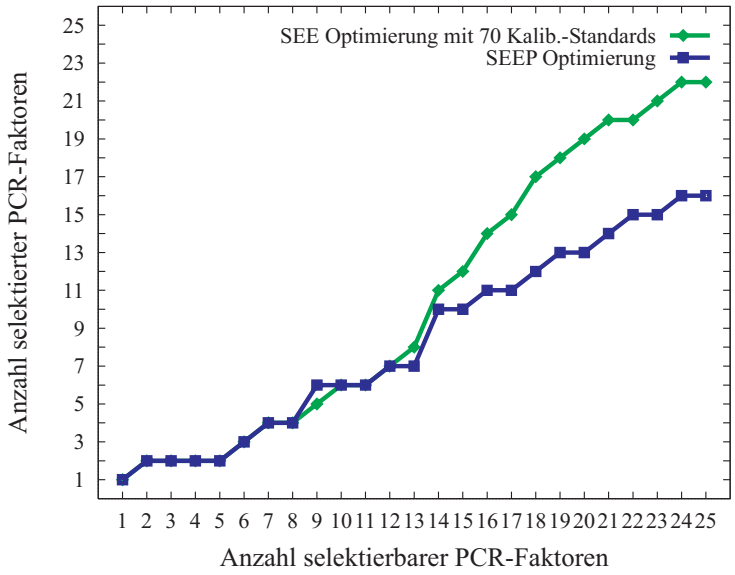

Abbildung 6.5: Vergleich der Anzahl selektierter latenter Variablen auf Basis des SEEP-Werts und 50 Kalibrationsstandards und des SEE-Werts mit 70 Kalibrationsstandards.

Tabelle 6.1: Liste selektierter Faktoren der Protein–Kalibrationen auf Basis unterschiedlicher Fitnessfunktionen. Der Wert von n_f gibt die Anzahl der selektierten Faktoren an.

$\mathcal{F}$	n_f	SEEP–Wert basierte Optimierung mit 50 Kal.-Standards	n_f	SEE–Wert basierte Optimierung mit 70 Kal.-Standards
1	1	1	1	1
2	2	1,2	2	1,2
3	2	1,2	2	1,2
4	2	1,2	2	1,2
5	2	1,2	2	1,2
6	3	1,2,6	3	1,2,6
7	4	1,2,6,7	4	1,2,6,7
8	4	1,2,6,7	4	1,2,6,7
9	6	1-3,6,7,9	5	1,2,6,7,9
10	6	1-3,6,7,9	6	1,2,6,7,9,10
11	6	1-3,6,7,9	6	1,2,6,7,9,10
12	7	1-3,6,7,9,12	7	1,2,6,7,9,10,12
13	7	1-3,6,7,9,12	8	1,2,6,7,9,10,12,13
14	10	1-3,5-9,12,14	11	1-3,5-7,9,10,12-14
15	10	1-3,5-9,12,14	12	1-3,5-7,9,10,12-15
16	11	1,2,5-9,12-14,16	14	1-7,9,10,12-16
17	11	1,2,5-9,12-14,16	15	1-7,9,10,12-17
18	12	1,2,5-9,12-14,16,18	17	1-7,9,10-18
19	13	1,2,5-9,12-14,16,18,19	18	1-7,9-19
20	13	1,2,5-9,12-14,16,18,19	19	1-7,9-20
21	14	1,2,5-9,12-14,16,18,19,21	20	1-7,9-21
22	15	1,2,5-9,12-14,16,18,19,21,22	20	1-7,9-21
23	15	1,2,5-9,12-14,16,18,19,21,22	21	1-7,9-21,23
24	16	1,2,5-9,12-14,16,18,19,21,22,24	22	1-7,9-21,23,24
25	16	1,2,5-9,12-14,16,18,19,21,22,24	22	1-7,9-21,23,24

6.1.1 Sortierung nach Korrelationskoeffizienten

Die bisher durchgeführten Kalibrationen zeigen, daß es mit GA–basierten Faktor–Selektionen allein nicht möglich ist, Overfitting zu vermeiden. Im weiteren werden daher die Selektionsmöglichkeiten eingeschränkt, indem nur diejenigen latenten Variablen berücksichtigt werden, die mit der zu kalibrierenden Eigenschaft korrelieren. Der Einfluß von Faktoren, die überwiegend Rauschen repräsentieren und sich somit negativ auf die Robustheit des Kalibrationsmodells auswirken, kann so stark eingeschränkt werden.

Im Falle der Faktor–Selektion ist eine solche Vorauswahl durch die Korrelation der Faktorgewichte (Scores) mit der zu kalibrierenden Eigenschaft der Standards erreichbar. Die einzelnen Korrelationskoeffizienten r_i sind die Einträge des Korrelationsvektors $\mathbf{r} \in \mathbb{R}^{1\times k}$. Sie lassen sich aus der Score-Matrix $\mathbf{C}$ und dem Eigenschafsvektor $\mathbf{p}$ wie folgt errechnen

$$r_i(\mathbf{c_i}, \mathbf{p}) = \frac{\mathbf{c_i}^{zT}\,\mathbf{p}^z}{\|\mathbf{c_i}^z\|\,\|\mathbf{p}^z\|} \quad \text{für} \quad i = 1, \ldots, k, \text{ oder} \tag{6.3}$$

$$= \frac{\beta_i}{\|\mathbf{c_i}^z\|\,\|\mathbf{p}^z\|} \quad \text{mit} \quad \beta_i = \mathbf{c_i}^{zT}\,\mathbf{p}^z. \tag{6.4}$$

Dabei ist $\|\cdot\|$ die Euklidische Norm eines Vektors mit

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}} \tag{6.5}$$

und $\mathbf{a}^z$ ein zentrierter Spaltenvektor $\mathbf{a}$ der Länge n dessen Elemente a_i sich nach

$$a_i^z = a_i - m_a \quad \text{mit} \quad m_a = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{6.6}$$

berechnen. Der Vektor $\boldsymbol{\beta} \in \mathbb{R}^k$ bezeichnet den Regressionsvektor der Faktoren aus der Principal Component Regression (*s.* 3.2.1). Die skalare Größe β_i steht also für den Regressionskoeffizienten des i-ten Faktors in der PCR.

Durch Sortierung der Faktoren nach fallenden r^2–Werten (Korrelations–Ranking) finden insbesondere bei einer kleinen und mittleren Zahl der Freiheitsgrade vorrangig Faktoren Eingang ins Kalibrationsmodell, die eine hohe Korrelation zu der zu kalibrierenden Eigenschaft aufweisen.

Dem Genetischen Algorithmus steht in diesem Fall, neben der Information aus der Fitnessfunktion durch die Position der latenten Variablen in der Korrelationsrangliste, auch Information über die potentielle Bedeutung der einzelnen Faktoren für das Kalibrationsmodell zur Verfügung. Dieser zusätzliche Informations–Input ist ein Beispiel für *Hybridisierung* (*s.* 7.2).

Für die Modellbildung bringt dies folgende Konsequenzen mit sich: Die Rangfolge der orthogonalen Principal Components (PCs) nach ihren Korrelationskoeffizienten entscheidet bei einer vorgegebenen Anzahl der Freiheitsgrade über das beste Modell. Der Einsatz eines Genetischen Algorithmus ist in diesem Kontext bei einer ausschließlich kalibrationsinternen Validation unnötig. In Verbindung mit einer *top–down variable selection* wird diese Form der Selektion als *correlated principal component regression* (CPCR) bezeichnet [45] (*s.* 4.1.1).

Der Genetische Algorithmus ist dann erforderlich, wenn, wie im Falle der SEEP–Fitnessfunktion, eine Kalibrationsoptimierung unter Berücksichtigung von Validationsdatensätzen erfolgen soll. Die Score-Matrizen der Validationsdatensätze

sind im Gegensatz zum Kalibrationsdatensatz *nicht* orthogonal[2]. Damit ergeben sich durch den Einsatz des GA Lösungen, die in der Vorhersage von Properties zukünftig zu messender Spektren potentiell besser sind als Modelle, die ausschließlich unter Berücksichtigung des Kalibrationsdatensatzes erstellt werden.

Für das Beispiel der Kalibration des Proteingehaltes in Weizen ergeben sich die im Anhang Seite 161 in Tabelle A.3 angegebenen Korrelationen. Eine Sortierung der latenten Variablen nach fallenden r^2–Werten führt zu der in der dritten und vierten Spalte dieser Tabelle dargestellten Rangfolge. An dem Datensatz wird deutlich, daß sich die einzelnen Korrelationskoeffizienten oft nur geringfügig voneinander unterscheiden. Dieser problematische Aspekt tritt besonders bei großen Datensätzen regelmäßig in Erscheinung und erschwert die Differenzierung zwischen relevanter spektraler Information und zu vermeidenden Minoritätseffekten.

Das Ergebnis der Faktor–Selektion wird in Abbildung 6.6 wiedergegeben. Der Vorteil des Korrelations–Rankings wird durch die stetige Verbesserung der Standardabweichungen des Kalibrationsdatensatzes (SEE–Wert) und des internen Validationsdatensatzes (RMSEP_{V1}) deutlich. Dennoch tritt auch in diesem Fall ein Overfittingeffekt in bezug auf den externen Validationsdatensatz auf. Die Standardabweichung (RMSEP_{V2}) durchläuft bei neun Freiheitsgraden ein Minimum und steigt dann wieder an. Ab etwa elf Freiheitsgraden verändert sich der RMSEP_{V2}–Wert kaum noch und bleibt auf einem konstant hohen Niveau. Ähnliches gilt für den RMSEP_{V1}–Wert des internen Datensatzes, der bereits ab dem neunten Freiheitsgrad keine signifikanten Veränderungen mehr aufweist.

Tabelle 6.2: Selektierte Faktoren der Protein–Kalibration nach Korrelations–Ranking

SEE	RMSEP		$\mathcal{F}$	n_f	selektierte Faktoren
	$V1$	$V2$			
0.664	0.761	0.641	1	1	2
0.602	0.684	0.585	2	2	1,2
0.536	0.636	0.507	3	3	1,2,14
0.471	0.655	0.462	4	4	1,2,6,14
0.410	0.572	0.464	5	5	1,2,6,14,16
0.373	0.550	0.481	6	6	1,2,6,12,14,16
0.332	0.519	0.460	7	7	1,2,6,12,14,16,18
0.292	0.464	0.396	8	8	1,2,5,6,12,14,16,18
0.267	0.409	0.349	9	9	1,2,5,6,9,12,14,16,18
0.250	0.398	0.394	10	10	1,2,5,6,9,12-14,16,18
0.235	0.405	0.421	11	11	1,2,5-7,9,12-14,16,18

Fortsetzung auf der folgenden Seite

[2]Ein Beispiel findet sich im Anhang A.3.

Selektierte Faktoren der Protein–Kalibration (Fortsetzung)

SEE	RMSEP $V1$	RMSEP $V2$	$\mathcal{F}$	n_f	selektierte Faktoren
0.235	0.405	0.421	12	11	1,2,5-7,9,12-14,16,18
0.235	0.405	0.421	13	11	1,2,5-7,9,12-14,16,18
0.218	0.384	0.435	14	13	1,2,5-7,9,12-14,16-18,35
0.223	0.376	0.409	15	13	1,2,5-7,9,12-14,16,18,33,35
0.212	0.374	0.412	16	14	1,2,5-7,9,12-14,16-18,21,35
0.212	0.374	0.412	17	14	1,2,5-7,9,12-14,16-18,21,35
0.207	0.375	0.397	18	15	1,2,5-7,9,12-14,16-18,21,35,46
0.207	0.375	0.397	19	15	1,2,5-7,9,12-14,16-18,21,35,46
0.207	0.375	0.397	20	15	1,2,5-7,9,12-14,16-18,21,35,46
0.203	0.358	0.411	21	16	1,2,5-9,12-14,16-18,21,35,46
0.199	0.359	0.409	22	17	1,2,5-9,12-14,16-18,21,24,35,46
0.199	0.359	0.409	23	17	1,2,5-9,12-14,16-18,21,24,35,46
0.201	0.350	0.411	24	17	1,2,5-9,12-14,16-18,21,26,35,46
0.202	0.345	0.404	25	18	1,2,5-9,12-14,16-18,21,24,26,29,35

Das Risiko, das bei der Sortierung der latenten Variablen nach fallenden Korrelationen besteht, wird aus der Auflistung der selektierten Faktoren in Tabelle 6.2 ersichtlich. Je größer die Zahl der Freiheitsgrade $\mathcal{F}$ wird, desto mehr steigt die Gefahr, Faktoren im Kalibrationsmodell zu berücksichtigen, die mit größter Wahrscheinlichkeit nur noch spektrale Minoritätseffekte repräsentieren (z.B. die Faktoren 29, 35 oder 46). Auch die Verwendung der SEEP-Fitnessfunktion bietet hier keinen ausreichenden Schutz vor Overfitting.

