4 Chemometrische Grundlagen

Computer sind in der Lage, große Datenmengen effizient und schnell zu verarbeiten. Viele Verarbeitungsschritte können mit ihrer Hilfe automatisiert werden. Daher eignen sich Computer zur Durchführung von wissenschaftlichen Berechnungen, wenn eine große Anzahl an Messwerten ausgewertet werden soll oder eine Auswertung nach einer komplexen Rechenvorschrift durchgeführt werden muss. Um elektronische Rechensysteme effizient einsetzen zu können, sind umfangreiche Kenntnisse über die eingesetzten mathematischen Verfahren notwendig. Dieses Wissen über mathematische Algorithmen ist insbesondere dann wichtig, wenn für die Datenverarbeitung keine fertig entwickelte Software zur Verfügung steht, sondern diese erst noch in Eigenarbeit erstellt werden muss.

4.1 Messfehler

4.1.1 Statistische Fehler

In der Analytik wird bei der Erfassung von Messwerten zwischen systematischen und statistischen Fehlern unterschieden. Systematische Fehler sind Fehler, die auf der falschen Erfassung und Auswertung von Messdaten beruhen. Statistische Fehler haben ihre Ursache in der mangelnden Präzision einer Messwerterfassung auf Grund von Rauschen und anderen Ursachen, die zur Verfälschung der einzelnen Messwerte nach dem Zufallsprinzip führen. Die Streuung von Messwerten wird als Standardabweichung oder ein Vielfaches der Standardabweichung angegeben. Da in der Analytik nicht unendlich viele Wiederholungsmessungen zur Verfügung stehen, muss man sich hierbei mit dem Schätzwert der Standardabweichung begnügen. Es gilt:

$$\Delta u = \pm t_{f,p} \cdot s \qquad (4.1\text{-}1)$$

mit

Δu = Vertrauensintervall = statistischer Fehler der Messgröße u,

$t_{f,p}$ = Student - Faktor für f Freiheitsgrade und ein Wahrscheinlichkeits - niveau von p und

s = Schätzwert der Standardabweichung.

Im Student-Faktor $t_{f,p}$ (siehe Tabelle 4.1-1) ist über seine Abhängigkeit von den Freiheitsgraden f (= Anzahl unabhängiger Variablen, über die u ermittelt worden ist) berücksichtigt, dass die Schätzung der Standardabweichung um so schlechter ist, je weniger Wiederholungsmessungen vorgenommen worden sind. Des weiteren hängt der Student-Faktor von p, dem Wahrscheinlichkeitsniveau, ab. Man unterscheidet Student-Faktoren für einseitige und zweiseitige Fragestellungen. Die zweiseitige Fragestellung lautet: Wie hoch soll die Wahr-

scheinlichkeit mindestens sein, dass das Ergebnis einer später durchgeführten Wiederholungsmessung innerhalb der Grenzen des Fehlerintervalls liegt? Die einseitige Fragestellung lautet entweder, wie hoch soll die Wahrscheinlichkeit mindestens sein, dass das Ergebnis einer später durchgeführten Wiederholungsmessung kleiner als die obere Grenze des Fehlerintervalls ist, oder wie hoch soll die Wahrscheinlichkeit mindestens sein, dass das Ergebnis einer später durchgeführten Wiederholungsmessung größer als die untere der Grenze des Fehlerintervalls ist? Wegen der Symmetrie der Gauß-Verteilung sind die Student-Faktoren der beiden einseitigen Fragestellungen gleich. p ist die für die entsprechende Fragestellung geforderte Wahrscheinlichkeit.

Das Wahrscheinlichkeitsniveau, mit dem statistische Fehler berechnet werden, wird oft mit dem Signifikanzniveau verknüpft. Werte die mit $p = 95\%$ berechnet worden sind, werden dem Signifikanzniveau **wahrscheinlich** zugeordnet, die mit $p = 99\%$ ermittelt worden sind, dem Signifikanzniveau **signifikant**. Ein Wert unterscheidet sich von u wahrscheinlich, wenn er mehr als der mit $p = 95\%$ berechnete Δu-Wert abweicht. Signifikant wird der Unterschied dann, wenn die Abweichung den mit $p = 99\%$ berechneten Δu-Wert übersteigt. **Diese Signifikanzniveaus sind aber frei wählbar, und müssen daher in statistischen Abhandlungen definiert werden, bevor man sie einsetzt**.

Tabelle 4.1-1: Der Studentfaktor $t_{f,p}$ in Abhängigkeit vom Freiheitsgrad f und dem Wahrscheinlichkeitsniveau p [8]

f	$p = 95\%$, einseitig	$p = 95\%$, zweiseitig	$p = 99\%$, einseitig	$p = 99\%$, zweiseitig
1	6,31	12,706	31,821	63,657
2	2,92	4,303	6,965	9,925
3	2,35	3,182	4,541	5,841
4	2,13	2,776	3,747	4,604
5	2,02	2,571	3,365	4,032
6	1,94	2,447	3,143	3,707
7	1,89	2,365	2,998	3,499
8	1,86	2,306	2,896	3,355
9	1,83	2,262	2,821	3,250
10	1,81	2,228	2,764	3,169
15	1,75	2,131	2,602	2,947
20	1,72	2,086	2,528	2,845
25	1,71	2,060	2,485	2,787
30	1,70	2,042	2,457	2,750
∞	1,65	1,960	2,326	2,576

4.1.2 Nachweis-, Erfassungs- und Bestimmungsgrenzen

In der quantitativen Analytik werden Stoffmengen oder mit einer Stoffmenge zusammenhängende Größen wie Konzentrationen, Flüsse oder Massen bestimmt. Diese Größen werden oftmals auf Basis einer fehlerbehafteten Messgröße ermittelt, wobei die Messgröße und die Konzentration einer Substanz im funktionalen Zusammenhang stehen:

$$\begin{aligned} K_{(u)} &= \text{Funktion in Abhängigkeit von der fehlerbehafteten Messgröße } u \\ &= \text{Stoffmenge, Konzentration, Fluss oder Masse einer Substanz.} \end{aligned} \tag{4.1-2}$$

Als Blindwert wird die Größe von u bezeichnet, bei der die Stoffmenge, die Konzentration, der Fluss oder die Masse der untersuchten Substanz Null ist:

$$K_{(u)} = 0 \quad \Rightarrow \quad u = u_{\mathrm{B}} = \text{Blindwert.} \tag{4.1-3}$$

Aus der Streuung von u_{B} kann die sogenannte kritische Größe ermittelt werden.

$$\begin{aligned} u_{\mathrm{K}} &= u_{\mathrm{B}} + \Delta u_{\mathrm{B}} = u_{\mathrm{B}} + t_{f,p} \cdot s_{\mathrm{B}} \\ &\text{mit} \\ u_{\mathrm{K}} &= \text{kritische Größe,} \\ \Delta u_{\mathrm{B}} &= \text{obere Vertrauensintervallgrenze von } u_{\mathrm{B}}, \\ t_{f,p} &= \text{Student - Faktor für } f \text{ Freiheitsgrade und ein Wahrscheinlich -} \\ &\quad \text{keitsniveau von } p \text{ für die einseitige Fragestellung und} \\ s_{\mathrm{B}} &= \text{Schätzwert der Standardabweichung des Blindwertes,} \\ &\text{falls} \\ u_1 > u_2 &\Rightarrow K_{(u_1)} > K_{(u_2)} \text{ (für alle weiteren Betrachtungen gültiger Regelfall).} \end{aligned} \tag{4.1-4}$$

Es ist zu erwarten, dass bei zukünftigen Blindwertmessungen sich nur 100%-p (z. B. 5% oder 1%) der Messwerte oberhalb (Regelfall, ansonsten unterhalb) der kritischen Größe befinden. Nach der Deutschen Norm DIN 32645[1] wird über diese kritische Größe die **Nachweisgrenze** als Konzentration, die der kritischen Messgröße entspricht, definiert. [9] Dieser Sachverhalt lässt sich verallgemeinern:

$$K_{(u_{\mathrm{K}})} = \text{Nachweisgrenze.} \tag{4.1-5}$$

Die **Erfassungsgrenze** ist nach DIN 32645 die Konzentration, bei der sich ein Anteil von p' aller Messwerte oberhalb der Nachweisgrenze befindet. [9] Auch diese Definition lässt sich auf

---

[1] DIN 32645 (herausgegeben vom **D**eutsches **I**nstitut für **N**ormung e. V.) beschreibt das Verfahren zur Bestimmung von Nachweis-, Erfassungs- und Bestimmungsgrenzen für eine Konzentrationsbestimmung, wobei $K_{(\mathrm{u})} = b \cdot \mathrm{u} + a$ mit b = Steigung und a = Achsenabschnitt.

andere aus einer fehlerhaften Messgröße hergeleitete Größen (Stoffmenge, Fluss, Masse etc.) anwenden. Im Regelfall wird das Wahrscheinlichkeitsniveau p' wie p auf 95% oder 99% gesetzt. Unter der Voraussetzung, dass $p' = p$ gilt:

$$K_{(u_K + t_{f,\,p'} \cdot s_B)} = K_{(u_B + 2t_{f,\,p'} \cdot s_B)} = \text{Erfassungsgrenze.} \tag{4.1-6}$$

Beschreibt die Funktion K einen linearen Zusammenhang, so bedeutet dies für die Erfassungsgrenze:

$$K_{(u_K + t_{f,\,p'} \cdot s_B)} = K_{(u_B + 2t_{f,\,p'} \cdot s_B)} = 2 \cdot K_{(u_K)}. \tag{4.1-7}$$

Die **Bestimmungsgrenze** ist laut DIN 32645 die Konzentration, oberhalb der sich der Gehalt einer Substanz mit einer bestimmten relativen Ergebnisunsicherheit bestimmen lässt. Verallgemeinert bedeutet dies:

$$\frac{\Delta K_{(u_{BG})}}{K_{(u_{BG})}} = \text{relative Ergebnisunsicherheit}$$

mit

$K_{(u_{BG})}$ = Bestimmungsgrenze, (4.1-8)

$\Delta K_{(u_{BG})}$ = obere Grenze des Vertrauensintervalls der Konzentration an der Erfassungsgrenze für die zweiseitige Fragestellung und

u_{BG} = Messwert, aus dem die Bestimmungsgrenze berechnet wird.

Die Bestimmungsgrenze muss aus dem statistischen Fehler der fehlerbehafteten Messgröße u_{BG} hergeleitet werden. Diese Herleitung ist abhängig von der Art des funktionellen Zusammenhangs zwischen Konzentration und Messgröße.

4.1.3 Quartile und Dezile

Die Berechung von Vertrauensintervallen beruht auf der Annahme, dass die Fehler, die auf die untersuchte Größe einwirken, statistische Zufallsfehler sind. Solche Zufallsfehler führen dazu, dass die Wahrscheinlichkeitsdichte, einen bestimmten Messwert zu reproduzieren, eine Gaußsche Glockenkurve ergibt:

$$d = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(u-\mu)^2}{2\sigma^2}}$$

mit

σ = Standardabweichung, (4.1-9)

μ = wahrer Wert der fehlerbehafteten Größe und

d = Wahrscheinlichkeitsdichte.

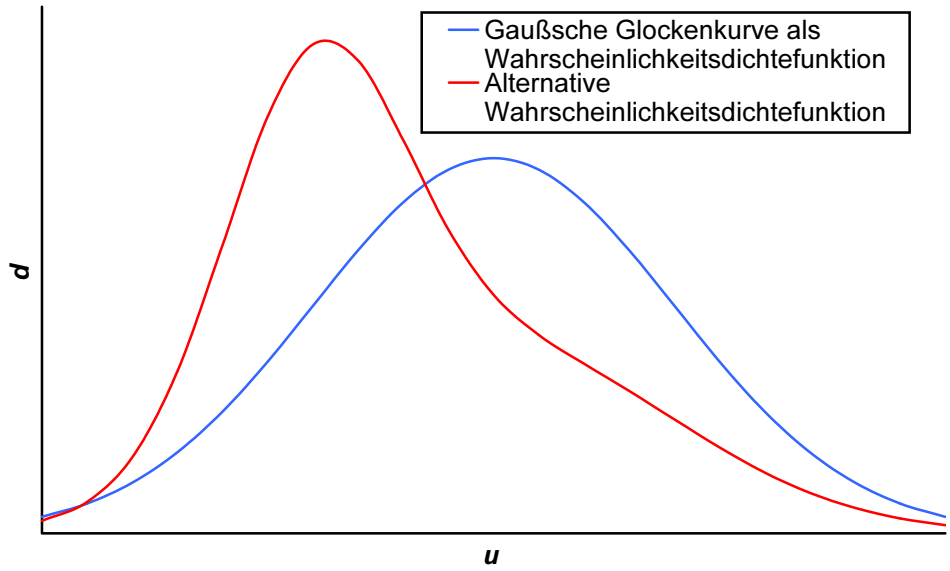

Abbildung 4.1-1: Gaußsche Wahrscheinlichkeitsdichtefunktion

Die Einzelmessungen, auch wenn es endlich viele Messwerte sind, werden in einem solchen Fall als normalverteilt bezeichnet. Spielen andere Fehlerquellen als klassische Zufallsfehler für die Streuung der Messwerte eine Rolle (z. B. wenn sich Zufallsfehler und systematische Fehler überlagern), so folgt die Wahrscheinlichkeitsdichtefunktion nicht mehr der Gauß-Kurve (siehe Abbildung 4.1-1). In einem solchen Fall spiegelt das in Abschnitt 4.1.1 definierte Vertrauensintervall keine verlässliche Fehlerabschätzung wieder. Um dennoch den Fehler einer Messgröße abschätzen zu können, können Quartil- und Interdezilbereiche verwendet werden. Voraussetzung ist, dass die Messgröße durch eine größere Anzahl an Messungen abgeschätzt werden kann. Der Quartilbereich ist das Intervall, im welchen sich 50% aller Messwerte befinden. 25%, also ¼ der Messwerte befinden sich oberhalb dieses Intervalls und ¼ bzw. 25% unterhalb. Beim Interdezilbereich sind die Intervallgrenzen so abgesteckt, dass sich 10% aller Messwerte oberhalb und 10% unterhalb der Grenzen des Intervalls befinden. Statt des Mittelwertes wird hierbei der Median als Abschätzung für den wahren Wert verwendet. Bei den Grenzen des Fehlerintervalls ist man nicht auf Quartile oder Dezile festgelegt. Allgemein kann ein Fehlerintervall folgendermaßen definiert werden:

$$
\begin{aligned}
I_{\mathrm{O},p} &= \text{obere Intervallgrenze und} \\
I_{\mathrm{U},p} &= \text{untere Intervallgrenze, in der sich } p\text{-Prozent aller Messwerte befin-} \\
&\quad \text{den, wobei sich } (100\% - p)/2 \text{ aller Messwerte jeweils oberhalb und} \\
&\quad \text{unterhalb der Grenzen des so definierten Intervalls befinden.} \\
I_p &= I_{\mathrm{O},p} - I_{\mathrm{U},p} = \text{Spannweite des Fehlerintervalls.}
\end{aligned}
\tag{4.1-10}
$$

Der Interdezilbereich wird beispielsweise über die Grenzen $I_{\mathrm{O},\,80\%}$ und $I_{\mathrm{U},\,80\%}$ definiert, wobei die Spannweite $I_{\,80\%}$ beträgt. Die Wahrscheinlichkeit, dass auch zukünftige Messungen mit der Wahrscheinlichkeit von p innerhalb des in (4.1-10) definierten Intervalls liegen, wächst, je größer die Anzahl an Messungen ist, auf der dieses Intervall beruht. Daher entspricht das

Intervall bei unendlich vielen normalverteilten Messwerten dem bezüglich der zweiseitigen Fragestellung ermittelten Vertrauensintervall mit dem entsprechenden Wahrscheinlichkeitsniveau p (siehe Gleichung (4.1-1)). Bei einer hohen Anzahl an Messwerten ist es eine recht gute Schätzung des Vertrauensintervalls.

4.1.4 Die Fehlerfortpflanzung

Oftmals wird ein Analysenergebnis auf der Grundlage mehrerer fehlerbehafteter Größen berechnet, die voneinander unabhängig sind:

$$
\begin{aligned}
u &= f_{(x_1, x_2, \ldots)} \\
&\text{mit} \\
u &= \text{Analysenergebnis und} \\
x_1, x_2, \ldots &= \text{voneinander unabhängige fehlerbehaftete Messgrößen.}
\end{aligned}
\tag{4.1-11}
$$

Kann für die fehlerbehafteten Messgrößen ein maximaler Fehler Δx_1, Δx_2, … angegeben werden, so ist es möglich, den maximalen Fehler Δu des Analysenergebnisses zu nähern:

$$
\Delta u = \left| \left(\frac{\partial u}{\partial x_1} \right)_{x_2, \ldots} \right| \cdot \Delta x_1 + \left| \left(\frac{\partial u}{\partial x_2} \right)_{x_1, \ldots} \right| \cdot \Delta x_2 + \ldots \tag{4.1-12}
$$

Dieses Verfahren zur Berechnung der Fehlerfortpflanzung vernachlässigt, dass sich die Fehler unter Umständen gegenseitig aufheben können. Daher wird es häufig dazu benutzt, die Auswirkung systematischer Fehler auf das Endergebnis abzuschätzen.

Die Standardabweichung eines Analysenergebnisses, welches aus mehreren Messwerten errechnet ist, kann nach dem Gaußschen Fehlerfortpflanzungsgesetz genähert werden:

$$
u = f_{(x_1, x_2, \ldots)} \Rightarrow s_u = \sqrt{\left(\frac{\partial u}{\partial x_1} \right)^2_{x_2, \ldots} \cdot s^2_{x_1} + \left(\frac{\partial u}{\partial x_2} \right)^2_{x_1, \ldots} \cdot s^2_{x_2} + \ldots} \tag{4.1-13}
$$

s_u = Schätzwert der Standardabweichung des Analysenergebnisses.

s_{x_1}, s_{x_2} = Schätzwerte der Standardabweichungen der verschiedenen Messgrößen, aus denen das Analysenergebnis berechnet wird.

Gleichung (4.1-13) darf auch zur Abschätzung des Vertrauensbereiches eines Analysenergebnisses verwendet werden und lautet dann: [2]

$$u = f_{(x_1,\, x_2,\, \ldots)} \Rightarrow \Delta u = \sqrt{\left(\frac{\partial u}{\partial x_1}\right)^2_{x_2,\ldots} \cdot \left(t_{f_1,\, p_1} \cdot s_{x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2_{x_1,\ldots} \cdot \left(t_{f_2,\, p_2} \cdot s_{x_2}\right)^2 + \ldots} \tag{4.1-14}$$

mit

$t_{f_1,\, p_1}$ = Studentfaktor zur Berechung des Vertrauensintervalls von x_1 und

$t_{f_2,\, p_2}$ = Studentfaktor zur Berechung des Vertrauensintervalls von x_2.

4.2 Die Speicherung von Messdaten

Messdaten werden in Computersystemen in sogenannten Variablen gespeichert. Variablen sind reservierte Speicherplätze, in denen Zahlen mit einem vorher definierten Wertebereich Platz finden. Auf einem Computer mit einem zum Intel 386® kompatiblen Prozessor und Intel 387® kompatibler Fließkommaeinheit sind folgende Variablentypen am wichtigsten:

1. Der 32-Bit (2 Byte) große Integertyp, der ganze Zahlen im Bereich von –2147483648 bis 2147483647 repräsentiert.

2. Der 80 Bit (10 Byte) große reelle Typ, der reelle Zahlen in einem Bereich von $\pm 3{,}6 \cdot 10^{-4951}$ bis $1{,}1 \cdot 10^{4932}$ mit 19 bis 20 signifikanten Stellen repräsentiert.

Mit beiden Variablentypen können die genannten Prozessoren direkt, d. h. ohne Konvertierung in einen anderen Variablentyp, rechnen. 4 und 8 Byte große reelle Typen können ohne großen Konvertierungsaufwand in den 10 Byte großen Typ umgewandelt werden. Mit dem 4 Byte großen Datentyp kann auf Kosten des Wertebereiches und der Genauigkeit $\pm 1{,}5 \cdot 10^{45}$ bis $3{,}4 \cdot 10^{38}$ bei 7 bis 8 signifikanten Stellen Speicherplatz gespart werden. Der 8 Byte große Typ wird von der arithmetischen Recheneinheit vieler Prozessortypen bereitgestellt. Daher wird er oft verwendet, wenn Programme auch auf nicht Intel 386/387-kompatiblen Prozessoren laufen sollen. Nachteil ist der gegenüber den 10 Byte großen reellen Typen geringere Wertebereich von $\pm 5{,}0 \cdot 10^{-324}$ bis $1{,}7 \cdot 10^{308}$ und mit 15-16 signifikanten Stellen geringere Genauigkeit. Da Intel 386/387-kompatible Prozessoren grundsätzlich mit dem 10 Byte großen Typ rechnen, bietet die Verwendung kleinerer reeller Typen keinen Vorteil bezüglich der Rechengeschwindigkeit, sondern nur bezüglich des benötigten Speicherplatzes. Nicht vergessen werden darf aber, dass das Kopieren großer Variablen mehr Zeit in Anspruch nimmt, als das Kopieren kleiner Variablen. Die hohe Anzahl an Nachkommastellen und der enorme Wertebereich der 10 Byte großen Fließkommavariable erscheint auf den ersten Blick für viele wissenschaftliche Berechnungen übertrieben. Aber gerade bei größeren mathematischen Berechnungen beispielsweise mit Matrizen, wird bei den

Zwischenergebnissen der Wertebereich der Fließkommavariablen stark beansprucht. Eine hohe Anzahl verfügbarer signifikanter Stellen hat schon oft einen Programmabbruch wegen einer Division durch Null verhindert. Daher ist die Verwendung des 10 Byte großen reellen Typs zu bevorzugen.

Bei der wissenschaftlichen Messwerterfassung werden innerhalb eines Messvorgangs oftmals mehrere Messwerte ermittelt. Mathematisch können solche Sätze von Messwerten durch Vektoren oder Matrizen beschrieben werden. Mathematische Vektor- und Matrizenkonstrukte werden in der elektronischen Datenverarbeitung durch Arrays beschrieben. Arrays sind Sätze gleicher Variablen, die Array-Elemente genannt werden. Jedem Array-Element sind definitionsgemäß einer oder mehrere ganzzahlige Indizes i_1, i_2, i_3, … i_n zugeordnet, so dass jedes Element eindeutig anhand seines Index oder seiner Indizes identifiziert werden kann. Die Zahl der Indizes beschreibt die Dimension des Arrays. Bei eindimensionalen Arrays wird jeder Variablen nur ein Index zugeordnet, so dass es sich mathematisch um einen Vektor handelt. Bei zweidimensionalen Arrays sind zwei Indizes zur Identifizierung eines Elementes erforderlich, so dass seine Definition mathematisch mit der einer Matrix kompatibel ist.