Vorteile ergeben sich dagegen durch die Anwendung des Korrelations-Rankings vor allem für kleine Freiheitsgrade ($\mathcal{F} \leq 10$), da nur Faktoren in die Berechnung Eingang finden, die eine hohe Korrelation zur Property aufweisen. Mit der SEEP-Fitnessfunktion und den beiden vorgestellten Methoden zur Sortierung der Faktoren (nach Eigenwerten oder Korrelation) können für ausreichend große Datensätze robuste Kalibrationen erstellt werden. Dabei findet die Faktoranalyse nach wie vor ausschließlich auf Basis des Kalibrationsdatensatzes statt. Nur zur Bewertung der Modellanpassung wird der interne Validationsdatensatz mit herangezogen. Auf diese Weise wird unter Berücksichtigung möglichst weniger latenter Variablen sowohl der Kalibrationsfehler als auch der in Zukunft zu erwartende Analysenfehler möglichst klein gehalten.

Untersuchungen zu diesem Themenkreis wurden auch von *A. Niemöller* im Zusammenhang mit der Selektion von Wavelet–Koeffizienten im Rahmen der Wavelet–Coefficient–Regression (WCR) angestellt [82].

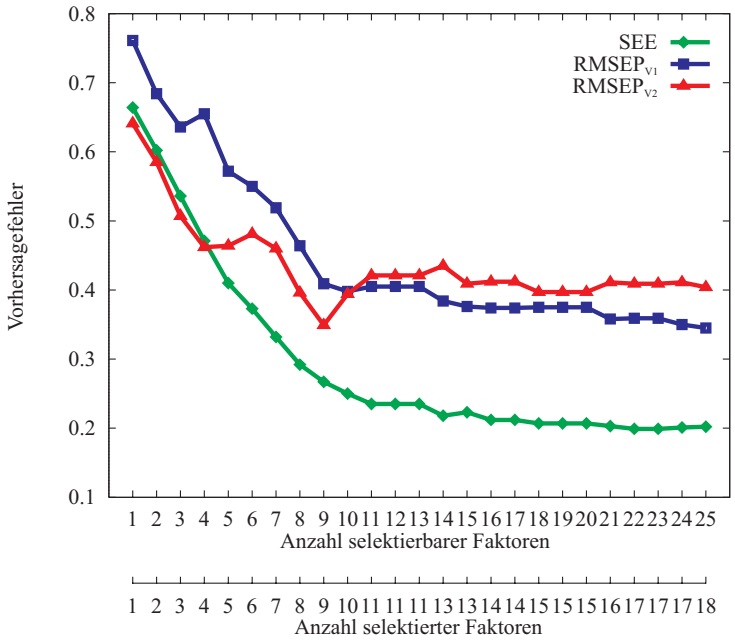

Abbildung 6.6: Optimierung der Protein–Kalibration des Weizen–Datensatzes auf Basis des SEEP-Werts mit Korrelations–Ranking.

6.2 Gegenüberstellung von Genetischen Algorithmen und klassischen Selektionsverfahren

Will man robuste Kalibrationen mit automatisierten Selektionsverfahren erstellen, dann ist eine Differenzierung zwischen unterschiedlichen Kalibrationsmodellen anhand geeigneter Kriterien unumgänglich. Die folgenden Abschnitte befassen sich daher mit der Bewertung der Kalibrationsmodelle, die aus den Optimierungen für die einzelnen Freiheitsgrade $\mathcal{F}$ resultieren. Am Beispiel der Olfen- und Weizen–Datensätze wird verdeutlicht, wie sich die Modelle im Vergleich bewerten lassen.

Ein Kriterium zum Vergleich von Kalibrationsmodellen [100] läßt sich gemäß 4.4 sehr leicht wie folgt formulieren: Sind S_1 und S_2 die Analysenfehler zweier zu vergleichender Kalibrationsmodelle und d_1, d_2 die korrespondierenden relativen Varianzen, dann sind S_1 und S_2 signifikant verschieden, wenn sich die beiden Intervalle

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

nicht überlappen. Für die relativen Varianzen d_i gilt die folgende asymptotische Näherung:

$$d_i \approx \frac{1}{\sqrt{2n}} \quad . \tag{6.7}$$

n steht für die Anzahl der Freiheitsgrade, die in die Berechnung des Root Mean Square Errors of Prediction (RMSEP) in den Nenner eingehen.

Modelle, bei denen sich die Intervallgrenzen der zu vergleichenden Standardabweichungen überschneiden, werden als gleichwertig betrachtet. Überschneiden sich die Intervallgrenzen nicht, dann sind die zugehörigen Kalibrationen tatsächlich unterscheidbar.

Vergleichend sollen hier die Kalibrationsergebnisse unterschiedlicher Faktor–Selektionsverfahren anhand des Weizen– und des Olfen–Datensatzes aufgezeigt werden:

(i) *top–down variable selection* (TD),

(ii) *stepwise variable selection* (SV),

(iii) *correlated principal component regression* (CPCR),

(iv) *partial least squares* Typ1 (PLS-1) und

(v) auf Basis von Genetische Algorithmen: Selektion von Faktoren,

- die nach fallenden Eigenwerten (Eigenwert–Ranking) oder
- nach fallenden Korrelationskoeffizienten (Korrelations–Ranking) sortiert sind.

Die Berechnungen wurden mit einer verschieden großen Anzahl der Freiheitsgrade durchgeführt und die Ergebnisse für die einzelnen Datensätze in Abbildung 6.7 veranschaulicht. Mit Ausnahme der Fitnessfunktion stellen die Ordinatenwerte den Root-Mean-Square Error of Prediction für den zweiten Validationsdatensatz dar (RMSEP_{V2}). Bedingt durch das Verfahren der *forward-backward–stepwise variable selection* (SV) wird für das Verfahren nur ein Selektionsergebnis erhalten (*s.* 4.1.1), welches durch eine horizontale Linie verdeutlicht wird. Alle übrigen Ergebniswerte sind von der Anzahl der berücksichtigten Freiheitsgrade ($\mathcal{F}$) abhängig und werden durch unterschiedliche Linientypen gekennzeichnet.

6.2.1 Kalibration des Weizen–Datensatzes

Der Datensatz setzt sich aus Spektren zusammen, die in diffuser Reflexion vermessen wurden und nur geringes Rauschen aufweisen (*s.* 5.2.1).

Kalibration des Feuchtigkeitsgehaltes im Weizen

Eigenwert–Sortierung

Für den Weizen–Datensatz zeigt Abbildung 6.7 die RMSEP–Werte des Feuchtigkeitsgehaltes in bezug auf den externen Validationsdatensatz (WHT_{V2}) und die Fitnessfunktion η = SEEP in Abhängigkeit der Anzahl der Freiheitsgrade $\mathcal{F}$. Aus

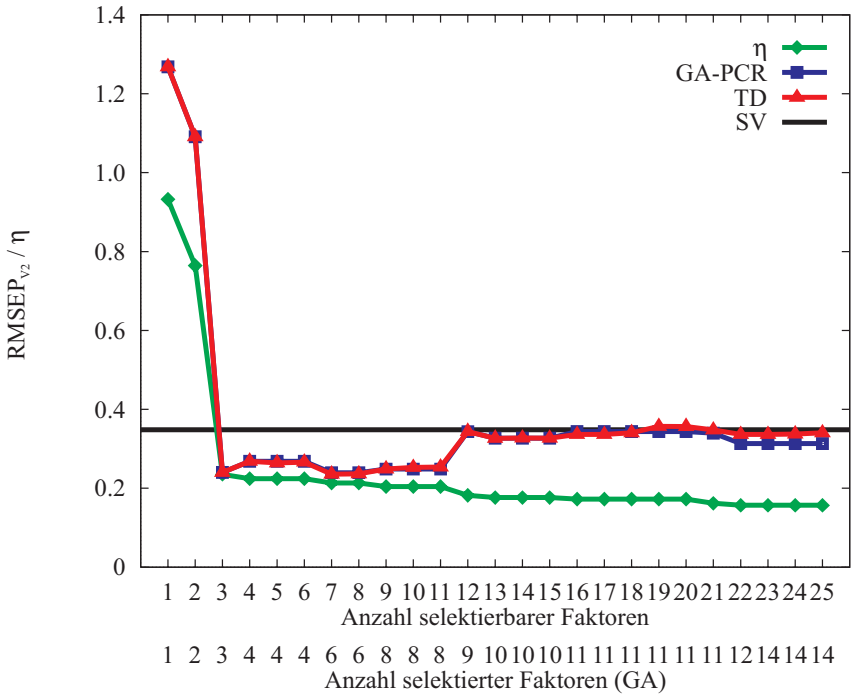

Abbildung 6.7: Ergebnisse der externen Validation (RMSEP_{V2}) in Abhängigkeit von der Anzahl der Freiheitsgrade für den Feuchtigkeitsgehalt von Weizen. η = SEEP ist die Fitnessfunktion. Es wird die Eigenwert–Sortierung eingesetzt.

dieser Abbildung wird auf Anhieb deutlich, daß das vom Genetischen Algorithmus (GA) und der Top–Down Variable Selection (TD) gefundene 3–Faktor–Modell eine gute Lösung darstellt. Nichtsdestoweniger findet die Stepwise–Variable Selection (SV) unter Verwendung des SEE-Werts als Optimierungskriterium ein Modell, das den 12. Faktor mitberücksichtigt. Dieser Zusammenhang spiegelt sich auch in der leichten Abnahme des Fitnessfunktionswerts η beim Schritt von $\mathcal{F} = 11$ nach $\mathcal{F} = 12$ wider. Jedoch wird durch die externe Validation deutlich, daß es sich hier um einen Overfitting–Effekt handelt, der zu einem Anstieg des Fehlers (RMSEP_{V2}) führt.

Die beiden Graphiken in Abbildung 6.8 stellen die zugehörigen Property–Weighting–Spektren (Regressionsvektoren) dar. Es ist offenkundig, daß die Hinzunahme des 12. Faktors zum Modell einen deutlich stärkeren Anteil an Rauschen in

Tabelle 6.3: Faktor–Selektionen für die Feuchtigkeit in Weizen auf Basis unterschiedlicher Methoden (Korrelations–Ranking).

Methode	$\mathcal{F}$	Faktor–Selektion		SEE	RMSEP_{V1}	RMSEP_{V2}
		n_f[a]	Selektion			
CPCR	3	3	1–3	0.238	0.230	0.240
SV	25	4	1–3,12	0.209	0.234	0.345
GA	3	3	1–3	0.238	0.230	0.240
GA	4	4	1–3,12	0.209	0.234	0.345

[a] n_f steht für die Anzahl (oder Nummer) selektierter Faktoren

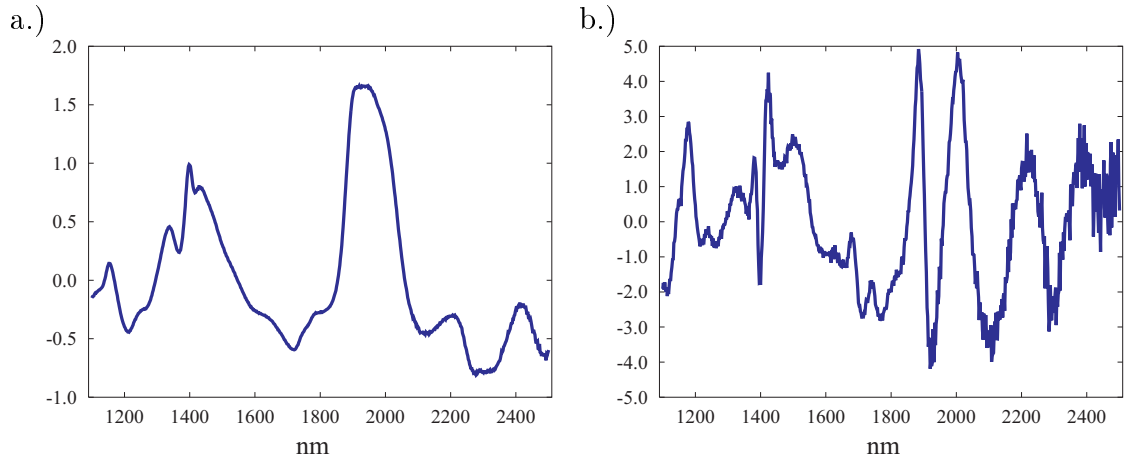

Abbildung 6.8: Property–Weighting–Spektren für den Feuchtigkeitsgehalt von Weizen des 3–Faktor–Modells a.) mit den Faktoren 1–3 und des 4–Faktor–Modells b.) mit den Faktoren 1–3, 12.

die Kalibration einbringt. Der Regressionsvektor verändert sich drastisch, vor allem in den für die quantitative Bestimmung von Wasser relevanten Bereichen der OH–Kombinationsschwingungen bei ca. 1400 nm (7140 cm^{-1}) und um ca. 2000 nm (5000 cm^{-1}). Der relativ unstetige und verrauschte Verlauf des Property–Weighting–Spektrums bei Hinzunahme des 12. Faktors unterstreicht, daß sich dieser negativ auf die Kalibration auswirkt.