Will man in elektronischen Datenverarbeitungssystemen Messgrößen in Form einer stetigen Funktion wie Chromatogramme und Spektren weiterverarbeiten, so steht man vor dem Problem, dass eine mathematisch exakte Beschreibung des Zusammenhangs zwischen der Unabhängigen und dem Funktionswert ein vom Aufwand her unlösbares Problem ist. Daher bleibt nichts anderes übrig, als die unendlich große Schar an Wertepaaren, die sich aus dem stetigen funktionalen Zusammenhang ergibt, auf eine endliche Zahl an Datenpunkten zu reduzieren. Ein solcher Vorgang ist beispielsweise die Analog-/Digitalwandlung (A/D-Wandlung) der Ausgangsspannung eines Meßsystems. Aus der permanent anliegenden Ausgangsspannung werden z. B. 10 Wertepaare pro Sekunde aus Spannung und Zeit erzeugt und mittels eines Computers gespeichert. Am leichtesten kann eine Speicherung solcher Daten erfolgen, wenn der Abstand zwischen zwei aufeinander folgenden unabhängigen Variablen immer gleich bzw. äquidistant ist. In diesem Fall können die Wertepaare in einem Array, also als Vektor, gespeichert werden. Der Abstand zweier Array-Elemente ist durch die Differenz ihrer Indizes definiert. Der Index ist ein Maß für die unabhängige Variable. Bei Chromatogrammen ist der Index ein Maß für die Retentionszeit und bei Spektren ein Maß für die Wellenlänge. Jedem Index ist genau ein Element zugeordnet, im welchem die zugehörige Messgröße, bei Spektren beispielsweise die Spektraldichte, gespeichert ist.

Die Anzahl der Arrayelemente pro Einheit der Unabhängigen (Spektraldichteeinheit, Retentionszeiteinheit) wird als Auflösung bezeichnet. Dieser Auflösungsbegriff bezeichnet unabhängig von der chromatograpischen oder spektralen Auflösung die Güte der digital gespeicherten Daten. Je höher diese digitale Auflösung ist, desto mehr Arrayelemente und somit um so mehr Speicher werden benötigt. Die digitale Auflösung sollte höher sein als die Auflösung der Signale, die in dem Array gespeichert sind, da sonst analytische Informationen bei der Speicherung verloren gehen.

Wichtig in diesem Zusammenhang ist, dass auch der in einem Arrayelement gespeicherte Messwert keine stetige Größe mehr ist, sobald er aus einer A/D-Wandlung stammt. Bei einer solchen Wandlung wird jedem Messwert eine ganze Zahl zugeordnet. Die Spannweite an ganzen Zahlen, die zur Verfügung steht, hängt von der Auflösung der verwendeten elektronischen Schaltung, dem Analog-/Digitalwandler (A/D-Wandler), ab. Dieser wird meistens in Bit angegeben. Beim 12 Bit-Wandler liegt die Spannweite der zur Verfügung stehenden Zahlen bei Null bis 2^{12}-1. Es können also 2^D-Zustände unterschieden werden, wobei D die Auflösung des Wandlers in Bit ist. Um bei der Wandlung keine Informationen zu verlieren, muss die Auflösung des A/D-Wandlers so groß sein, dass sich das Rauschen auch nach der Wandlung in den gespeicherten Messwerten widerspiegelt.

4.3 Distanzbestimmungen

Um einen Sachverhalt aufzuklären oder zu quantifizieren, werden in der Analytik bei einzelnen Messvorgängen häufig nicht nur ein Messwert sondern ein ganzer Satz von Messwerten bestimmt. Man erhält also einen Messwertvektor:

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \tag{4.3-1}$$

mit

$\vec{\mathbf{x}}$ = Messwertvektor und

x_i = Einzelmesswert mit dem Index $i = 1, 2, 3, ..., n$.

Hierbei werden die Messwerte in n Klassen eingeteilt. Klasse 1 sind beispielsweise die Spektraldichten einer bestimmten Linie, Klasse 2 die Spektraldichten einer weiteren Linie u. s. w.. Soll der Unterschied zwischen zwei Messungen quantifiziert werden, so kann dies mit Hilfe der Bestimmung der Länge des Differenzvektors geschehen:

$$\left|\vec{\mathbf{x}} - \vec{\mathbf{z}}\right| = \left| \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{pmatrix} \right| = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + (x_3 - z_3)^2 + \cdots + (x_n - z_n)^2} \tag{4.3-2}$$

mit

$\left|\vec{\mathbf{x}} - \vec{\mathbf{z}}\right|$ = Betrag des Differenzvektors von $\vec{\mathbf{x}}$ und $\vec{\mathbf{z}}$.

Sind die Einzelmesswerte zu einer Messung unterschiedlich skaliert, so ist diese Differenzvektorlänge kein sinnvolles Maß für den Unterschied zweier Messungen. Liegt die Ursache der unterschiedlichen Skalierung in der Verwendung verschiedener Einheiten wie Sekunden und Millisekunden, so kann dieses Problem durch Angleichen der Einheiten gelöst werden. Werden jedoch verschiedene Messgrößen wie Zeiten und Spektraldichten als Einzelmesswerte aufgezeichnet, so muss eine Standardisierung vorgenommen werden.[22] Bei einem Satz von Messwerten

$$\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_N \quad \text{mit} \quad \vec{\mathbf{x}}_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \\ \vdots \\ x_{i,n} \end{pmatrix} = \text{Messwertevektor der } i\text{-ten Messung} \tag{4.3-3}$$

wird üblicherweise zunächst die folgende Quadratsumme bestimmt:

$$\begin{aligned} Q_k &= \sum_i \left(x_{i,k} - \bar{x}_k\right)^2 \\ &\text{mit} \\ \bar{x}_k &= \text{Mittelwert aller Einzelmessungen der Klasse } k \text{ und} \\ Q_k &= \text{zur Standardisierung der Messwertklasse } k \text{ herangezogene} \\ &\quad\ \text{Quadratsumme.} \end{aligned} \tag{4.3-4}$$

Die Standardisierung erfolgt durch Division aller entsprechenden Messwerte durch die Wurzel der jeweiligen Quadratsumme. In vielen Fällen ist es vorteilhaft die Messwerte zu zentrieren. Unter Zentrierung von Messwerten versteht man die Subtraktion mit dem Mittelwert aller Einzelmesswerte der Klasse k der Messreihe:

$$\begin{aligned} x_{i,k}^{\mathrm{s}} &= \frac{x_{i,k}}{\sqrt{Q_k}} \quad = \text{standardisierter Messwert.} \\ x_{i,k}^{\mathrm{sz}} &= \frac{x_{i,k} - \bar{x}_k}{\sqrt{Q_k}} = \text{standardisierter und zentrierter Messwert.} \end{aligned} \tag{4.3-5}$$

Führt man die Differenzvektorlängenbestimmung mit standardisierten Messwerten durch, so sind die Messwerte jeder Klasse auf den Wertebereich der gesamten Messreihe skaliert.

Für die folgende Fragestellung existiert eine alternative Methode, die über einen Indizierungs- und Sortieralgorithmus arbeitet: Gegeben seien eine Einzelmessung mit dem Messvektor $\vec{\mathbf{v}}_{\mathrm{E}}$ und die Ergebnisse einer Messreihe bestehend aus vielen Messvektoren ($\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots$). Die Frage lautet, welche Messung aus der Messreihe bezüglich der erfassten Messdaten der Einzelmessung am ähnlichsten ist. Die Einzelmesswerte sind hierbei unterschiedlich skaliert. Um diese Frage zu beantworten sortiert man die Daten aus der Messreihe jeweils nach jeder Klasse k bezüglich der Ähnlichkeit zum Einzelmesswert. Das Sortierkriterium ist die Betrags-

differenz des zu der Klasse k gehörenden Einzelwertes der Einzelmessung und des jeweiligen Einzelwertes der Messung aus der Messreihe:

$$\text{Seien } \left|\vec{\mathbf{v}}_{\mathrm{E}}\right| = \begin{pmatrix} v_{\mathrm{E},1} \\ v_{\mathrm{E},2} \\ v_{\mathrm{E},3} \\ \vdots \\ v_{\mathrm{E},n} \end{pmatrix} = \text{Messvektor und } \left|\vec{\mathbf{v}}_{\mathrm{i}}\right| = \begin{pmatrix} v_{\mathrm{i},1} \\ v_{\mathrm{i},2} \\ v_{\mathrm{i},3} \\ \vdots \\ v_{\mathrm{i},n} \end{pmatrix} \text{ beliebiger Vektor der Messreihe} \tag{4.3-6}$$

$$\Rightarrow \text{die Höhe von } \left|v_{\mathrm{E},k} - v_{i,k}\right| \text{ ist das Sortierkriterium.}$$

Die Position eines jeden Messvektors in der sortierten Liste aller Messvektoren wird für jede Klasse k gespeichert. Eine solche Liste liegt in einem Computerprogramm meistens als Array vor, so dass die Position des Messvektors dem Arrayindex des sortierten Arrays entspricht. Die Indizierung sollte hierbei von 1 an aufsteigend erfolgen. Auf diese Weise erhält man für jeden Messvektor $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots$ einen Indexvektor $\vec{\mathbf{i}}_1, \vec{\mathbf{i}}_2, \ldots$ in dem dessen Position in der sortierten Liste für alle k Sortiervorgänge gespeichert ist. Das Maß für die Distanz ist die Länge der Indexvektoren:

$$\left|\vec{\mathbf{i}}\right| = \left\|\begin{pmatrix} i_1 \\ i_2 \\ i_3 \\ \vdots \\ i_n \end{pmatrix}\right\| = \sqrt{(i_1)^2 + (i_2)^2 + (i_3)^2 + \cdots + (i_n)^2} \tag{4.3-7}$$

mit

$\left|\vec{\mathbf{i}}\right|$ = Betrag des Indexvektors $\vec{\mathbf{i}}$.

Tabelle 4.3-1 enthält als Beispiel die Zahlenwerte für einen Einzelvektor $\vec{\mathbf{v}}_{\mathrm{E}}$, für den herausgefunden werden soll, welcher der beiden Vektoren $\vec{\mathbf{v}}_1$ oder $\vec{\mathbf{v}}_2$ ihm ähnlicher ist. Nach der verwendeten Rechenmethode ist $\vec{\mathbf{v}}_1$ dem Vektor $\vec{\mathbf{v}}_{\mathrm{E}}$ ähnlicher.

Tabelle 4.3-1: Beispiel für eine Distanzbestimmung über Sortieren und Indizieren

$\vec{\mathbf{v}}_{\mathrm{E}}$	$\vec{\mathbf{v}}_1$	$\vec{\mathbf{v}}_2$	$\vec{\mathbf{i}}_1$	$\vec{\mathbf{i}}_2$	$\left\|\vec{\mathbf{i}}_1\right\|$	$\left\|\vec{\mathbf{i}}_2\right\|$
2	5	8	1	2	2,65	3,61
4	7	11	1	2		
17	2	16	2	1		
20000	12000	30000	1	2		

Manchmal ist es erforderlich die einzelnen Klassen k zu gewichten. Dies geschieht mit sogenannten Gewichtungsfaktoren:

$$
\begin{aligned}
\left|\vec{\mathbf{x}}-\vec{\mathbf{z}}\right|_{\text{gewichtet}} &= \sqrt{[g_1(x_1-z_1)]^2+[g_2(x_2-z_2)]^2+[g_3(x_3-z_3)]^2+\cdots+[g_n(x_n-z_n)]^2} \\
&\text{bzw.} \\
\left|\vec{\mathbf{i}}\right|_{\text{gewichtet}} &= \sqrt{[g_1\cdot i_1]^2+[g_2\cdot i_2]^2+[g_3\cdot i_3]^2+\cdots+[g_n\cdot i_n]^2} \qquad (4.3\text{-}8) \\
&\text{mit} \\
g_{1,2,3,\ldots,n} &= \text{Gewichtungsfaktoren.}
\end{aligned}
$$

Werden standardisierte Messwerte verwendet, so ist es hierbei unwesentlich, ob diese auch zentriert worden sind oder nicht. Bisweilen wird bei Distanzvergleichen in Computerprogrammen auf das Ziehen der Quadratwurzel verzichtet, um Rechenzeit zu sparen. Man verwendet dann die Quadrate der Vektorbeträge als Distanzmaße:

$$
\begin{aligned}
\left|\vec{\mathbf{x}}-\vec{\mathbf{z}}\right|^2_{\text{gewichtet}} &= [g_1(x_1-z_1)]^2+[g_2(x_2-z_2)]^2+[g_3(x_3-z_3)]^2+\cdots+[g_n(x_n-z_n)]^2 \\
&\text{bzw.} \qquad (4.3\text{-}9) \\
\left|\vec{\mathbf{i}}\right|^2_{\text{gewichtet}} &= [g_1\cdot i_1]^2+[g_2\cdot i_2]^2+[g_3\cdot i_3]^2+\cdots+[g_n\cdot i_n]^2
\end{aligned}
$$

4.4 Sortieren von Zahlenreihen

Die Distanzbestimmung mittels eines Indexvektors (siehe vorheriger Abschnitt) ist einer von vielen Softwarealgorithmen, in denen ein Sortiervorgang vorkommt. Effektiv gestaltete Sortiervorgänge sind oftmals die Voraussetzung für eine zufriedenstellend arbeitende Software. Einer der einfachsten und zusätzlich sehr effizienten Sortieralgorithmen ist das Shell-Sort-Verfahren. [10]

Voraussetzung für den Einsatz des Shell-Sort-Algorithmus ist eine indizierte Liste von Variablen, welche häufig über ein Array realisiert wird:

$$
\begin{aligned}
\mathbf{A} &= \text{Array} \\
&\text{mit.} \qquad (4.4\text{-}1) \\
\mathbf{A}_{[i]} &= \text{Arrayelement mit dem ganzzahligen Index } i. \\
i &= 0,1,2,\ldots,L \;=\; \text{Indexvariable (höchster Index} = L).
\end{aligned}
$$

Zuerst speichert man den ganzzahlig halbierten höchsten Indexwert in der Variable l ab:

$$l \quad = \quad \text{ganzzahliger Anteil von } \frac{L}{2}. \tag{4.4-2}$$

Anschließend vergleicht man die Elemente $\mathbf{A}_{[0]}$ und $\mathbf{A}_{[l]}$, $\mathbf{A}_{[1]}$ und $\mathbf{A}_{[l+1]}$, ... miteinander. Will man absteigend sortieren, so werden die verglichenen Arrayelemente dann getauscht, wenn das mit dem höheren Index den höheren Wert enthält. Soll aufsteigend sortiert werden, so wird dann getauscht, wenn das Arrayelement mit dem höheren Index den niedrigeren Wert enthält. Hierbei ist zu beachten, dass sobald ein Tauschvorgang bei einem Element mit dem Index i durchgeführt wird, das i-l-te mit dem i-ten Element, danach das i-$2l$-te mit dem i-l-ten Element usw. verglichen wird, bis man wieder am vordersten Arrayelement ankommt oder dem Vergleich kein Tauschvorgang mehr folgt. Für ein Beispielarray aus 8 Elementen, welches absteigend zu sortieren ist, bedeutet dies (Abbildung 4.4-1):

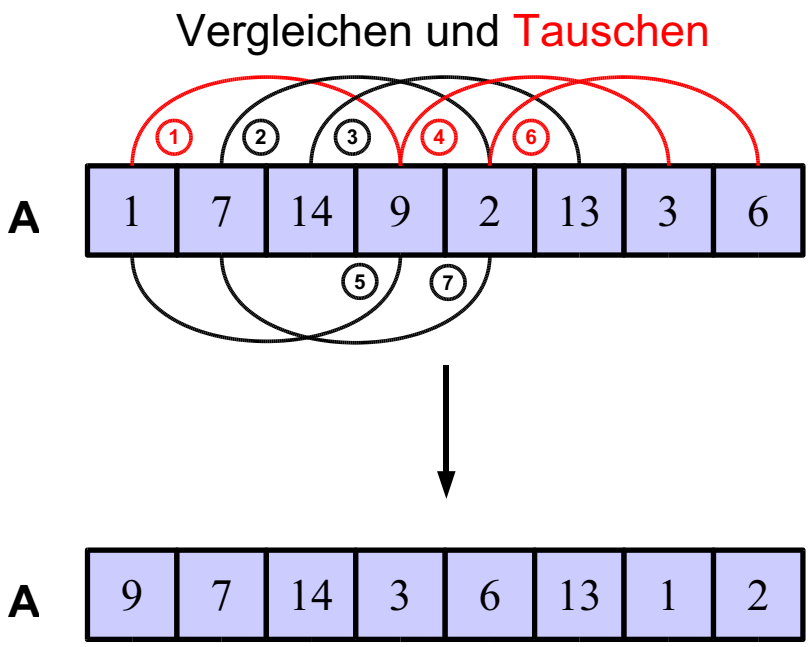

Abbildung 4.4-1: Erste Tauschreihe des Shell-Sort-Algorithmus (L/2 = 7/2 = 3,5 => l = 3)

Anschließend wird der Arrayelementeabstand l noch einmal ganzzahlig halbiert und das Vergleichs- und Tauschverfahren noch einmal durchgeführt. Abbildung 4.4-2 verdeutlicht den Vorgang anhand des Array-Beispiels.

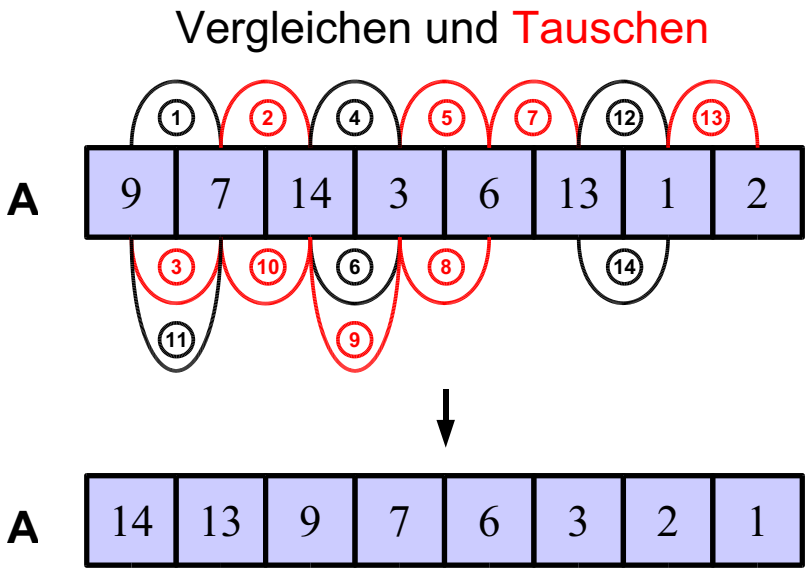

Abbildung 4.4-2: Zweite Tauschreihe des Shell-Sort-Algorithmus (l = 1)

Es müssen insgesamt so viele ganzzahlige Halbierungen von *l* mit anschließendem Tauschen der Arrayelemente nach dem beschriebenen Verfahren durchgeführt werden, bis *l* Eins ist. Bei dem Beispielarray muss demnach keine weitere Tauschreihe mehr durchgeführt werden. Die Elemente des Arrays sind vom niedrigen zum hohen Index hin absteigend sortiert.

Es existieren noch einfachere Sortierverfahren als der Shell-Sort-Algorithmus wie das Austauschverfahren (Exchange-Sort) oder das Auswahlverfahren.[10] Diese haben jedoch den Nachteil, dass die Anzahl der Tausch- und Vergleichsoperationen mit wachsender Arraygröße viel stärker ansteigen, als beim Shell-Sort-Verfahren.

4.5 Glättungsalgorithmen

Analytische Messwerte sind immer mehr oder weniger durch Rauschen gestört. In Arrays gespeicherte Chromatogramme oder Spektren sind hierbei keine Ausnahme. Um die in den Arrayelementen gespeicherten Messwerte zu glätten stehen mit der Bewegtsegment-mittelung[31][35] und der Scharmittelung[35] zwei einfache und effiziente Verfahren zur Verfügung. Bei der Scharmittelung wird nicht nur eine Messung eines Chromatogrammes oder Spektrums pro Arrayelement in das Array übertragen, sondern mehrere Messungen aufaddiert. Jedes Element des Arrays wird anschließend durch die Anzahl der Messungen geteilt. Das Signal-Rauschverhältnis verbessert sich proportional zu der Quadratwurzel aus der Anzahl der Messungen. Nachteil dieses Verfahrens ist, dass für jedes Arrayelement mehrere Wiederholungsmessungen benötigt werden. Es kann daher zum einen nicht auf bereits aufgezeichnete Chromatogramme oder Spektren angewendet werden. Zum anderen steigt die Zeit, die für die Gewinnung der Daten erforderlich ist, um den Faktor der Anzahl der durchgeführten Wiederholungsmessungen an.

Bereits 1964 haben A. Savitzky und M. J. E Golay [31] grundlegende Arbeiten zur Bewegtsegmentmittelung veröffentlicht. Die Bewegtsegmentmittelung gehört zu den sogenannten digitalen Filterungsverfahren. Grundlage dieser Glättungsmethode ist, dass das im Array gespeicherte Chromatogramm oder Spektrum besser aufgelöst ist, als es für die enthaltenen Informationen erforderlich wäre. Es müssen also mehr Datenpunkte im Array enthalten sein, als benötigt. Unter dieser Bedingung können Veränderungen von benachbarten Arrayelementen dem Rauschen zugeordnet werden. Je größer der Abstand zwischen zwei Arrayelementen ist, desto höher ist die Wahrscheinlichkeit, dass die Veränderung ihrer Zahlenwerte nicht nur ein Ergebnis des Signalrauschens ist, sondern mit analytisch relevanten Informationen korrelieren. Bei der Bewegtsegmentmittelung wird daher jedes Array-Element durch einen Mittelwert aus ihm selbst und seiner bezüglich des Index nächsten Nachbarn ersetzt. Mathematisch kann die Bewegtsegmentmittelung folgendermaßen beschrieben werden [31]:

$$\mathbf{E}^*_{[j]} = \frac{\sum_{i=0}^{i=h}\left(\mathbf{C}_{[i]}\mathbf{E}_{[j-h/2+i]}\right)}{\sum_{i=0}^{i=h}\left(\mathbf{C}_{[i]}\right)} \tag{4.5-1}$$

$\mathbf{E}^*_{[j]}$ = Element des rauschgefilterten Arrays mit dem Index j.
$\mathbf{C}_{[i]}$ = Element i des Filterarrays **C**.
$\mathbf{E}_{[j-h/2+i]}$ = Element mit dem Index $j-(h/2)+i$ des Datenarrays.
h = höchster Index des Filterarrays; eine positive gerade Zahl.