Dieses Beispiel zeigt, daß Property–Weighting–Spektren in gewissem Grade einer chemischen Interpretation zugänglich sind und sich aus ihrer Form auch Rückschlüsse auf die Robustheit einer Kalibration ziehen lassen. Auch das über das Property–Weighting–Spektrum abgeleitete skalare *Net Analyte Signal* (NAS) stellt in Form der 'Empfindlichkeit' ein Gütekriterium für die Kalibration dar. Eine höhere Empfindlichkeit kann allgemein so verstanden werden, daß die für die zu kalibrie-

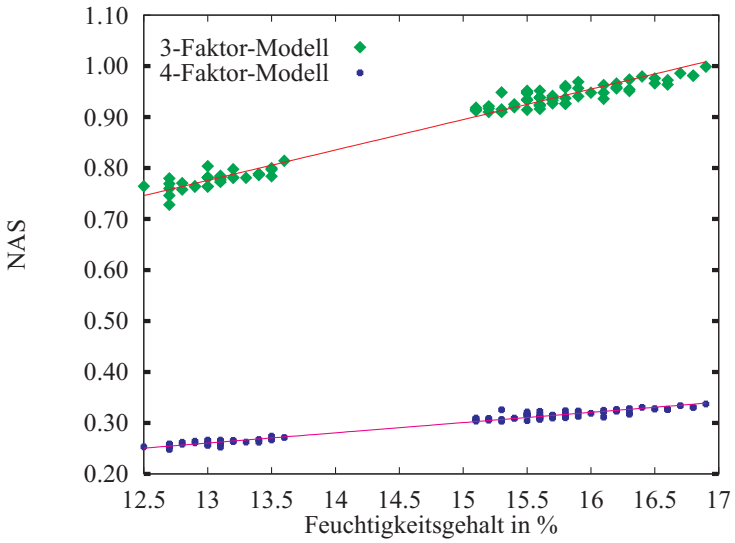

Abbildung 6.9: Net Analyte Signal für die Kalibration der Feuchtigkeit von Weizen.

rende Probeneigenschaft spezifischen spektralen Merkmale selektiver erfaßt werden als im Falle einer niedrigeren Empfindlichkeit. Wie Abbildung 6.9 zeigt, ist die Empfindlichkeit des 3–Faktor–Modells höher als die des 4–Faktor–Modells.

Ebenfalls ein 3–Faktor–Modell wird durch die Anwendung der Correlated Principal Component Regression (CPCR) und GA basierten Optimierung nach Korrelations–Ranking erzielt[3] (*s.* Tab. 6.3).

Vergleich zu Partial Least Squares Berechnungen

Ein Vergleich mit Kalibrationsergebnissen aus *Partial Least Squares* Rechnungen (PLS) und GA basierter *Principal Component Regression* (GA–PCR) belegt die gute Vorhersagequalität des zuvor beschriebenen 3–Faktor–Modells (*s.* Abb. 6.10). In der graphischen Darstellung werden die Standardabweichungen des Kalibrations- und der beiden Validationsdatensätze mit den Vorhersagefehlern des externen Validationsdatensatzes (RMSEP_{V2}) der GA–PCR Rechnung in Beziehung gesetzt. Die relative Unsicherheit des RMSEP_{V2}–Werts aus dem im vorigen Absatz diskutierten Kalibrationsmodell mit 3 Faktoren (1–3) bildet ein Intervall, das in der Abbildung als schraffierte Fläche dargestellt ist. Wie die GA basierte PCR erreicht auch die PLS mit einem 3–Faktor–Modell ein Minimum im Vorhersagefehler des WHT_{V2} Datensatzes. Darüber hinaus tritt auch hier ein deutlicher Overfitting–Effekt auf, der sich negativ auf die Standardabweichung des externen Validationsdatensatzes

[3]Die Numerierung der Faktoren entspricht in allen Abbildungen und Tabellen der Rangfolge im Hinblick auf die Eigenwert–Sortierung.

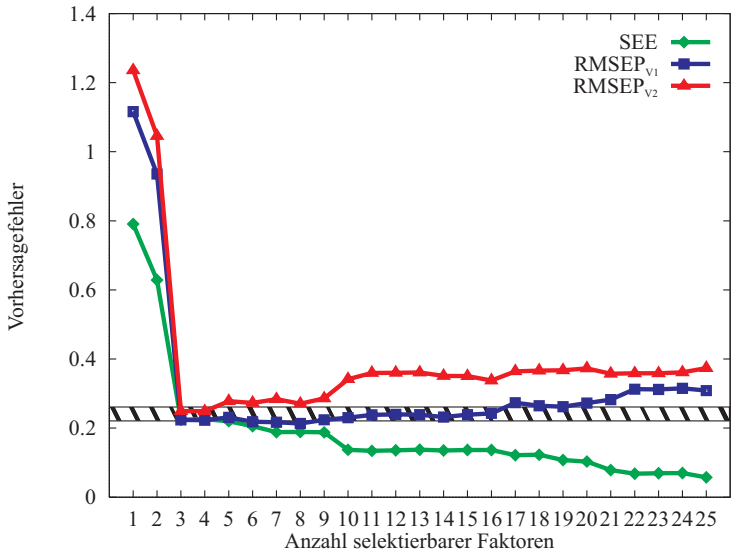

Abbildung 6.10: Kalibrationsergebnisse der PLS–Berechnung für die Feuchtigkeit in Weizen. Die schraffierte Fläche stellt die Varianzgrenzen des Analysenfehlers (RMSEP_{V2}) im 3–Faktor–Modell auf GA–PCR Basis dar.

auswirkt, wenn höhere Faktoren (Schritt vom neunten zum zehnten PLS–Faktor) in der Kalibration Berücksichtigung finden. In diesem Fall scheint es daher keinen grundsätzlichen Unterschied zwischen den Ergebnissen beider Kalibrationsverfahren zu geben.

Kalibration des Proteingehaltes im Weizen

Wie bereits in 6.1 dargelegt, ist die Situation im Falle der Protein–Kalibration weniger eindeutig.

Eigenwert–Sortierung

Bei der GA–basierten Faktor–Selektion mit Eigenwert–Ranking ändert sich der Verlauf der Fitnessfunktion $\eta = \mathrm{SEEP}$ im Bereich der Freiheitsgrade $2 \leq \mathcal{F} \leq 13$ relativ wenig (*s.* Abb. 6.11). Für $\mathcal{F} > 13$ sinkt η dann sehr stark bis zum Erreichen eines Plateaus bei $\mathcal{F} \geq 18$. Die Faktorkombination für $\mathcal{F} = 18$ (12 selektierte Faktoren) kann als eine potentiell gute Lösung des Kalibrationsmodells angesehen werden. Unter Berücksichtigung des externen Validationsfehlers ist im Sinne des Varianz–Kriteriums (6.7) allerdings keine weitere signifikante Verbesserung über $\mathcal{F} \geq 16$ hinaus feststellbar. Im Hinblick auf die externen Validationsdaten stellt das Modell mit $\mathcal{F} = 16$ (11 selektierte Faktoren) ebenfalls eine optionale Lösung des Kali-

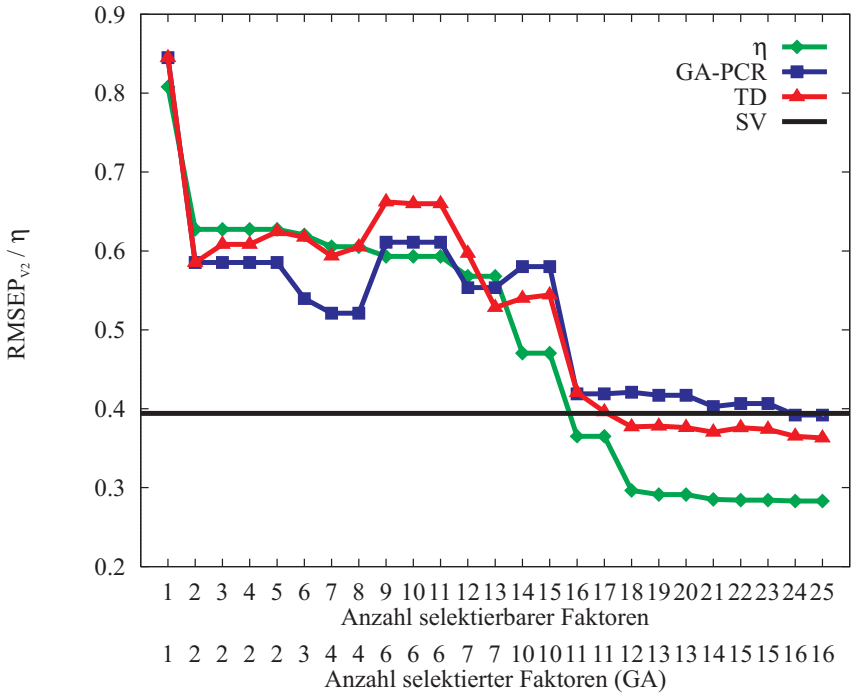

Abbildung 6.11: Ergebnisse der externen Validation (RMSEP$_{V2}$) in Abhängigkeit von der Anzahl der Freiheitsgrade für den Proteingehalt in Weizen. η = SEEP ist die Fitnessfunktion. Es wird die Eigenwert–Sortierung eingesetzt.

brationsproblems dar. In Tabelle 6.4 sind neben den Ergebnissen der GA–basierten Faktor–Selektion auch die Ergebnisse der alternativen Selektionsmethoden einander gegenübergestellt.

Tabelle 6.4: Faktor–Selektionen für Protein in Weizen auf Basis unterschiedlicher Methoden (Eigenwert–Sortierung).

Methode	$\mathcal{F}$	Faktor–Selektion		SEE	RMSEP_{V1}	RMSEP_{V2}
		n_f [a]	Selektion			
TD	16	16	1–16	0.308	0.482	0.420
SV	25	13	1–3,5–7,9,12–14,16–18	0.218	0.436	0.394
GA	16	11	1,2,5–9,12–14,16	0.298	0.467	0.419
GA	18	12	1,2,5–9,12–14,16,18	0.233	0.387	0.421

[a] n_f steht für die Anzahl (oder Nummer) selektierter Faktoren

Sortierung nach Korrelationskoeffizienten

Stellt man die Sortierung der Faktoren von der Eigenwertsortierung zur Sortierung nach fallenden Korrelationskoeffizienten um (*s.* Tab. A.6), dann entspricht die Top–Down Methode der Correlated Principal Component Regression (CPCR).

Sowohl der GA als auch die CPCR finden bis zum 11. Freiheitsgrad die gleichen Kombinationen latenter Variablen (Abb. 6.12). Für $\mathcal{F} \geq 11$ sinken die Werte der Fitnessfunktion η (SEEP) nur noch sehr wenig, und der Vorhersagefehler (RMSEP_{V2}) erreicht für die GA–basierte Selektion ein annähernd konstantes Niveau. Das aus den Berechnungen für $\mathcal{F} = 11$ resultierende Kalibrationsmodell (selektierte Faktoren: 1–11) kann daher als eine potentiell gute Lösung betrachtet werden. In derselben Abbildung zeigt die Standardabweichung der externen Validation (RMSEP_{V2}) ein Minimum für $\mathcal{F} = 9$. Die zugehörige Faktorkombination muß ebenfalls als Lösungsmöglichkeit des Kalibrationsproblems in Betracht gezogen werden, da η von $\mathcal{F} = 11$ nach $\mathcal{F} = 9$ nicht wesentlich größer wird. Die Modelle für $\mathcal{F} = 11$ und $\mathcal{F} = 9$ sind für die CPCR und die GA basierte Optimierung identisch. Eine Übersicht über die einzelnen Parameter gibt Tabelle 6.5.