Gleichung (4.5-1) liegt ein Algorithmus zu Grunde, bei dem aus dem Datenarray **E** ein neues Array $\mathbf{E}^*$ erzeugt wird, welches die geglätteten Daten enthält. Das Filterarray **C** verfügt über eine ungerade Anzahl an Elementen. Der höchste Index h ist jedoch eine gerade Zahl, da die Indizierung mit Null beginnt. Die in ihm enthaltenen Zahlenwerte repräsentieren die Gewichtung der einzelnen Arrayelemente, aus denen jeder gefilterte Zahlenwert errechnet wird. Ein einfaches Beispiel ist ein Filterarray, welches sich aus den drei Elementen $C_{[0]} = 1$, $\mathbf{C}_{[1]} = 2$ und $\mathbf{C}_{[2]} = 1$ zusammensetzt. Das Blockschema in Abbildung 4.5-1 verdeutlicht den Glättungsalgorithmus unter Einsatz dieses Filterarrays. Es werden jeweils drei benachbarte Elemente gemittelt, wobei das mittlere doppelt so stark gewichtet wird, wie die beiden äußeren.

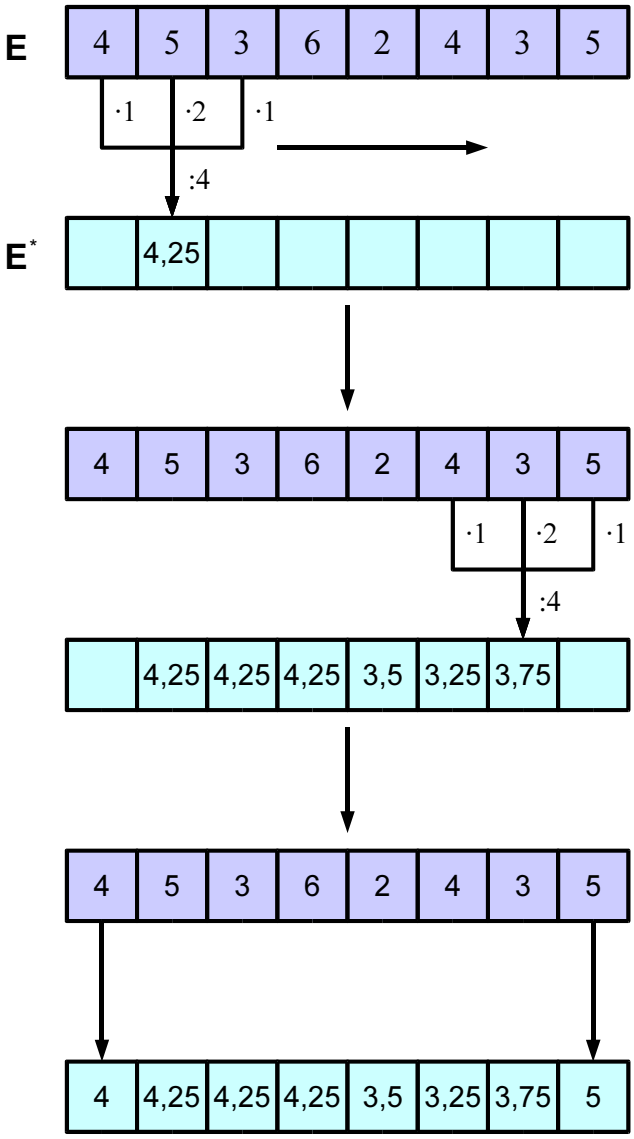

Abbildung 4.5-1: Blockschema zur Bewegtsegmentmittelung

Abbildung 4.5-2 zeigt verschiedene 7 Elemente große Filterarrays zur Bewegtsegmentmittelung. Den einfachsten Filter erhält man, wenn jedes Element des Arrays den gleichen Zahlenwert besitzt. Der Algorithmus in Gleichung (4.5-1) führt in diesem Fall zur Bildung eines einfachen Durchschnitts. Es kann sinnvoll sein, die mittleren Elemente des Filterarrays stärker zu gewichten. Durch eine derartige Gewichtung kann beispielsweise berücksichtigt werden, dass mit zunehmenden Abstand der Arrayelemente im Datenarray die Wahrscheinlichkeit steigt, dass die Unterschiede in den Zahlenwerten nicht vom Rauschen stammen, sondern analytische Informationen bergen. Zu diesem Zweck kann eine Dreiecksfunktion, welche aus zwei Geraden mit den Steigungen 1 und –1 zusammengesetzt ist, in das Filterarray eingelesen werden. Die quadrierte Form dieser Dreiecksfunktion, welche man auch als eine aus zwei Ästen einer Exponentialfunktion zusammengesetzte Funktion betrachten kann, gewichtet die mittleren Elemente des Filterarrays noch stärker. Eine weitere Möglichkeit ist das Einlesen einer symmetrischen Binominalverteilungsfunktion, welche die Form einer Gauß-Glockenkurve besitzt. Von A. Savitzky und M. J. E Golay [31] sind

Filterarrays berechnet worden die angewandt auf den durch Gleichung (4.5-1) repräsentierten Algorithmus das gleiche Ergebnis liefern, als ob man mit den entsprechenden Zahlenwerten eine Polynomregression durchgeführt hätte. Tabelle 4.5-2 enthält Filterarrays, denen eine Polynomregression mit Polynomen 2. bzw. 3 Grades zu Grunde liegt. D. h. der Filter auf Grundlage einer Polynomregression mit einem Polynom 2. Grades gleichen dem auf Grundlage einer Polynomregression mit Polynomen 3. Grades.

Tabelle 4.5-1: Filterarrays mit Binominalkoeffizienten

Elementindex	*C*, 3 Elemente	*C*, 5 Elemente	*C*, 7 Elemente	*C*, 9 Elemente
0				1
1			1	8
2		1	6	28
3	1	4	15	56
4	2	6	20	70
5	1	4	15	56
6		1	6	28
7			1	8
8				1

Tabelle 4.5-2: Filterarrays mit Koeffizienten, die nach der Methode der kleinsten Quadrate für Polynome 2. und 3. Gerades ermittelt worden sind

Elementindex	*C*, 5 Elemente	*C*, 7 Elemente	*C*, 9 Elemente	*C*, 11 Elemente
0				-36
1			-21	9
2		-2	14	44
3	-3	3	39	69
4	12	6	54	84
5	17	7	59	89
6	12	6	54	84
7	-3	3	39	69
8		-2	14	44
9			-21	9
10				-36

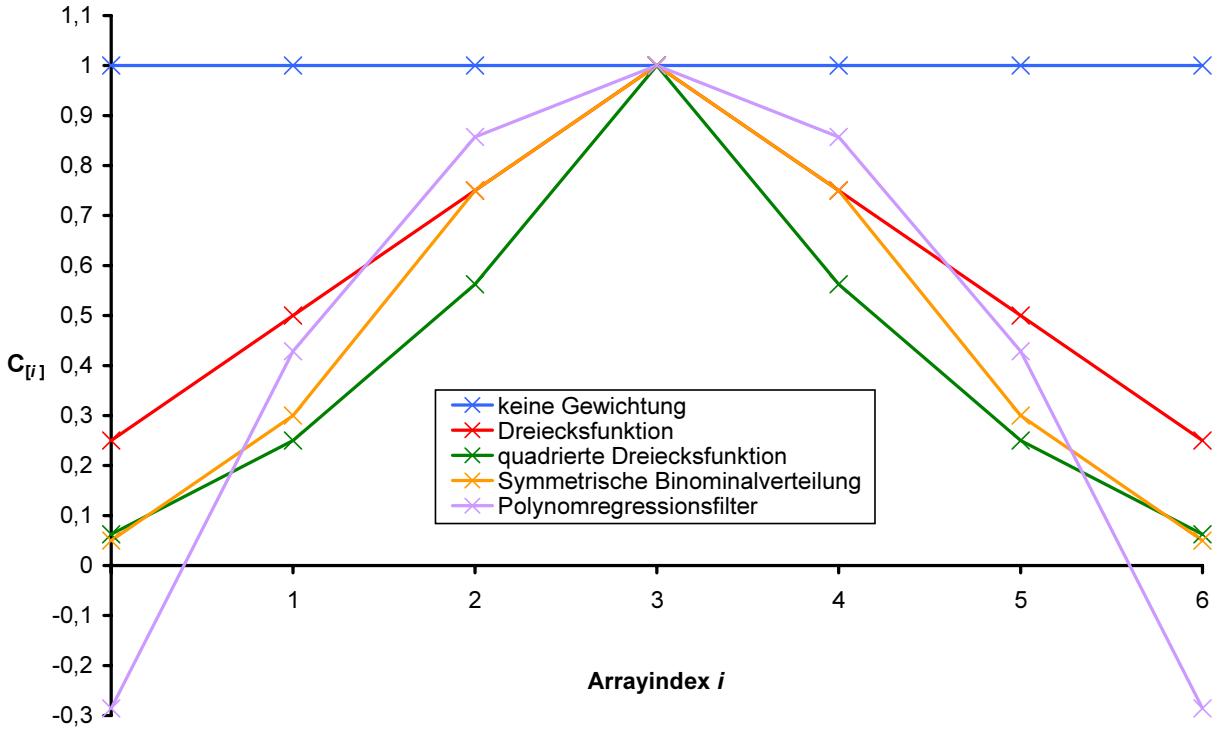

Abbildung 4.5-2: Verschiedene Filterarrays zur Glättung von Datenarrays (das Element mit dem höchsten Zahlenwert ist auf 1 normiert worden)

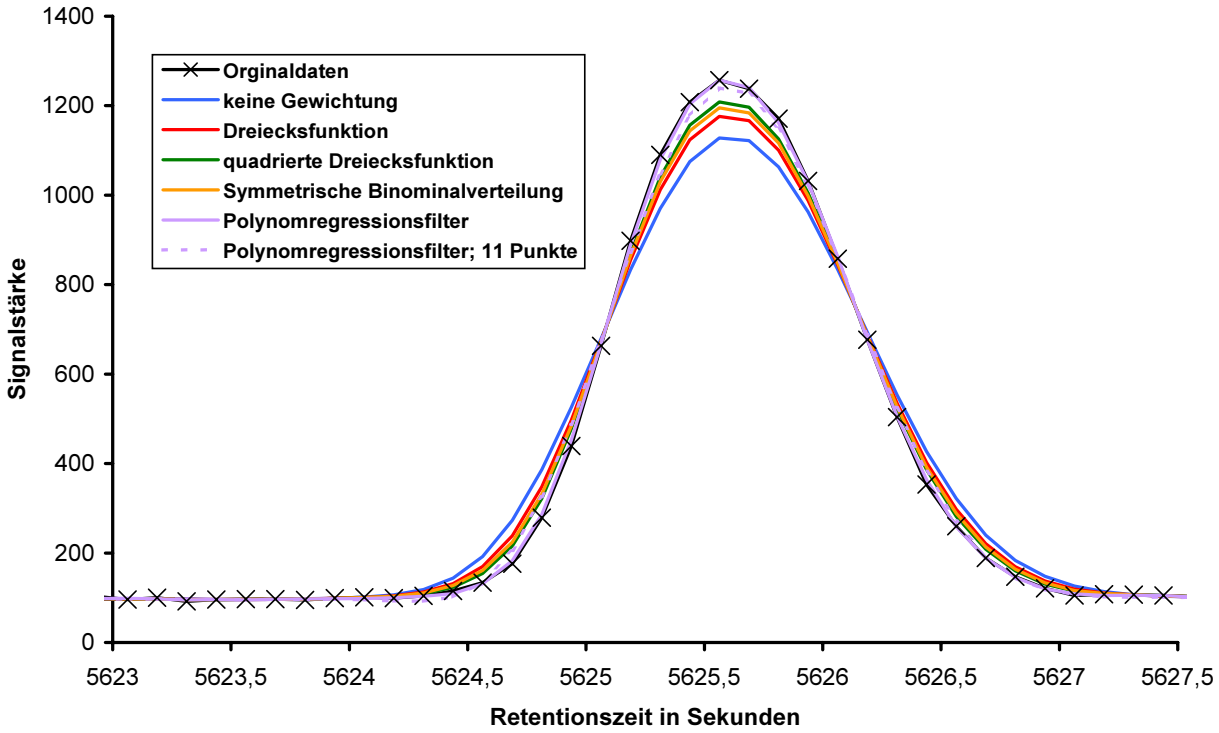

Abbildung 4.5-3: Die Auswirkung der Bewegtsegmentglättung (falls nicht anders angegeben 7 Filterarrayelemente) auf ein chromatographisches Signal

Wendet man die in Abbildung 4.5-2 graphisch dargestellten Filter auf ein gut aufgelöstes Signal an, so führt dies in fast allen Fällen zu einer Stauchung des Signals. Das Ausmaß der Stauchung hängt von der Art der Filterfunktion ab. Nur beim Filter auf der Basis der Polynomregression ist eine solche Stauchung kaum zu beobachten (siehe Abbildung 4.5-3 und Abbildung 4.5-4). [31] Bei Anwendung dieser Filter auf einen sogenannten Spike wird dieser mehr oder weniger stark gedämpft. Seine Form nähert sich der der Filterfunktion an (siehe Abbildung 4.5-6). [31] Ein Spike ist ein einzelner Wert im Datenarray, welcher sich von seinen Nachbarn sehr stark unterscheidet. Spikes in analytischen Messreihen werden oft durch elektronische Störungen des Messsystems verursacht. Die Signaldeformation, die dem Spike wiederfährt, tritt in abgeschwächter Form auch bei Filterung von schlecht aufgelösten Spektren oder Chromatogrammen auf (siehe Abbildung 4.5-7). Während bei allen anderen Filtern sich die Deformation weitgehend auf eine Stauchung des Signals mit einhergehender Verbreiterung der Signalbasis beschränkt, wird bei Polynomregressionsfiltern die Basislinie verfälscht. Die Dämpfung des Rauschens ist bei gleich großen Filterarrays um so schwächer, je stärker die mittleren Elemente des Filterarrays gewichtet werden (siehe Abbildung 4.5-5 und Tabelle 4.5-3). Am schwächsten ist sie für ein Filterarray auf Basis der Polynomregression (Polynomregressionsfilter). Es bedarf beispielsweise eines Polynomregressionsfilterarrays von 11 Punkten um den gleichen Rauschglättungseffekt zu erzielen wie mit einem Filter auf Basis einer symmetrischen Binominalverteilung mit 7 Elementen.

Tabelle 4.5-3: Die Auswirkung der Bewegtsegmentglättung auf die Standardabweichung des Rauschens bei einer chromatographischen Messung

Filtertyp (falls nicht anders angegeben 7 Filterarrayelemente)	Standardabweichung
Gleitender Durchschnitt	2,38
Dreiecksfunktion	0,81
Quadrierte Dreiecksfunktion	0,92
Symmetrische Binominalverteilung	1,07
Polynomregressionsfilter	1,29
Polynomregressionsfilter auf Basis von 11 Filterarrayelementen	1,01

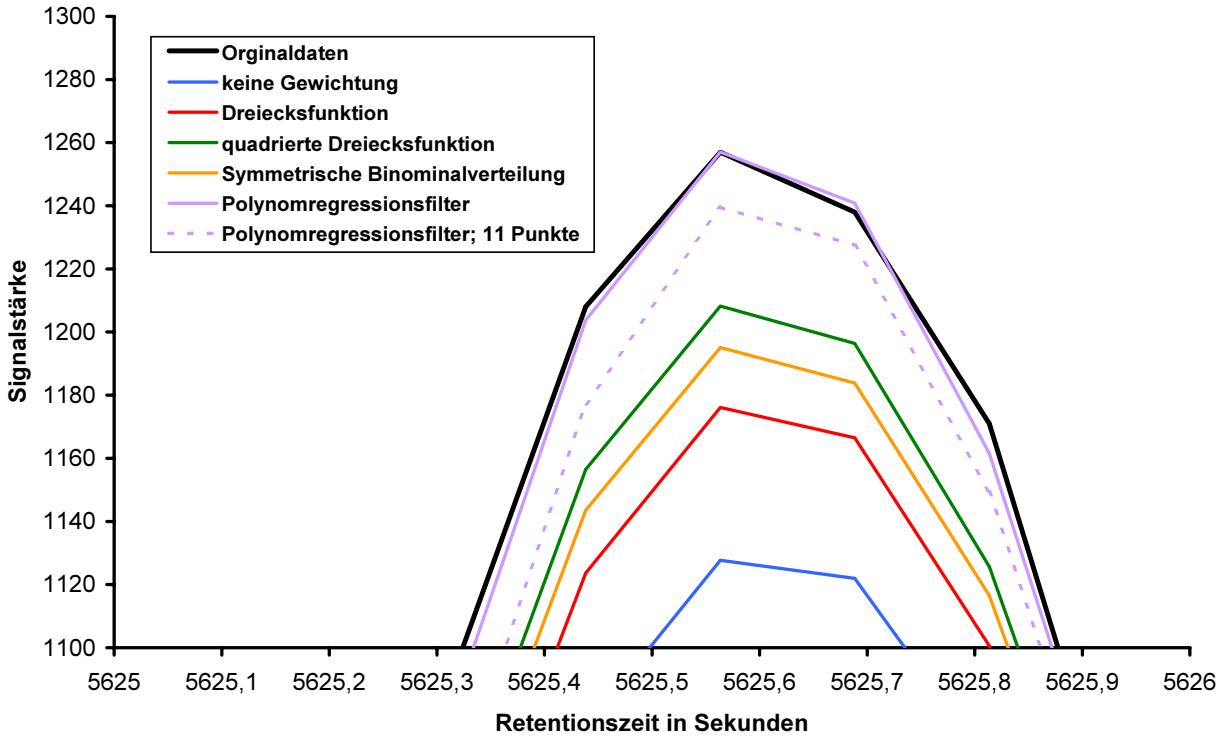

Abbildung 4.5-4: Die Auswirkung der Bewegtsegmentglättung (falls nicht anders angegeben 7 Filterarrayelemente) auf die Spitze eines chromatographischen Signals

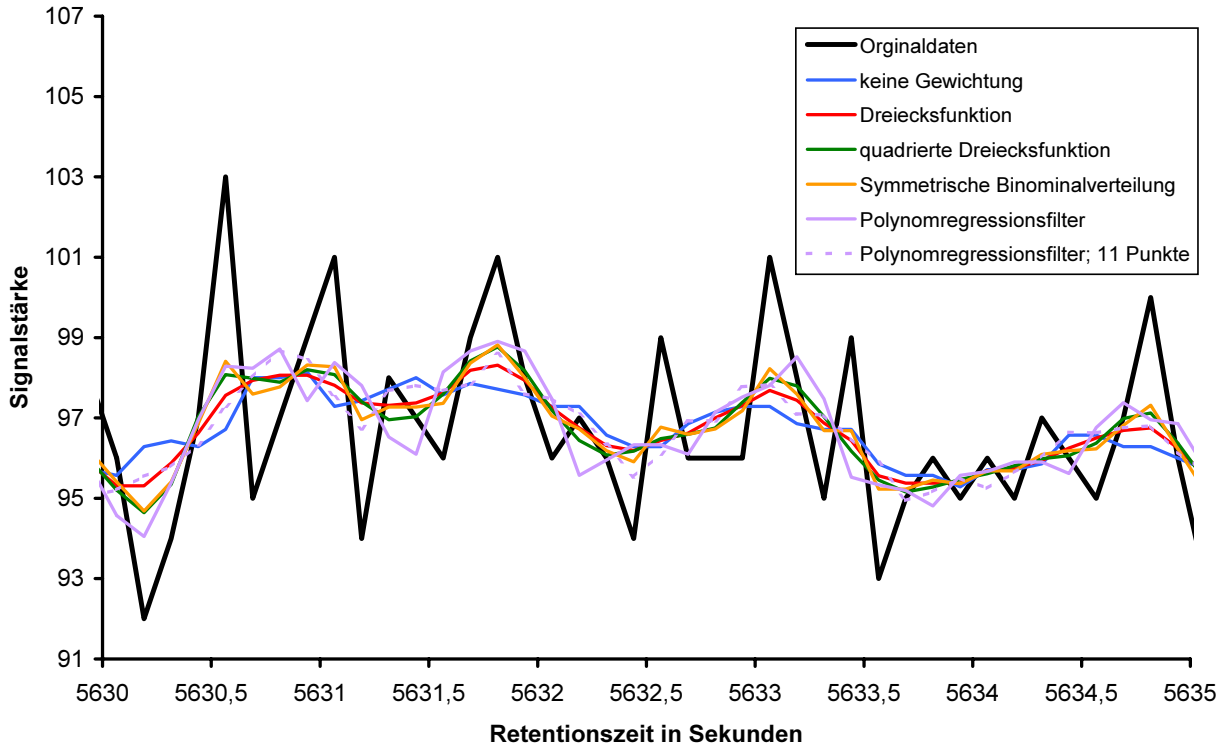

Abbildung 4.5-5: Die Auswirkung der Bewegtsegmentglättung (falls nicht anders angegeben 7 Filterarrayelemente) auf das Rauschen in einem Chromatogramm

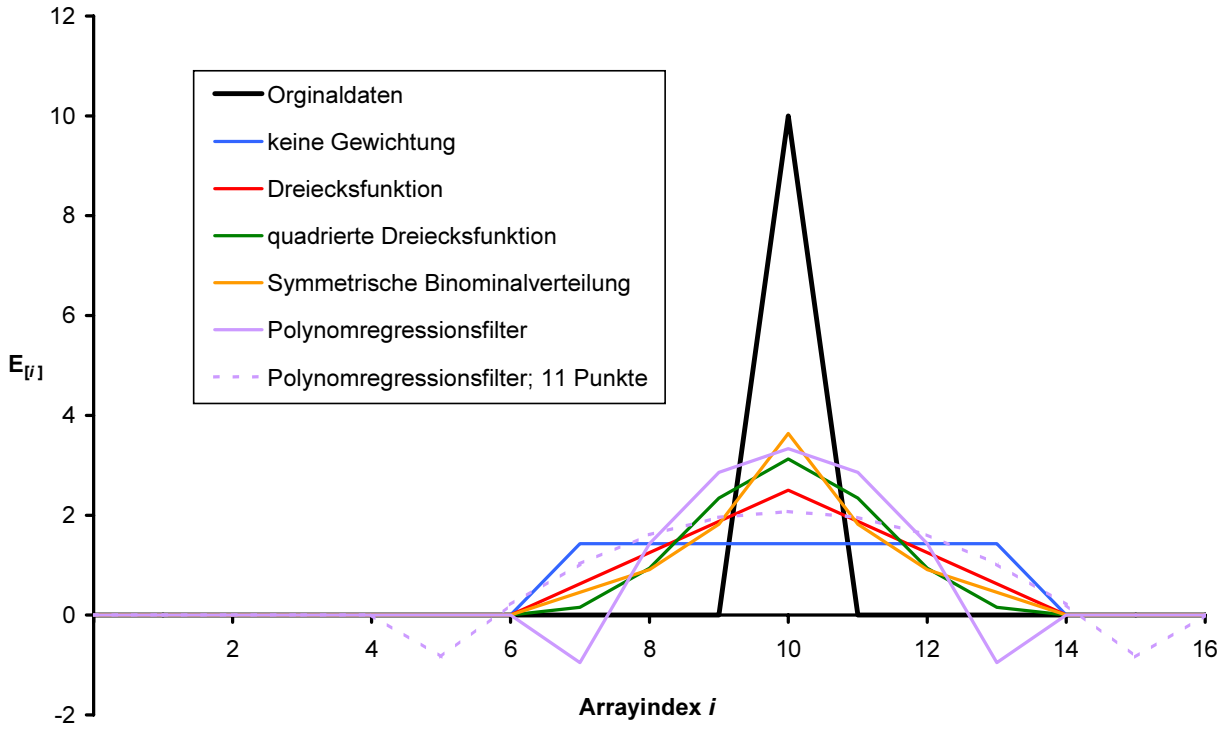

Abbildung 4.5-6: Die Auswirkung der Bewegtsegmentglättung (falls nicht anders angegeben 7 Filterarrayelemente) auf einen Spike