Vergleicht man die Ergebnisse, die durch Eigenwert–Ranking erhalten werden (Tab. 6.4) mit jenen auf Basis von Korrelations–Ranking (Tab. 6.5), dann fällt auf, daß im Falle des Korrelations–Rankings alle Lösungsmodelle den 18. Faktor beinhalten, wohingegen der 8. Faktor fehlt. Dies hängt mit dem Umstand zusammen, daß im Fall des Korrelations–Rankings der 18. Faktor an 7. Position der Rangliste, der 8. Faktor aber erst an der 21. Position steht (*s.* Tab. A.6). Der 8. Faktor findet daher nur in Kalibrationsmodellen Berücksichtigung, für die $\mathcal{F}$ größer ist als 20. In

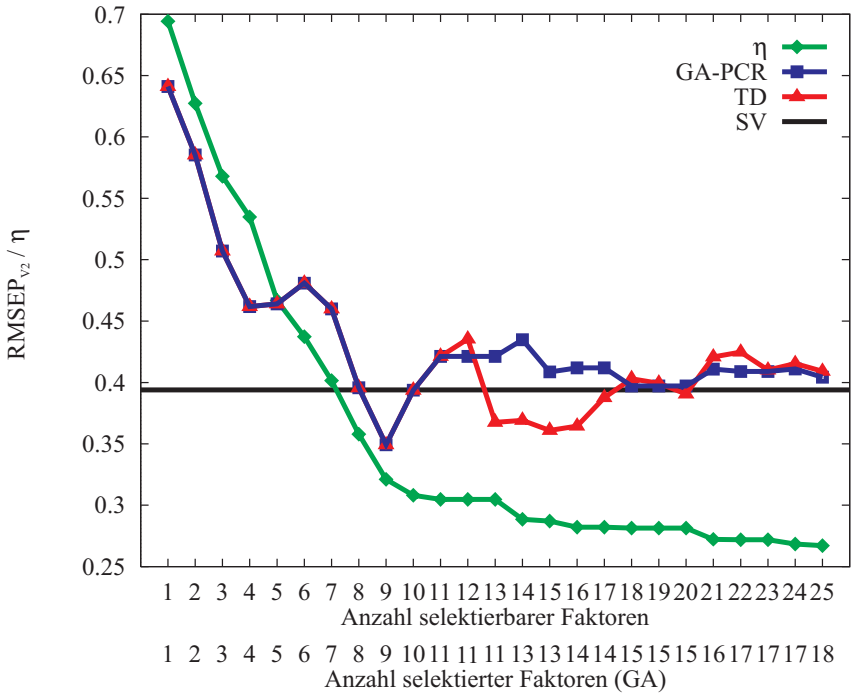

Abbildung 6.12: Ergebnisse der externen Validation (RMSEP$_{V2}$) in Abhängigkeit von der Anzahl der Freiheitsgrade $\mathcal{F}$ für den Proteingehalt von Weizen. η = SEEP ist die Fitnessfunktion. Es wird eine Sortierung nach fallender Korrelation eingesetzt.

Tabelle 6.5: Faktor–Selektion für Protein in Weizen auf Basis unterschiedlicher Methoden (Korrelations–Ranking).

Methode	$\mathcal{F}$	Faktor–Selektion		SEE	RMSEP_{V1}	RMSEP_{V2}
		n_f	Selektion			
SV	25	13	1–3,5–7,9,12–14,16–18	0.218	0.436	0.394
GA / CPCR	9	9	1,2,5,6,9,12,14,16,18	0.267	0.409	0.349
GA / CPCR	11	11	1,2,5–7,9,12–14,16,18	0.235	0.405	0.421

Tabelle 6.6: Vergleich von Faktor–Selektionen für Protein in Weizen. Fehlerschwellwerte gemäß des Varianz–Kriteriums

	Ranking Methode				
	Eigenwerte			Korrelation	
Methode	SV	GA	GA	GA/CPCR	GA/CPCR
$\mathcal{F}$	25	16	18	9	11
n_f Faktoren	13	11	12	9	11
Selektion	1–3,5–7,9, 12–14,16–18	1,2,5–9, 12–14,16	1,2,5–9, 12–14,16,18	1,2,5,6,9, 12,14,16,18	1,2,5–7,9, 12–14,16,18
η	0.314	0.365	0.296	0.321	0.305
obere Grenze	0.341	0.395	0.321	0.348	0.331
untere Grenze	0.287	0.334	0.271	0.294	0.279
RMSEP_{V2}	0.394	0.419	0.421	0.349	0.421
obere Grenze	0.456	0.485	0.488	0.404	0.488
untere Grenze	0.331	0.353	0.354	0.294	0.354

Tabelle 6.6 sind alle bisher beschriebenen potentiellen Lösungen mit RMSEP_{V2} und η zusammengefaßt. Mit Hilfe des Varianz–Kriteriums ist es möglich, die unterschiedlichen Kalibrationsmodelle zu bewerten. Für die Größen η und RMSEP_{V2} lassen sich mit Ausnahme einer Kalibration keine gemäß des Varianz–Kriteriums signifikanten Unterschiede zwischen den Modellen feststellen. Das aus der Eigenwert–Sortierung resultierende Modell für $\mathcal{F} = 16$ weist einen signifikant höheren Wert für die Fitnessfunktion η auf. Dies äußert sich auch im Vergleich zu den übrigen Modellen in den hohen Analysefehlern des Kalibrationsdatensatzes (SEE) und des internen Validationdatensatzes (RMSEP_{V1}) (vgl. Tab. 6.4). Dieses Kalibrationsmodell erfaßt offensichtlich die Varianz im Datensatz nur unzureichend und stellt daher keine brauchbare Lösung dar.

Für die Entscheidung, welche die beste unter den verbleibenden Kalibrationen ist, wird letztendlich dasjenige Modell ausgewählt, welches die kleinste Anzahl an Faktoren und damit den kleinsten Anteil spektroskopischer Information beinhaltet, der zur Berechnung einer robusten Kalibration notwendig ist. Dies ist das Kalibrationsmodell mit neun Faktoren (1,2,5,6,9,12,14,16,18).

Das Modell erreicht eine Datenkompressionsrate ($\mathfrak{D}$) von 82 % und repräsentiert die spektrale Varianz ($\mathfrak{I}$) der Kalibrationsspektren zu über 99 %.

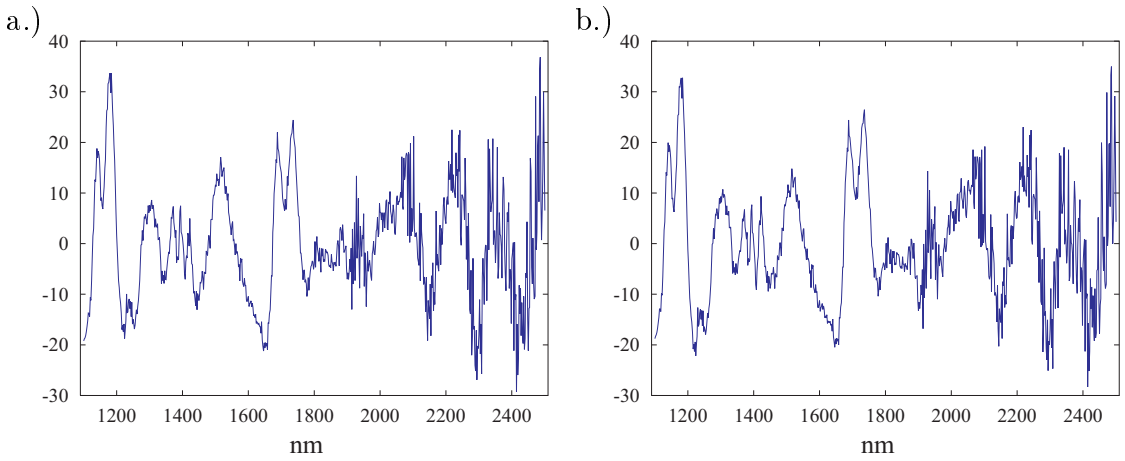

Abbildung 6.13: Property–Weighting–Spektren für den Proteingehalt von Weizen des 9–Faktor–Modells a.) und des 12–Faktor–Modells b.).

Ein Vergleich der Property–Weighting–Spektren (PWS, *s.* Abb. 6.13) stützt dieses Ergebnis. Die Kalibrationsmodelle mit $\mathcal{F} = 9$ und $\mathcal{F} = 18$ weisen praktisch keine Unterschiede im PWS auf. Auch die Empfindlichkeiten, die sich aus der Auftragung des NAS gegen die Proben–Konzentration ergeben (*s.* Abb. 6.14), sind nahezu identisch. Die Berücksichtigung der Faktoren 7, 8 und 13 im 12–Faktor–Modell bringt gegenüber dem 9–Faktor–Modell keine Vorteile. Diese Faktoren beinhalten Information, die nicht zur Verbesserung der Kalibration führt. Zusammenfassend ist daher das 9–Faktor–Modell als das beste Kalibrationsmodell anzusehen.

Vergleich zu Partial Least Squares Berechnungen

In Abbildung 6.15 sind die Standardabweichungen des Kalibrationsdatensatzes und der beiden Validationdatensätze einer PLS–Rechnung dargestellt. Die schraffierte Fläche stellt die Varianzgrenzen des Analysenfehlers (RMSEP_{V2}) für das oben diskutierte 9–Faktor–Kalibrationsmodell aus der GA basierten Optimierung dar. Die von der GA basierten PCR und der CPCR erzielte Lösung unterscheidet sich unter Berücksichtigung dieses Varianz–Kriteriums nicht signifikant von einem 11–Faktor PLS–Modell. Auch hier führt die weitere Hinzunahme von PLS–Faktoren in das Kalibrationsmodell zu Overfitting, das heißt der Vorhersagefehler des Kalibrationsdatensatzes nimmt weiter erheblich ab, wohingegen die Fehler in beiden Validationsdatensätzen einen zunehmenden Trend ausweisen.

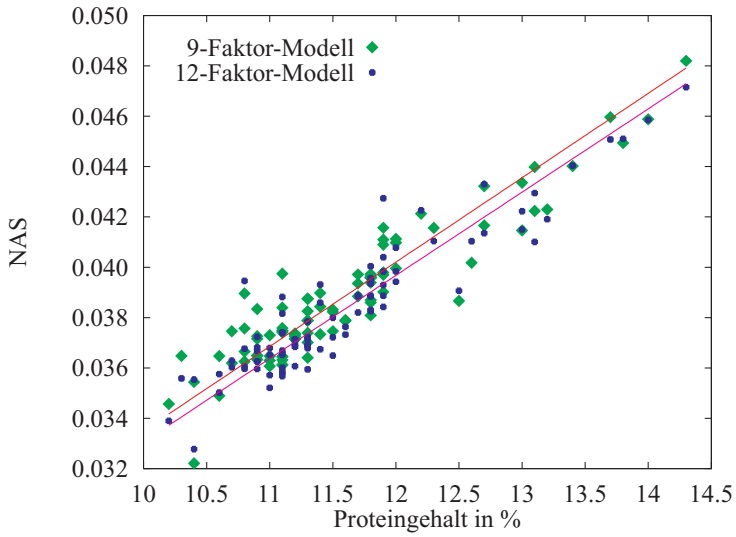

Abbildung 6.14: Net Analyte Signal für die Kalibration von Protein in Weizen.

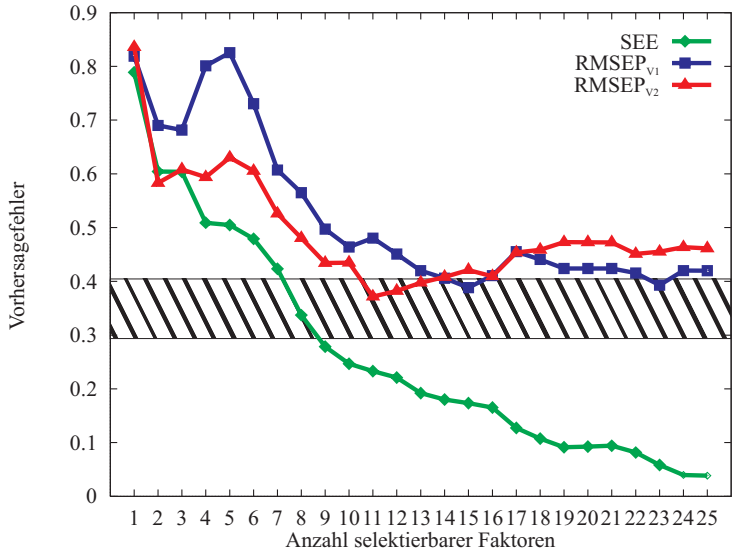

Abbildung 6.15: Kalibrationsergebnisse der PLS–Berechnung für den Proteinanteil in Weizen. Die schraffierte Fläche stellt die Varianzgrenzen des Analysenfehlers (RMSEP_{V2}) im 9–Faktor–Modell auf GA–PCR Basis dar.