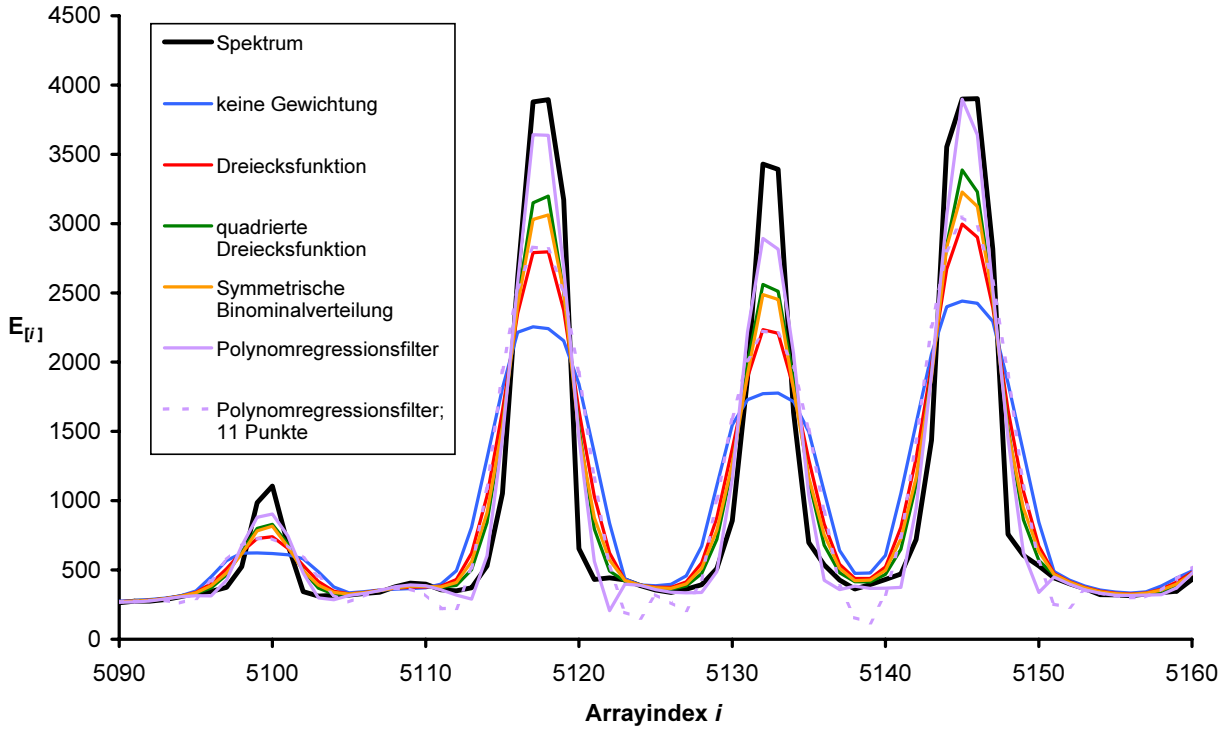

Abbildung 4.5-7: Die Auswirkung der Bewegtsegmentglättung (falls nicht anders angegeben 7 Filterarrayelemente) auf ein schlecht aufgelöstes Spektrum

4.6 Der Gauß-Algorithmus

Der Gauß-Algorithmus eignet sich sowohl zur Lösung linearer Gleichungssysteme, als auch zur Vereinfachung von Determinantenberechnungen und Matrixinversionen. Die Berechnung von Determinanten basiert auf dem Laplaceschen Entwicklungssatz. Wendet man diesen Entwicklungssatz bei der Lösung von mehrzeiligen Determinanten an, ohne die Determinantenmatrix zwecks Optimierung umzustellen, so führt dies zu einem exorbitant hohen Rechenaufwand. Der Gauß-Algorithmus kann helfen, den Rechenaufwand drastisch zu senken. Bei Matrixinversionen hilft der Gauß-Algorithmus aufwendige Determinantenberechnungen zu vermeiden.

4.6.1 Das Gaußsche Diagonalisierungsverfahren

Eines der bekanntesten Verfahren, mit denen ein beliebiges lineares Gleichungssystem mit $n \in \mathrm{IN}^{>1}$ lineare Gleichungen mit n Unbekannten gelöst werden kann, ist der Gauß-Algorithmus. Ein wichtiger Vorteil dieses Lösungsverfahrens ist seine gute Eignung zur Implementierung in ein Computerprogramm. Ausgangspunkt des Algorithmus ist folgende allgemeine Form eines linearen Gleichungssystems:

$$\begin{aligned}
a_{1,1}x_1 &+ a_{1,2}x_2 + \cdots + a_{1,n}x_n = c_1 \\
a_{2,1}x_1 &+ a_{2,2}x_2 + \cdots + a_{2,n}x_n = c_2 \\
a_{3,1}x_1 &+ a_{3,2}x_2 + \cdots + a_{3,n}x_n = c_3 \\
\vdots \quad & \quad \vdots \qquad\qquad\quad \vdots \\
a_{n,1}x_1 &+ a_{n,2}x_2 + \cdots + a_{n,n}x_n = c_n
\end{aligned} \tag{4.6-1}$$

wobei $a_{1,1} \neq 0$ und Koeffzientendeterminante $D \neq 0$.

Das lineare Gleichungssystem kann auch als Matrix aufgefasst werden:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & \cdots & a_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{pmatrix} \quad \text{(Koeffizientenmatrix mal Unabhängigen - vektor gleich Lösungsvektor)} \tag{4.6-2}$$

Man multipliziert die erste Gleichung von (4.6-1) mit $-a_{2,1}/a_{1,1}$ und addiert diese zur zweiten Gleichung. Anschließend wird erste Gleichung mit $-a_{3,1}/a_{1,1}$ multipliziert und zur dritten addiert. Verfährt man so bis zur n-ten Gleichung, so steht unter der ersten Gleichung ein lineares Gleichungssystem mit n-1 Gleichungen und $n-1$ Unbekannten:

$$\begin{array}{ccccccccc}
a'_{2,2}x_2 & + & a'_{2,3}x_3 & + & \cdots & + & a'_{2,n}x_n & = & c'_2 \\
a'_{3,2}x_2 & + & a'_{3,3}x_3 & + & \cdots & + & a'_{4,n}x_n & = & c'_3 \\
a'_{4,2}x_2 & + & a'_{4,3}x_3 & + & \cdots & + & a'_{4,n}x_n & = & c'_4 \\
\vdots & & \vdots & & & & \vdots & & \\
a'_{n,2}x_2 & + & a'_{n,3}x_3 & + & \cdots & + & a'_{n,n}x_n & = & c'_n
\end{array}$$

bzw. in Matrixschreibweise: (4.6-3)

$$\begin{pmatrix} a'_{2,2} & a'_{2,3} & \cdots & a'_{2,n} \\ a'_{3,2} & a'_{3,3} & \cdots & a'_{4,n} \\ a'_{4,2} & a'_{4,3} & \cdots & a'_{4,n} \\ \vdots & \vdots & \ddots & \vdots \\ a'_{n,2} & a'_{n,3} & \cdots & a'_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c'_2 \\ c'_3 \\ c'_4 \\ \vdots \\ c'_n \end{pmatrix}$$

Mit diesem Gleichungssystem führt man den für (4.6-1) bzw. (4.6-2) beschriebenen Eliminierungsschritt nochmals aus, so dass ein Gleichungssystem mit n - 2 Gleichungen und n - 2 Unbekannten erzeugt wird. Die Eliminierungsschritte werden so lange wiederholt, bis das lineare Gleichungssystem in der Dreiecksform vorliegt.

$$\begin{array}{ccccccccccc}
a_{1,1}x_1 & + & a_{1,2}x_2 & + & a_{1,3}x_3 & + & \cdots & + & a_{1,n}x_n & = & c_1 \\
 & + & a'_{2,2}x_2 & + & a'_{2,3}x_3 & + & \cdots & + & a'_{2,n}x_n & = & c'_2 \\
 & & & & a''_{2,3}x_3 & + & \cdots & + & a''_{3,n}x_n & = & c''_3 \\
 & & & & & & & & \vdots & & \vdots \\
 & & & & & & & = & a^{(n-1)}_{n,n}x_n & = & c^{(n-1)}_n
\end{array}$$

bzw.: (4.6-4)

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ 0 & a'_{2,2} & a'_{2,3} & \cdots & a'_{2,n} \\ 0 & 0 & a''_{2,3} & \cdots & a''_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & a^{(n-1)}_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c'_1 \\ c''_2 \\ c'''_3 \\ \vdots \\ c^{(n-1)}_n \end{pmatrix}$$

Jetzt kann x_n berechnet werden. Durch Einsetzen in die entsprechenden Gleichungen von unten nach oben ist die Berechnung der übrigen Unbekannten möglich. Dieses Verfahren wird auch als Rücksubstitution bezeichnet.

Das strikte Befolgen dieses Algorithmus führt zu Problemen, sobald bei irgendeinem Eliminierungsschritt ein Gleichungssystem erhalten wird, dessen erster Koeffizient in der ersten Zeile Null ist. Dies ist bei folgendem Gleichungssystem der Fall:

$$\begin{aligned} 3x_1 + 1x_2 + 7x_3 + 8x_4 &= 0 \\ 3x_1 + 1x_2 + 9x_3 + -2x_4 &= 0 \\ 1x_1 + 7x_2 + 5x_3 + 4x_4 &= 0 \\ -4x_1 + 4x_2 + 12x_3 + 8x_4 &= 4 \end{aligned} \xrightarrow[\text{nierung}]{\text{Elimi-}} \begin{aligned} 3x_1 + 1x_2 + 7x_3 + 8x_4 &= 0 \\ \mathbf{0}x_2 + 2x_3 + -10x_4 &= 0 \\ 6\tfrac{2}{3}x_2 + 2\tfrac{2}{3}x_3 + 1\tfrac{1}{3}x_4 &= 0 \\ 5\tfrac{1}{3}x_2 + 21\tfrac{1}{3}x_3 + 18\tfrac{2}{3}x_4 &= 4 \end{aligned} \tag{4.6-5}$$

$$\begin{aligned} 2x_2 + -10x_3 &= 0 \\ \rightarrow 6\tfrac{2}{3}x_2 + 2\tfrac{2}{3}x_3 + 1\tfrac{1}{3}x_4 &= 0 \\ 5\tfrac{1}{3}x_2 + 21\tfrac{1}{3}x_3 + 18\tfrac{2}{3}x_4 &= 4 \end{aligned}$$

Die Dreiecksform kann ohne Variation des Algorithmus nicht mehr erreicht werden, obwohl das Gleichungssystem eine eindeutige Lösung besitzt. In einem solchen Fall müssen Gleichungen getauscht werden.