In diesem Zusammenhang ist bemerkenswert, daß häufig empfohlen wird, die Anzahl berücksichtigter PLS–Faktoren analog zur Top–Down Variable Selection zu bestimmen (*s.* 4.1.1). Als Kriterium wird in diesen Fällen das erste Minimum des PRESS–Werts (*s.* 3.5.1 ff.) oder der Standardabweichung eines Validationsdatensatzes herangezogen [20]. Für die Kalibration des Proteingehaltes würde dies zu einem deutlichen Underfitting führen, da nur die ersten zwei oder drei PLS–Faktoren in der Kalibration Berücksichtigung fänden. Die resultierenden PLS–Modelle weisen einen Vorhersagefehler auf, der etwa um den Faktor 2 höher liegt als der Fehler der Modelle, die nach dem bisherigen Stand der Diskussion als optimal anzusehen sind. Hierbei wird mit dem GA–basierten PCR–Modell eine geringere Anzahl von Faktoren (9) benötigt und damit eine höhere Selektivität erzielt als mit dem PLS–Modell (11 Faktoren). Der Vergleich mit PLS–Kalibrationen zeigt auch in diesem Fall die Leistungsfähigkeit der PCR basierten Kalibrationsmethoden.

6.2.2 Kalibration des Olfen–Datensatzes

Dieser Datensatz beinhaltet Transmissionsspektren von Olfen–Tabletten (*s.* 5.2.2). Die Intensität der von einer Tablette noch durchgelassenen NIR–Strahlung ist sehr niedrig, was ein entsprechend schlechtes Signal–Rausch–Verhältnis bedingt. Für den Kalibrationsdatensatz wurden 30 Scans (Meßzeit ca. 1 Minute) zu einem Spektrum akkumuliert. Die Validationsdatensätze wurden mit 10 Scans im Falle des ersten und einem Scan für den zweiten Datensatz vermessen.

Aus Abbildung 6.16 geht hervor, daß die kleinste Standardabweichung für den externen (zweiten) Validationsdatensatz durch ein Kalibrationsmodell erreicht wird, das die ersten vier Faktoren berücksichtigt ($\mathcal{F} = 4$). Die Ergebnisse zu den Modellen sind in Tabelle 6.7 zusammengefaßt. Die Rangfolge der Faktoren nach Korrelationskoeffizienten kann der Tabelle A.9 im Anhang entnommen werden.

Tabelle 6.7: Faktor–Selektion für die Kalibration von Olfen–Tabletten auf Basis unterschiedlicher Methoden (Korrelations–Ranking).

Methode	$\mathcal{F}$	Faktor–Selektion		SEE	RMSEP_{V1}	RMSEP_{V2}
		n_f	Selektion			
SV	25	5	1–5	1.732	4.090	3.859
GA / CPCR	4	4	1–4	2.608	4.299	2.896
GA / CPCR	5	5	1–5	1.732	4.090	3.859

Das erste stabile Plateau der Fitnessfunktion wird für $5 \leq \mathcal{F} \leq 7$ erreicht. Daher kann das Kalibrationsmodell mit $\mathcal{F} = 5$ (dies entspricht einem 5–Faktor–Modell) ebenfalls als eine mögliche gute Lösung in Betracht gezogen werden. Darüber hinaus

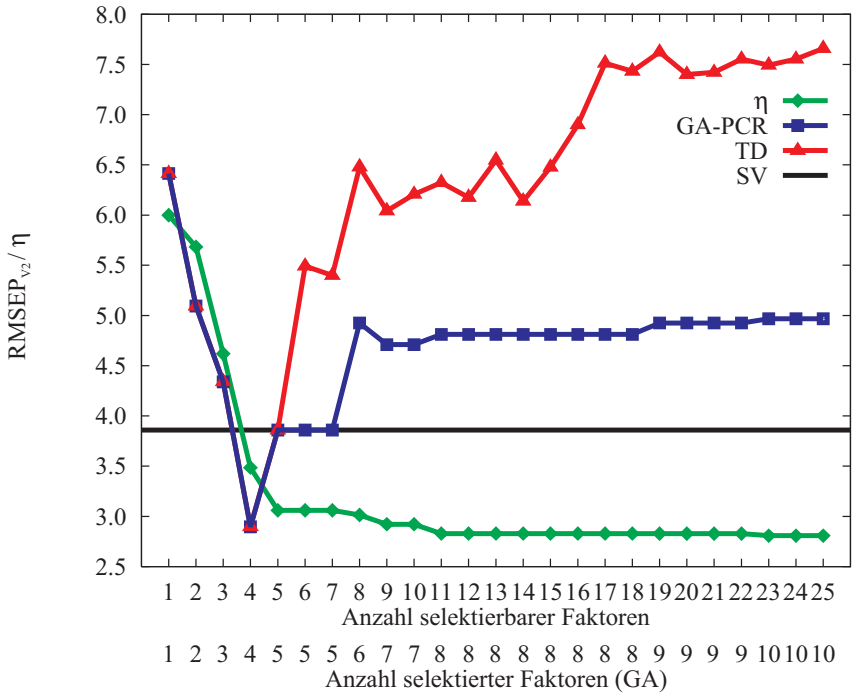

Abbildung 6.16: Ergebnisse der externen Validation ($RMSEP_{V2}$) in Abhängigkeit von der Anzahl der Freiheitsgrade für den Wirkstoffgehalt in Olfen–Tabletten. η = SEEP ist die Fitnessfunktion. Eine Sortierung nach fallender Korrelation wird eingesetzt.

ist keine weitere Verbesserung der Fitnessfunktion (SEEP) zu beobachten, die im Sinne des Varianz–Kriteriums (6.7) als signifikant zu bezeichnen wäre.

Aus Abbildung 6.16 und Tabelle 6.7 wird jedoch deutlich, daß die Hinzunahme des fünften Faktors zu einem Overfitting–Effekt im externen Validationsdatensatz führt ($RMSEP_{V2}$). Unter Berücksichtigung des Varianz–Kriteriums für diese Standardabweichung ist das 4–Faktor–Modell besser als das 5–Faktor–Modell.

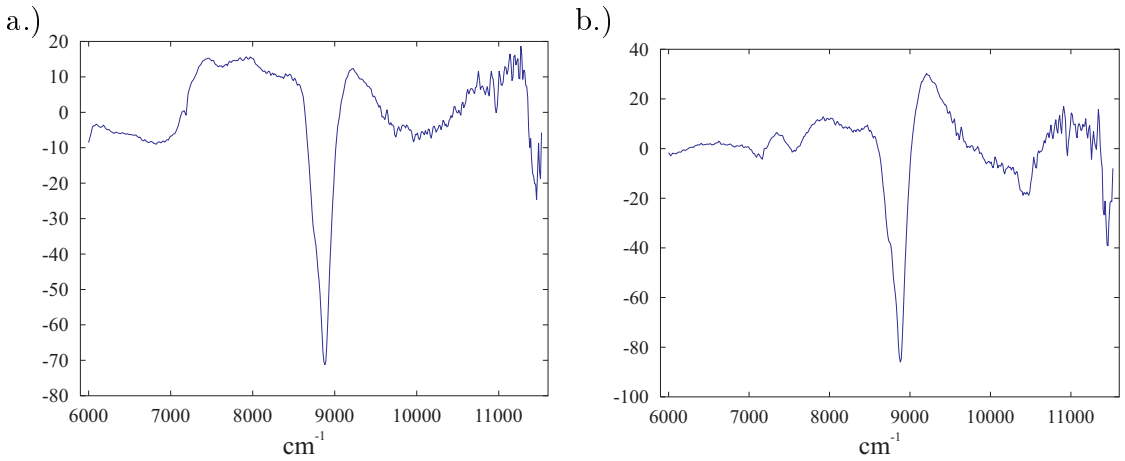

Abbildung 6.17: Property–Weighting–Spektren für den Diclofenac–Gehalt in Olfen–Tabletten des 4–Faktor–Modells a.) und des 5–Faktor–Modells b.).

Abbildung 6.17 zeigt die Property–Weighting–Spektren für beide Modelle, wobei im Falle des 5–Faktor–Modells das Rauschen etwas höher ist. Aus Abbildung 6.18 ist ersichtlich, daß die Empfindlichkeit des 4–Faktor–Modells etwas höher ist als die des 5–Faktor–Modells. Beides spricht zusätzlich dafür, daß das 5–Faktor–Modell gegenüber dem 4–Faktor–Modell keine Vorteile bietet.

Vergleich zu Partial Least Squares Berechnungen

Im Vergleich mit der PCR–Kalibration erreichen die in Abbildung 6.19 dargestellten PLS–Modelle nicht die gleiche Leistungsfähigkeit. Das oben angesprochene 4–Faktor–Modell (GA–PCR) ist in bezug auf die Vorhersage des externen Validationsdatensatzes durchgängig besser als alle alternativen PLS–Modelle. An diesem Beispiel zeigt sich deutlich die bessere spektrale Selektivität der Principal Component Regression im Vergleich zur PLS.

Anhand der beschriebenen Beispiele wird deutlich, daß es für das Kalibrationsergebnis einen Unterschied macht, ob die Faktoren bei der PCR nach fallenden Eigenwerten oder Korrelationen sortiert werden. Wenn die Eigenwert–Sortierung

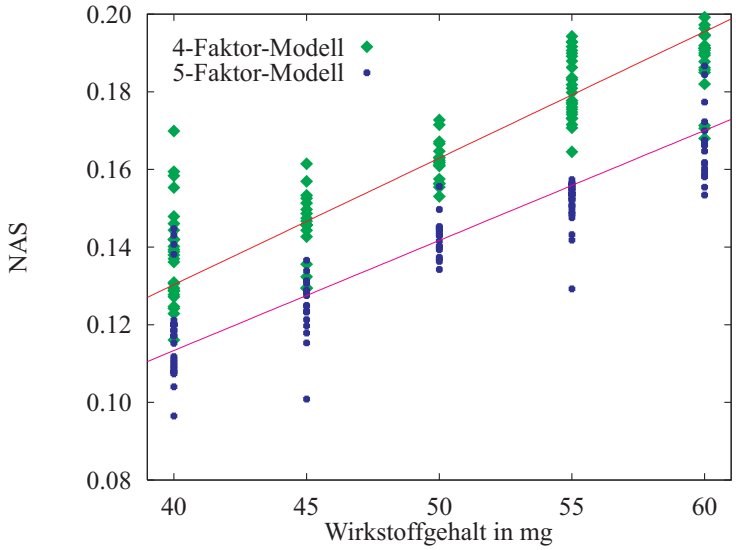

Abbildung 6.18: Net Analyte Signal für die Kalibration des Diclofenac–Gehalts in Olfen–Tabletten.

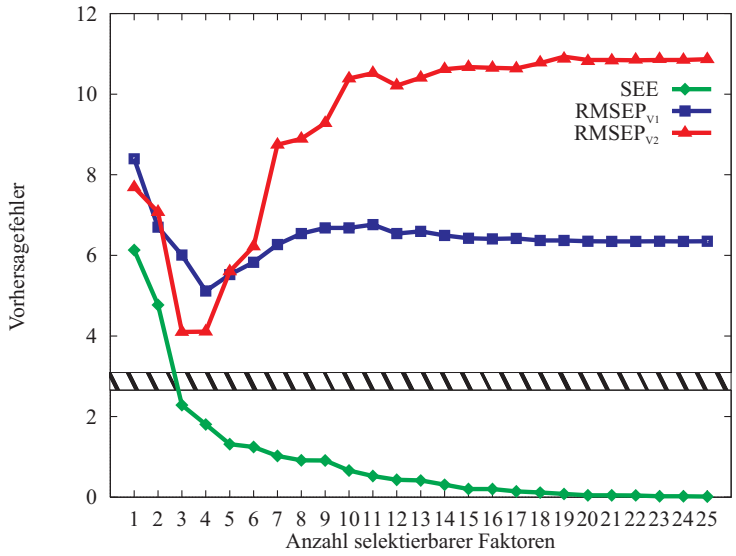

Abbildung 6.19: Kalibrationsergebnisse der PLS–Berechnung für den Diclofenac–Gehalt in Olfen–Tabletten. Die schraffierte Fläche stellt die Varianzgrenzen des Analysenfehlers (RMSEP_{V2}) im 4–Faktor–Modell auf GA–PCR Basis dar.

angewandt wird, steigt mit sinkender Anzahl der Freiheitsgrade die Gefahr, für die Güte der Kalibration wichtige Faktoren außer acht zu lassen (Underfitting). Wird dagegen ohne Berücksichtigung der durch einen Faktor repräsentierten Varianz die Sortierung nach fallenden Korrelationskoeffizienten eingesetzt, dann steigt die Gefahr von Overfitting mit zunehmender Zahl der Freiheitsgrade an.

Aus den Ergebnissen läßt sich ableiten, daß die Verwendung einer bestimmten Sortiermethode zur Optimierung des Kalibrationsmodells einen direkten Einfluß auf die Qualität des Ergebnisses eines Genetischen Algorithmus oder jedes anderen Selektionsverfahrens hat. Wie bereits dargestellt, erfolgt die endgültige Auswahl eines Kalibrationsmodells nach folgenden (in der Reihenfolge ihrer Bedeutung aufgelisteten) Kriterien:

1. möglichst geringer Vorhersagefehler (bezüglich eines möglichst umfassenden Sets unabhängiger Validationsstandards),
2. möglichst geringe Anzahl latenter Variablen / Faktoren,
3. möglichst hohe Empfindlichkeit (zu ermitteln über das NAS) und
4. einigermaßen kontinuierlicher und rauschfreier Verlauf des Property–Weighting–Spektrums.

Bleiben nach Auswertung dieser Kriterien und des Varianz–Kriteriums noch mehrere gleichwertige Kalibrationsmodelle zur Auswahl, so ist die Entscheidung zwischen diesen immer mehr oder weniger subjektiv.

Grundsätzlich sollte man sich immer vor Augen halten, daß Kalibrationsmodelle nur *Modelle* sind, und daher eine absolute Aussage über ihre Qualität unmöglich ist. Das Ziel einer Kalibrationsoptimierung ist es daher, eine brauchbare und zuverlässige Beschreibung der betrachteten Datensätze zu finden, die insgesamt ausgewogen ist und einen Parameter zuverlässig und eindeutig optimiert. «*All models are wrong, but some models are very useful.* », *H. Martens* und *T. Næs* [20].

6.3 Modifizierte Fitnessfunktionen

Zwei Hauptaspekte der Implementierung von Genetischen Algorithmen zur Faktor–Selektion sind zum einen deren Fähigkeit, eine optimale (globale) Lösung für ein kombinatorisches Problem zu finden und zum anderen ihre weitgehende Unabhängigkeit von starren Vorannahmen wie im Fall der Verwendung konventioneller statistischer Verfahren.

Das in den vorangegangenen Abschnitten vorgestellte Verfahren zur Bewertung von Kalibrationsmodellen auf Basis der Varianz von Standardabweichungen stellt zwar eine pragmatische Lösung dar, ist aber im Sinne einer weitergehenden Automatisierung der Faktor–Selektion nur bedingt geeignet. Der Grund liegt darin, daß

zum Zwecke des Vergleichs zwischen den Kalibrationsmodellen für unterschiedliche Freiheitsgrade für jeden Freiheitsgrad ($\mathcal{F}$) ein separates Kalibrationsmodell berechnet werden muß. Für eine Automatisierung der Kalibrationsoptimierung ist es jedoch wünschenswert, daß das optimale Kalibrationsmodell aus einem einzigen Rechengang hervorgeht. Einer der wesentlichsten Punkte ist in diesem Zusammenhang die Bewertung des *Informationsgehaltes* der Faktoren, die in ein Kalibrationsmodell Eingang finden.

Anfang 1998 veröffentlichten *A.S. Barros* und *D.N. Rutledge* eine Arbeit, die sich mit der Anwendung von Genetischen Algorithmen zur Faktor–Selektion befaßt [41]. Die von den Autoren vorgeschlagene Fitnessfunktion berücksichtigt neben dem Vorhersagefehler der Validation auch die Anzahl eingesetzter Faktoren und ist damit den in dieser Arbeit verwendeten Fitnessfunktionen (SEE, RMSEP, SEEP) sehr ähnlich. Darüber hinaus verwenden Barros und Rutledge ein statistisches Kriterium, das durch *Von Neumann* 1941 entwickelt und von *Durbin* und *Watson* 1951 zum Testen von Regressions-Residuen eingesetzt wurde. Das Verfahren wird in der Literatur allgemein als *Durbin–Watson–Test* bezeichnet [101, 102] und zum Test auf Abweichungen einzelner Datenpunkte von linearen Regressionsmodellen verwendet.

6.3.1 Durbin–Watson–Test

Dieses Testverfahren kann herangezogen werden um zu überprüfen, ob Residuen eines linearen Modells einem autoregressiven Prozeß unterliegen, d.h. ob die Residuen einen systematischen Trend beinhalten. Für die Residuen aufeinanderfolgender Punkte x_i zu ihrem Mittelwert $\bar{x}$ gilt

$$e_i = x_i - \bar{x} \quad . \tag{6.8}$$

Zur Prüfung auf Autoregression können die Parameter ρ und u_i zweier aufeinanderfolgender Residuen ermittelt werden. Für sie gilt

$$e_i = \rho \cdot e_{i-1} + u_i \quad . \tag{6.9}$$

Der Parameter ρ bezieht sich auf den kompletten Datensatz. Je größer der ρ–Wert, desto höher ist der Grad der Autoregression. Je größer der Wert der normalverteilten Zufallsvariablen u_i ist, desto höher ist der Anteil an Rauschen (normalverteilte Residuen). Der Durbin–Watson–Testwert errechnet sich nach folgender Formel[4]:

$$W = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \quad . \tag{6.10}$$

[4] *Barros* und *Rutledge* verwenden in ihrer Fitnessfunktion den Ausdruck:
$W = \sum_{i=2}^{n}(e_i - e_{i-1})^2 / \sum_{i=2}^{n} e_i^2$.

Der Test ermöglicht eine Aussage darüber, inwieweit die Residuen eines linearen Modells normalverteilt sind oder einem Trend unterliegen.

Dieses Testverfahren läßt sich auch auf Faktoren aus einer PCA anwenden. Für die einzelnen Faktoren läßt sich ermitteln, ob in Gleichung (6.9) der Anteil von ρ oder von u_i dominiert. Ist u_i die dominierende Größe, so unterliegen die Differenzen aufeinanderfolgender Wertepaare des Faktors (d.h. x_{i-1}, x_i) einer Normalverteilung: Der Faktor repräsentiert überwiegend Rauschen. Dominiert umgekehrt ρ, so sind die Residuen nur zu einem geringen Anteil normalverteilt. Je weniger normalverteilt die Residuen sind, desto größer ist der Anteil spektraler Information, die durch den geprüften Faktor repräsentiert wird. Das Ergebnis des Durbin–Watson–Tests kann also als Meßgröße für den Gehalt spektraler Information eines Faktors interpretiert werden.

Die Werte des Durbin–Watson–Tests tendieren gegen Null, wenn eine hohe Autokorrelation zwischen aufeinanderfolgenden Residuen vorliegt (d.h. durch eine ausgeprägte spektrale Struktur des geprüften Faktors ein hoher Informationsgehalt vorliegt). Wenn eine geringe Autokorrelation vorliegt (z.B. bei überwiegendem Rauschen), tendiert W gegen 2.0. Die Werte von ρ und u_i werden für die Bewertung nie explizit berechnet. Die Verwendung des Testwerts W ist genügend aussagekräftig.

In Abbildung 6.20 sind die Werte des Durbin–Watson–Tests der ersten (nach fallenden Eigenwerten sortierten) 25 Faktoren für den Weizen- und den Olfen–Datensatz dargestellt. In beiden Datensätzen sind die Testwerte für die ersten

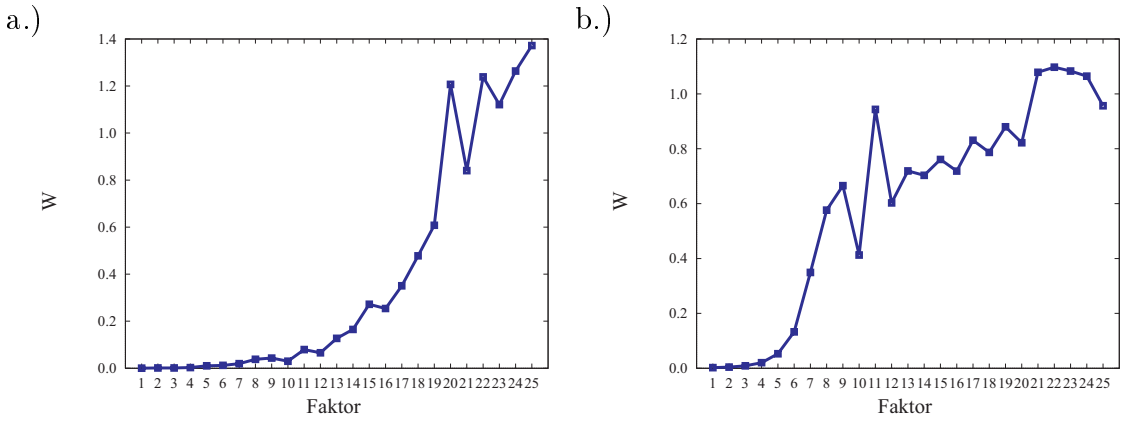

Abbildung 6.20: Durbin–Watson–Tests der ersten 25 Faktoren für den a.) Weizen–Datensatz und b.) Olfen–Datensatz.

Faktoren sehr klein und nehmen dann zu. Dies spiegelt die Tatsache wider, daß der Durbin–Watson–Test einen hohen spektralen Informationsgehalt durch ein niedriges Testergebnis und einen niedrigen Gehalt spektraler Information durch ein hohes Ergebnis ausweist. Der Anstieg in Abbildung 6.20 erklärt sich also dadurch, daß höhere Faktoren einen zunehmenden Anteil von Rauschen beinhalten. Der Anstieg ist aber

nicht streng monoton. Je kleiner der Eigenwert eines Faktors ist, desto kleiner ist zwar die durch ihn wiedergegebene spektrale Varianz, nicht unbedingt aber sein Informationsgehalt. Der unstetige Verlauf der Kurven in Abbildung 6.20 zeigt daher sehr schön, daß Informationsgehalt und Eigenwert zwar grob korrelieren, aber im einzelnen nicht direkt voneinander abhängig sind.

Die festen Intervallgrenzen 0 und 2 machen die Verwendung des Durbin–Watson–Tests, im Gegensatz zu den nach oben und unten offenen Eigenwerten, in einem GA einfach und ermöglichen eine Erweiterung der bisher in dieser Arbeit verwendeten Fitnessfunktionen.

6.3.2 Fitnessfunktionen

Die von *Barros* und *Rutledge* verwendete Fitnessfunktion (im folgenden mit B&R–Fitnessfunktion abgekürzt) berücksichtigt als Kriterien für die Güte einer Kalibration

- den Vorhersagefehler des Validationsdatensatzes (in diesem Fall der *Predicted Residual Sum of Squares* (PRESS–Wert)),
- die Anzahl berücksichtigter latenter Variablen (n_f), d.h. Faktoren und
- den Durbin–Watson–Testwert (W).

Als Durbin–Watson–Testwert wird das Ergebnis aus Gleichung (6.10) eingesetzt, das zum Faktor mit dem kleinsten Eigenwert in einer Selektion gehört (W_{max})[5]. Die B&R–Fitnessfunktion ist definiert als:

$$\eta = \frac{n_f \times W_{max} \times \mathrm{PRESS}}{2.0 - W_{max}} \quad . \tag{6.11}$$

Die Fitnessfunktion ist auf eine Minimierung ausgelegt. Das heißt, je kleiner η wird, desto besser ist das gefundene Kalibrationsmodell. Der Einfluß dieser Fitnessfunktion auf den Vorhersagefehler von Kalibrations- und Validationsspektren läßt sich im Rahmen der Modellbildung für den Weizen- und den Olfen–Datensatz verdeutlichen. Abbildung 6.21 zeigt den Verlauf der Kalibrationsergebnisse für den Feuchtigkeitsgehalt von Weizen über 25 Freiheitsgrade. Durch den Genetischen Algorithmus werden auf Basis der B&R–Fitnessfunktion die ersten drei Faktoren selektiert. Dieses Kalibrationsmodell entspricht dem zuvor in 6.2.1 als optimal gefundenen Modell. Es zeigen sich keine Overfitting–Effekte bei steigender Anzahl der Freiheitsgrade, d.h. die Verwendung von W als zusätzlichem Kriterium in der Fitnessfunktion führt im Gegensatz zu den früher verwendeten Fitnessfunktionen offensichtlich dazu, daß keine weiteren Faktoren zum Modell hinzugefügt werden. Die Kalibration des Pro-

[5] Der Faktor mit dem kleinsten Eigenwert hat den höchsten Index (d.h. Faktor–Nummer), daher die Bezeichnung *max* (*s.* Tab. A.3).

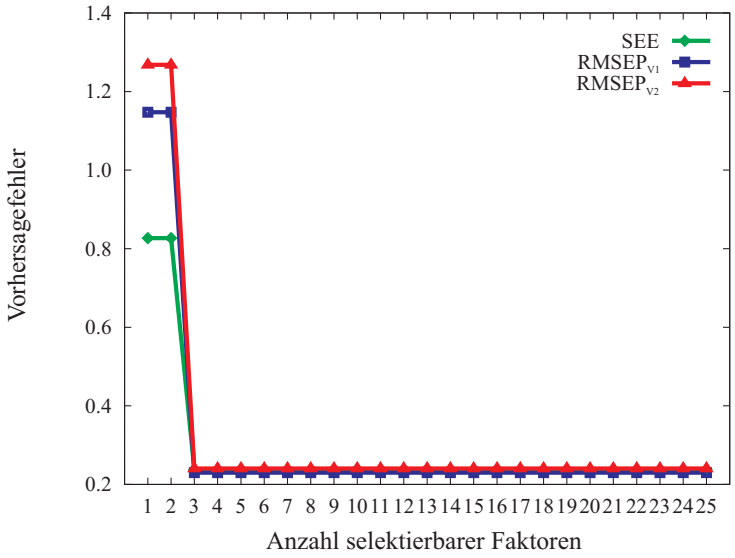

Abbildung 6.21: Optimierung der Kalibration des Feuchtigkeitsgehaltes des Weizen–Datensatzes auf Basis der B&R–Fitnessfunktion. Für $\mathcal{F} \geq 3$ werden die ersten 3 Faktoren selektiert.

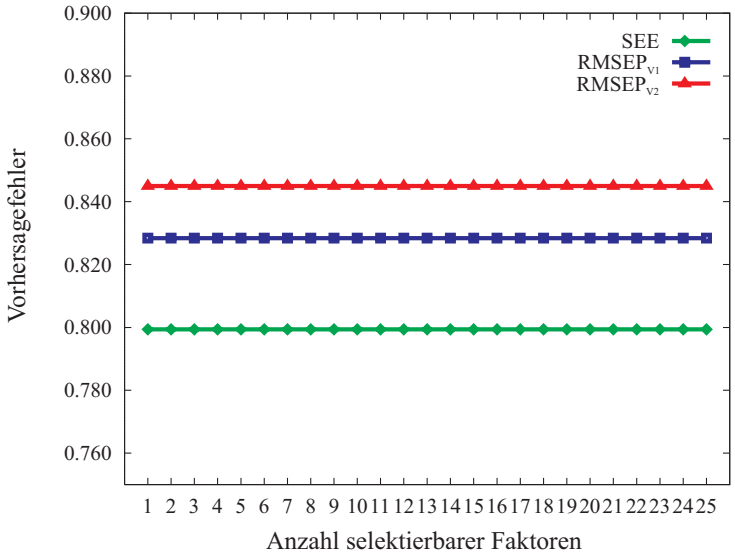

Abbildung 6.22: Optimierung der Kalibration des Proteingehaltes des Weizen–Datensatzes auf Basis der B&R–Fitnessfunktion. Es wird nur ein Faktor selektiert.

teingehaltes liefert dagegen ein mangelhaftes Ergebnis (*s.* Abb. 6.22). Auch bei zunehmender Anzahl der Freiheitsgrade wird nur der erste Faktor selektiert.

Die Kalibration des Olfen–Datensatzes bietet ein analoges Bild. Auch hier wird unabhängig vom Freiheitsgrad nur der erste Faktor im Kalibrationsmodell berücksichtigt. Die Ergebnisse der Vorhersagen sind suboptimal.

Die B&R–Fitnessfunktion ist offensichtlich in Fällen, in denen komplexe Kalibrationsmodelle zur Erfassung einer Eigenschaft notwendig sind oder dann, wenn Störungen durch starkes Rauschen im Datensatz auftreten, zu restriktiv und führen zu Kalibrationsmodellen mit deutlichem Underfitting. Im folgenden wird daher versucht, einen effizienteren Weg zu finden, auf dem der Durbin–Watson–Test in den GA integriert werden kann.

Bereits in den vorangegangenen Abschnitten wurde die Verwendung von Korrelationskoeffizienten bei der Faktor–Selektion beschrieben (*s.* 6.2*ff*). Der Korrelationskoeffizient $r_i(\mathbf{c_i}, \mathbf{p})$[6] gibt einen Hinweis auf die relative Bedeutung von Faktoren im Regressionsmodell. Bei großen Datensätzen ist es jedoch oft der Fall, daß sich vor allem höhere Faktoren nur geringfügig in den Korrelationskoeffizienten unterscheiden, und somit eine Differenzierung auf diesem Weg problematisch wird. Die Kombination mit dem Durbin–Watson–Testverfahren bietet hier die Möglichkeit einer Lösung. Neben der Korrelation eines Faktors mit der zu kalibrierenden Probeneigenschaft kann zusätzlich dessen spektraler Informationsgehalt durch den Durbin–Watson–Test ermittelt werden. Bei der Bewertung der Eignung von Faktoren für ein Kalibrationsmodell ergänzen sich Korrelationskoeffizient und Durbin–Watson–Test.

Ziel war es daher, die beiden statistischen Größen Korrelationskoeffizient und Durbin–Watson–Test in die bisherigen Fitnessfunktionen zu integrieren. Hauptziel war hierbei, die bei der Verwendung der bisherigen Fitnessfunktionen für die Optimierung der Kalibrationsmodelle erforderliche hohe Zahl von GA–Durchläufen (für unterschiedliche Freiheitsgrade $\mathcal{F}$) auf nur einen einzigen mit hohem Freiheitsgrad $\mathcal{F}$ zu reduzieren. Darüberhinaus sollten störende spektrale Effekte (insbesondere Rauschen) keine negativen Auswirkungen auf die Modellierung mit sich bringen. Nach einer Reihe empirischer Tests stellten sich zwei Fitnessfunktionen als besonders gut geeignet heraus. Beide Funktionen werden im folgenden in ihrem Aufbau und ihrer Wirkungsweise beschrieben.

Ausgehend von einer der bisher verwendeten Fitnessfunktionen η (SEE, RMSEP, SEEP) lassen sich die beiden neuen Funktionen wie folgt definieren:

$$\eta_a = \frac{\eta}{|2 - W_{max}| \cdot (|r_{max}| + 1)} \quad , \tag{6.12}$$

$$\eta_b = \frac{\eta}{\sqrt{|2 - W_{min}| \cdot (2 \cdot \frac{|r_{min}|}{||r||})}} \quad . \tag{6.13}$$

[6] $\mathbf{c_i}$ Eigenvektor i, $\mathbf{p}$ Vektor der Eigenschaftswerte der Kalibrationsstandards.

Dabei steht $|r_{max}|$ für den Betrag des Korrelationskoeffizienten, der zum höchsten Faktor[7] und $|r_{min}|$ für den Betrag des Korrelationskoeffizienten, der zum niedrigsten Faktor[8] in einer Selektion gehört. Der Wert $||r||$ repräsentiert die Euklidische Norm des Vekors r für die Korrelationskoeffizienten aller Faktoren:

$$||r|| = \sqrt{r^T r} \quad . \tag{6.14}$$

In Gleichung (6.13) steht $\frac{|r_{min}|}{||r||}$ damit für den zwischen 0 und 1 normierten Korrelationskoeffizienten, der zum Faktor mit dem höchsten Eigenwert (kleinstem Index) in einer Selektion gehört.

Setzt man für η die SEEP Funktion ein, dann ergeben sich SEEP_a und SEEP_b zu

$$\begin{aligned} \text{SEEP}_a &= \frac{\text{SEEP}}{|2 - W_{max}| \cdot (|r_{max}| + 1)} \\ &= \sqrt{\frac{\sum_{i=1}^{n_s}(p_i - \hat{p}_i)^2 + \sum_{i=n_s+1}^{n_s+n_e}(p_i - \hat{p}_i)^2}{(n_s + n_e - n_f - 1) \cdot (|2 - W_{max}| \cdot (|r_{max}| + 1))^2}} \quad , \end{aligned} \tag{6.15}$$

$$\begin{aligned} \text{SEEP}_b &= \frac{\text{SEEP}}{\sqrt{|2 - W_{min}| \cdot 2 \cdot \frac{|r_{min}|}{||r||}}} \\ &= \sqrt{\frac{\sum_{i=1}^{n_s}(p_i - \hat{p}_i)^2 + \sum_{i=n_s+1}^{n_s+n_e}(p_i - \hat{p}_i)^2}{(n_s + n_e - n_f - 1) \cdot |2 - W_{min}| \cdot 2 \cdot \frac{|r_{min}|}{||r||}}} \quad . \end{aligned} \tag{6.16}$$

Prinzipiell ist es möglich, daß die Nenner in Gleichung 6.15 und 6.16 den Wert Null annehmen, und somit kein gültiges Ergebnis aus der Berechnung resultiert. In dem Computer–Programm *VG* wird dieses Problem dadurch abgefangen, daß der minimal zulässige Wert des Nenners für Gleichung (6.15) auf 10^{-7} und für Gleichung (6.16) auf $2 \cdot 10^{-14}$ gesetzt wird.

Am Beispiel des Weizen- und des Olfen-Datensatzes lassen sich die Auswirkungen dieser beiden Fitnessfunktionen auf den Kalibrationsfehler gut verdeutlichen. Während unter Verwendung des SEEP-Wertes als Fitnessfunktion mit zunehmender Anzahl der Freiheitsgrade auch die Zahl selektierter Faktoren sukzessive ansteigt (vgl. Abb. 6.23 und Abb. 6.26), resultieren aus den beiden neuen Fitnessfunktionen

[7] Faktor mit dem kleinsten Eigenwert.
[8] Faktor mit dem größten Eigenwert.

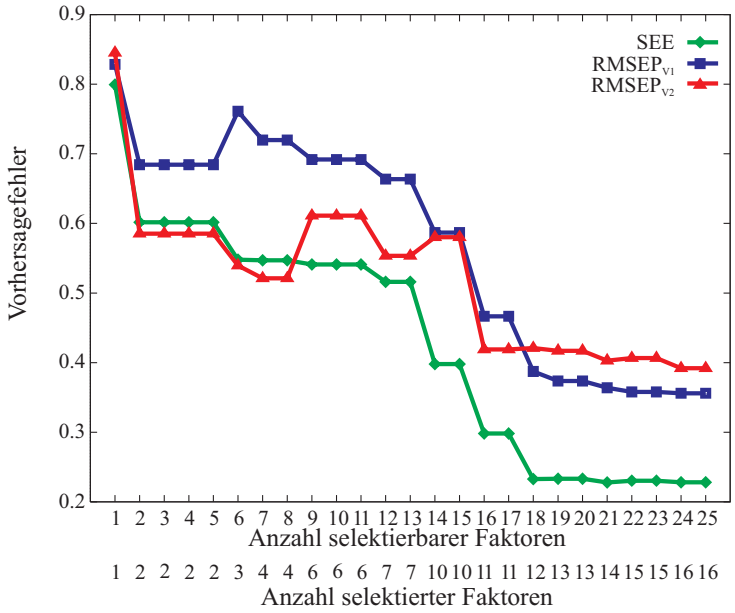

Abbildung 6.23: Ergebnisse der Protein–Kalibration des Weizen–Datensatzes. Fitnessfunktion: SEEP.

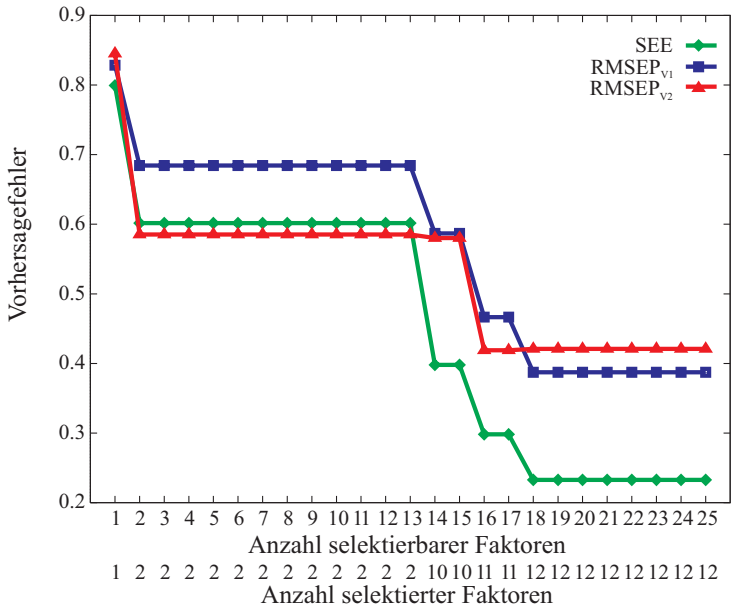

Abbildung 6.24: Ergebnisse der Protein–Kalibration des Weizen–Datensatzes. Fitnessfunktion: SEEP$_a$.

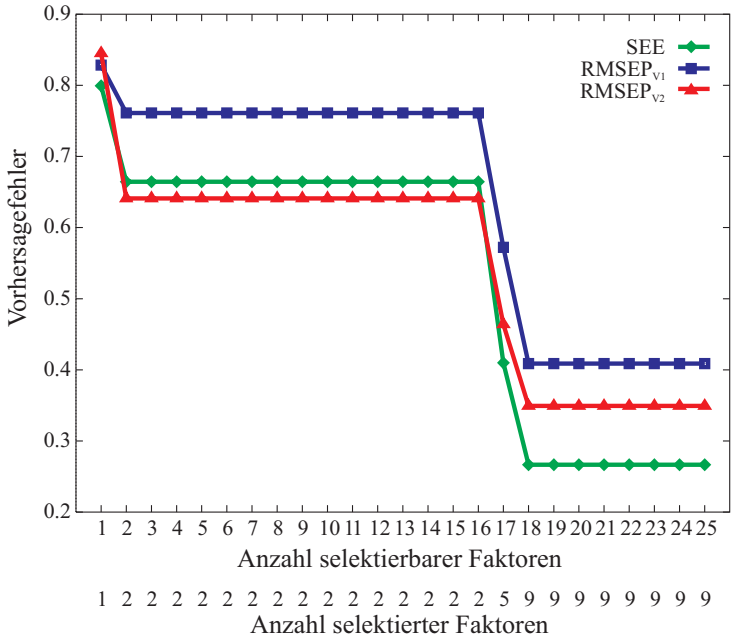

Abbildung 6.25: Ergebnisse der Protein–Kalibration des Weizen–Datensatzes. Fitnessfunktion: SEEP$_b$.

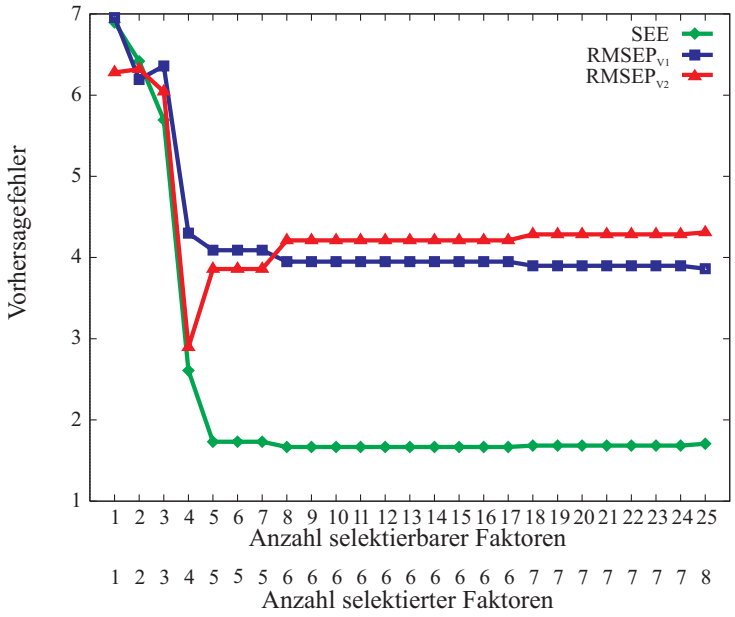

Abbildung 6.26: Kalibrationsergebnisse des Olfen–Datensatzes. Fitnessfunktion: SEEP.

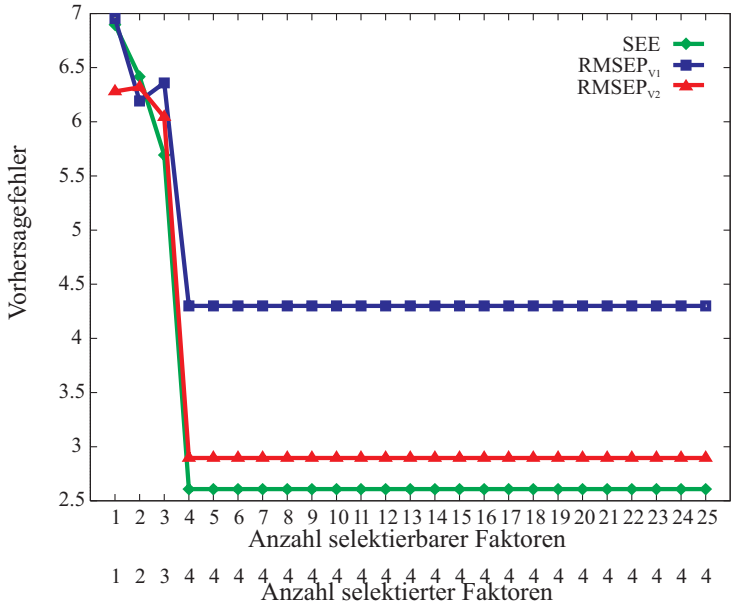

Abbildung 6.27: Kalibrationsergebnisse des Olfen–Datensatzes. Fitnessfunktion: SEEP$_a$.

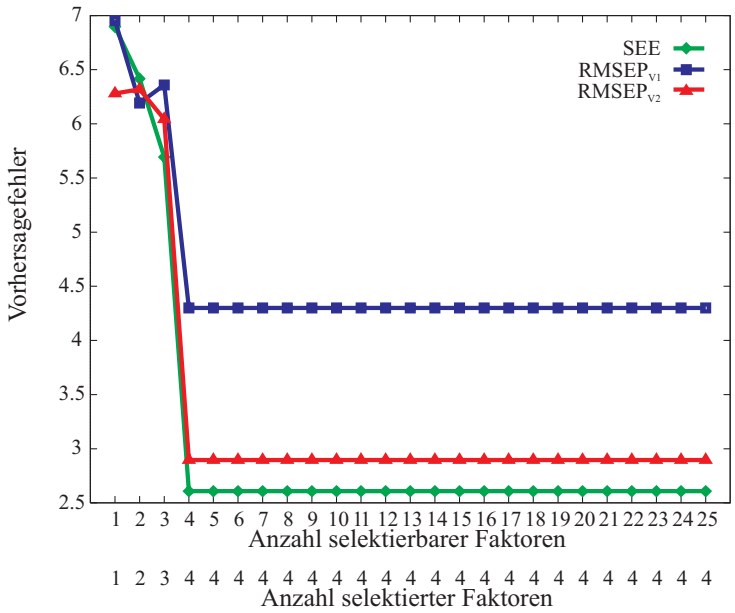

Abbildung 6.28: Kalibrationsergebnisse des Olfen–Datensatzes. Fitnessfunktion: SEEP$_b$.

nur noch wenige Faktorkombinationen auf stabilen Plateaus (Abb. 6.24, 6.25 und Abb. 6.27, 6.28.).

Die Tabellen 6.8 und 6.9 listen die Ergebnisse der Kalibrationsoptimierungen für den Weizen- und den Olfen–Datensatz auf. Die neuen Fitnessfunktionen führen zwanglos zur Selektion derjenigen Faktorkombinationen, die schon in den vorhergehenden Abschnitten als optimal ausgewählt wurden.

Tabelle 6.8: Überblick über die Faktor–Selektionen, die aus der Fitnessfunktion SEEP_a resultieren.

Eigenschaft	$\mathcal{F}$	n_f[a]	Faktoren	SEE	RMSEP_{V1}	RMSEP_{V2}
Weizen						
Feuchte	≥ 3	3	1–3	0.238	0.230	0.240
Protein	≥ 18	12	1,2,5–9,12–14,16,18	0.233	0.387	0.421
Olfen						
Diclofenac	≥ 4	4	1–4	2.608	4.299	2.896

[a] n_f steht für die Anzahl (oder Nummer) selektierter Faktoren

Tabelle 6.9: Überblick über die Faktor–Selektionen, die aus der Fitnessfunktion SEEP_b resultieren.

Eigenschaft	$\mathcal{F}$	n_f	Faktoren	SEE	RMSEP_{V1}	RMSEP_{V2}
Weizen						
Feuchte	≥ 3	3	1–3	0.238	0.230	0.240
Protein	≥ 18	9	1,2,5,6,9,12,14,16,18	0.267	0.409	0.349
Olfen						
Diclofenac	≥ 4	4	1–4	2.608	4.299	2.896

Der Vorteil der neuen Fitnessfunktionen ist darin zu sehen, daß ein optimales Kalibrationsergebnis (unter Vorgabe einer ausreichend hohen Anzahl der Freiheitsgrade $\mathcal{F}$) mit einer einzigen GA–Berechnung erreicht wird. Die bei Anwendung der alten Fitnessfunktionen notwendige aufwendige Differenzierung zwischen den einzelnen Modellen im Anschluß an einen kompletten Optimierungslauf entfällt vollständig.

Ein weiterer Aspekt bei der Umsetzung dieser neuen Fitnessfunktionen war die Forderung, die Kalibrationen weniger anfällig gegenüber spektralen Störungen zu

machen. Dies ist besonders wichtig bei Spektren, die stark verrauscht sind. Je stärker das Rauschen ist, desto größer wird auch sein Anteil an der Gesamtvarianz des Spektrums und desto mehr findet es sich in den niedrigen Faktoren aus einer PCA wieder. Grundsätzlich sollte in eine Kalibration so wenig Rauschen wie möglich eingebracht werden. Da der Durbin–Watson–Test die spektrale Signifikanz eines Faktors bewertet, ist damit zu rechnen, daß die neuen Fitnessfunktionen, die diesen Test beinhalten, auch bei Kalibrationen mit stark verrauschten Spektren zu einer adäquaten Faktor–Selektion führen. Die Ergebnisse einer entsprechenden Überprüfung am Weizen–Datensatz zeigt Abbildung 6.29. Zu den Spektren des Datensatzes wurde

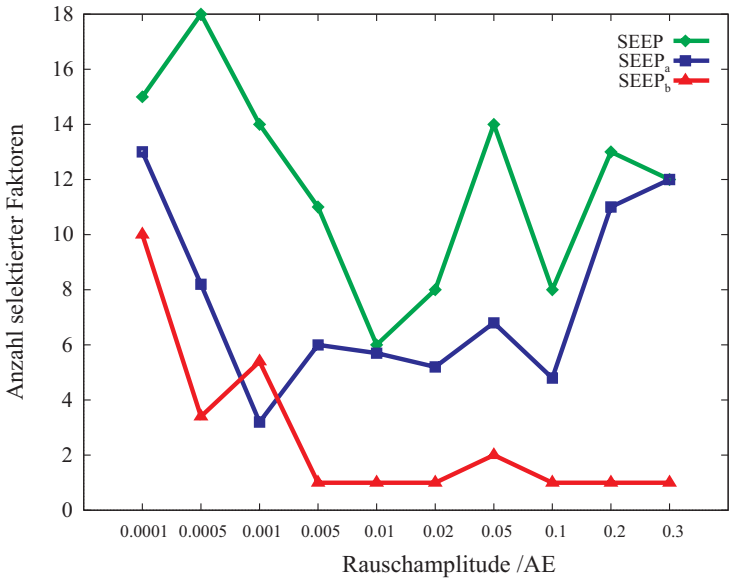

Abbildung 6.29: Anzahl selektierter Faktoren für Protein in Weizen mit verschiedenen Fitnessfunktionen bei 25 Freiheitsgrade. (Durchschnittswert aus zehn Berechnungen).

in unterschiedlicher Intensität Rauschen addiert, um spektrale Störungen zu simulieren. Die zehn gewählten Rauschamplituden (Tab. 6.10) reichen von nahezu keiner Störung (A) mit max. ±0.0001 Absorptionseinheiten (AE) bis zu extremer Beeinflussung der Spektren (J) mit max. ±0.3 AE. Am Beispiel der Protein–Kalibration wird deutlich, daß die Fitnessfunktionen SEEP und SEEP$_a$ bei starken Störungen dazu tendieren, viele Faktoren zu selektieren, wohingegen die Fitnessfunktion SEEP$_b$ dazu neigt, weniger Faktoren in der Kalibration zu berücksichtigen. Die SEEP$_b$–Funktion liefert die geforderte Sicherheit, d.h. Störungsunempfindlichkeit gegenüber Rauschen und robuste Kalibrationsmodelle.

Tabelle 6.10: Verwendete Rauschamplituden.

	Rauschamplitude / AE
A	± 0.0001
B	± 0.0005
C	± 0.001
D	± 0.005
E	± 0.01
F	± 0.02
G	± 0.05
H	± 0.1
I	± 0.2
J	± 0.3

AE = Absorptionseinheiten