4.6.2 Determinanten

Mit Hilfe des Laplaceschen Entwicklungssatzes können n-zeilige Determinanten mit $n \in \mathrm{IN}^{\geq 2}$ aus einer Schar 2-zeiliger Unterdeterminanten entwickelt werden. Es gilt für die Entwicklung der Determinante nach den Elementen einer Zeile z:

$$\det \mathbf{A} = \sum_{s=1}^{n} \left((-1)^{z+s} a_{z,s} \cdot \det \mathbf{A}_{z,s} \right)$$

$$\begin{aligned} z &= \text{Zeile der Determinante } \mathbf{A}. \\ s &= \text{Spalte der Determinante } \mathbf{A}. \\ n &= \text{Anzahl der Zeilen bzw. Spalten der Determinante } \mathbf{A}. \\ \det \mathbf{A}_{z,s} &= \text{Unterdeterminante, die durch durch Streichen der Zeile } z \text{ und} \\ &\quad \text{der Spalte } s \text{ gebildet aus } \mathbf{A} \text{ gebildet wird.} \end{aligned} \tag{4.6-6}$$

$(-1)^{z+s} a_{z,s} \cdot \det \mathbf{A}_{z,s}$ ist als algebraisches Komplement der Zeile z und der Spalte s definiert.

Besitzen die Unterdeterminanten $\mathbf{A}_{z,s}$ immer noch mehr als 2 Zeilen, so muss Gleichung (4.6-6) auch auf die Unterdeterminanten angewandt werden. Gelangt man schließlich zu zweizeiligen Unterdeterminante, so können diese leicht berechnet werden:

$$\det \mathbf{A}' = \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1} \cdot a_{2,2} - a_{1,2} \cdot a_{2,1} \tag{4.6-7}$$

Der Algorithmus lässt sich in einer Hochsprache wie Object Pascal über eine Rekursion umsetzen (siehe Abbildung 4.6-1). Die Variable *zaehler_i* enthält die Anzahl, wie oft sich die Funktion *det_e* selber wieder aufruft. Tabelle 4.6-1 enthält die Anzahl dieser Aufrufe in Abhängigkeit von der Zeilenanzahl der zu berechnenden Determinante. Bei 11 Zeilen erfordert die Determinante über 28 Millionen dieser Rekursionsschritte. Eine solche Determinante kann von einem AMD Athlon mit 1 GHz Taktfrequenz in etwa 25 Sekunden berechnet werden. Da sich mit jeder zusätzlichen Determinantenzeile die Anzahl der Funktionsaufrufe um den gleichen Faktor erhöht, wie die neue Determinante Zeilen hat, sind Determinantenberechnungen nach Laplace mit wesentlich mehr als 11 Zeilen auch für erheblich leistungsfähigere Rechner ein nicht in akzeptabler Zeit lösbares Problem. Die Lösung liegt in einer Umformung der Determinante in die Dreiecksform. Bei der Dreiecksform einer Matrix wie der Determinantenmatrix sind alle Elemente unterhalb der Hauptdiagonale Null. Zu diesem Zweck eignet sich der im vorherigen Abschnitt beschriebene Gauß-Algorithmus. Aus der Matrixdarstellung des linearen Gleichungssystem muss lediglich die Spalte mit den Lösungen gestrichen werden. Statt der Koeffizienten des linearen Gleichungssystems werden die Elemente der Determinante in die Matrix eingesetzt (siehe Gleichung (4.6-2)). Nach den Determinantengesetzen ändert die Addition eines Vielfachen der Elemente einer Reihe zu einer parallelen Reihe den Wert einer Determinante nicht. Durch die Diagonalisierung zur Dreiecksform bleibt also der Wert der Determinante gleich. Ist das Tauschen einer Zeile erforderlich, findet nach den Determinantengesetzen jedes mal eine Umkehr des Vorzeichens statt. Die Entwicklung nach Laplace vereinfacht sich für eine Determinante mit mehr als zwei Zeilen, die sich in Dreiecksform befindet, zu:

$$\det \mathbf{A} \quad = \quad \left(a_{n-1} a_{n,n} - a_{n,n-1} a_{n-1,n}\right) \cdot \prod_{i=1}^{n-2} a_{i,i} \tag{4.6-8}$$

Zur Berechnung dieses Produktes werden so viele Multiplikationen benötigt, wie die Determinante Zeilen hat. Hinzu kommt eine einzelne Subtraktion. Der größte Aufwand ist die erforderliche Diagonalisierung, welche aber den Computer wesentlich weniger fordert, als eine vollständige Entwicklung nach Laplace.

Tabelle 4.6-1: Die Anzahl der Funktionsaufrufe von det_e = $Anz_{(n)}$ in Abhängigkeit von der Zeilenzahl n der zu berechnenden Determinante

n	$Anz_{(n)}$	$Anz_{(n)}/Anz_{(n-1)}$
2	1	
3	4	4
4	17	4,25
5	86	5,06
6	517	6,01
7	3620	7,00
8	28961	8,00
9	260650	9,00
10	2606501	10,00
11	28671512	11,00

```
Unit Determinante;

Interface

Type tvektor_da_e = Array Of Extended;
     tmatrix_da_e = Array Of TVektor_da_e;

Function Det_e(Var matrix_da_e : tmatrix_da_e;
               Var zaehler_i   : Integer) : Extended;

Implementation

Function Det_e(Var matrix_da_e : tmatrix_da_e;
               Var zaehler_i   : Integer) : Extended;

Var i, j, k          : Integer;
    UnterMatrix_da_e : tmatrix_da_e;
    Vorzeichen_i     : Integer;

Begin
  Result := 0;
  Inc(zaehler_i);
  If Length(Matrix_da_e) =  2 Then
     Result :=   Matrix_da_e[0, 0] * Matrix_da_e[1, 1]
               - Matrix_da_e[0, 1] * Matrix_da_e[1, 0]
     Else
     Begin
       Vorzeichen_i := 1;
       //Speicher für Unterdeterminanten reservieren
       SetLength(UnterMatrix_da_e, Length(Matrix_da_e) - 1, Length(Matrix_da_e[0]) - 1);
       //Berechnung der Determinante aus den algebraischen Komplementen
       For i := 0 To High(Matrix_da_e) Do
           Begin
             //Bildung der Unterdeteminante in RedMatrix_da_e
             For j := 0 To i - 1 Do
                 For k := 1 To High(Matrix_da_e) Do
                     UnterMatrix_da_e[j, k - 1] := Matrix_da_e[j, k];
             For j := i + 1 To High(Matrix_da_e) Do
                 For k := 1 To High(Matrix_da_e) Do
                     UnterMatrix_da_e[j - 1, k - 1] := Matrix_da_e[j, k];
             //Addition des Produktes aus algebraischen Komplement und dem
             //entsprechenden Element aus der ersten Spalte
             Result :=   Result
                       + Vorzeichen_i * Matrix_da_e[i, 0] * det_e(UnterMatrix_da_e, zaehler_i);
             Vorzeichen_i := -Vorzeichen_i
           End;
       //Speicher für Unterdeterminanten wieder freigeben
       SetLength(UnterMatrix_da_e, 0, 0)
  End
End;

end.
```

Abbildung 4.6-1: Object Pascal Quellcode zur Berechnung von Determinanten

4.6.3 Die Inversion von Matrizen

Die Inversion einer $n \times n$-Matrix sollte nicht über ihre algebraischen Komplemente durchgeführt werden:

$$\begin{aligned} \mathbf{A}^{-1} &= \frac{1}{\det \mathbf{A}} \overline{\mathbf{A}}^{\mathrm{T}} \\ &\text{mit} \\ \overline{\mathbf{A}}^{\mathrm{T}} &= \text{transponierte Matrix der algebraischen Komplemente von A.} \end{aligned} \tag{4.6-9}$$

Bei diesem Algorithmus müssen für eine Matrixinversion n+1 Determinanten mit der entsprechenden Zahl an Gaußschen Diagonalisierungen berechnet werden. Besser eignet sich die direkte Anwendung des Gauß-Algorithmus. Es gilt:

$$\begin{aligned} \mathbf{A} \cdot \mathbf{A}^{-1} &= \mathbf{E} \\ &\text{mit} \\ \mathbf{A} &= \text{quadratische, nicht singuläre Matrix (Zeilenzahl} = \text{Spaltenzahl} \\ &\quad \text{sowie Determinante } \det(\mathbf{A}) = 0), \\ \mathbf{A}^{-1} &= \text{inverse Matrix und} \\ \mathbf{E} &= \text{Einheitsmatrix (Hauptdiagonalenelemente 1, alle anderern Null).} \end{aligned} \tag{4.6-10}$$

Bei Gleichung (4.6-10) handelt sich im Prinzip um n lineare Gleichungssysteme mit den Spaltenvektoren der inversen Matrix als Unabhängigenvektoren und den Spaltenvektoren der Einheitsmatrix als Lösungsvektoren. Im Unterschied zu Gleichung (4.6-2) ist hier statt des Unabhängigenvektors eine Unabhängigenmatrix zu finden und der Lösungsvektor ist ebenfalls zur Matrix, der Einheitsmatrix, erweitert worden. Die Lösung erfolgt analog zum Gaußalgorithmus in Abschnitt 4.6.1. Da die Koeffizientenmatrix **A** immer gleich bleibt, braucht sie nur einmal diagonalisiert werden. Die Anzahl der Rücksubstitutionsschritte ist allerdings genau so groß, als wenn alle n linearen Gleichungssysteme einzeln gelöst werden. In Analogie zu Gleichung (4.6-4) gilt:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ 0 & a'_{2,2} & a'_{2,3} & \cdots & a'_{2,n} \\ 0 & 0 & a''_{2,3} & \cdots & a''_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & a^{(n-1)}_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,n} \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,n} \end{pmatrix} = \begin{pmatrix} c'_{1,1} & c'_{1,2} & c'_{1,3} & \cdots & c'_{1,n} \\ c''_{2,1} & c''_{2,2} & c''_{2,3} & \cdots & c''_{2,n} \\ c'''_{3,1} & c'''_{3,2} & c'''_{3,3} & \cdots & c'''_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^{(n-1)}_{n,1} & c^{(n-1)}_{n,1} & c^{(n-1)}_{n,1} & \cdots & c^{(n-1)}_{n,1} \end{pmatrix}$$

bzw. im Falle der Matrixinversion ohne den Tausch von Gleichungen: (4.6-11)

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ 0 & a'_{2,2} & a'_{2,3} & \cdots & a'_{2,n} \\ 0 & 0 & a''_{2,3} & \cdots & a''_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & a^{(n-1)}_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,n} \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c''_{2,1} & 1 & 0 & \cdots & 0 \\ c'''_{3,1} & c'''_{3,2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^{(n-1)}_{n,1} & c^{(n-1)}_{n,1} & c^{(n-1)}_{n,1} & \cdots & 1 \end{pmatrix}.$$

Müssen bei der Diagonalisierung keine Gleichungen getauscht werden, wird die Berechnung der Lösungsmatrix dadurch erleichtert, dass die Hauptdiagonale nur aus Einsen bestehen kann und sich oberhalb der Hauptdiagonalen nur Nullen befinden. Dieser Sachverhalt hat seine Ursache darin, dass die Lösungsmatrix aus der Einheitsmatrix berechnet worden ist. Löst man Gleichung (4.6-11) über das Rücksubstitutionsverfahren nach den Unabhängigen in der Unabhängigenmatrix auf (siehe Abschnitt 4.6.1), so erhält man die inverse Matrix zur Koeffizientenmatrix.

4.7 Lineare Regression

4.7.1 Multiple und multivariate Regression [14]

Messwerte werden in der Analytik durch einen Satz von Variablen beschrieben. Bei einer Regressionsanalyse solcher Messwertsammlungen geht man davon aus, dass man die Variablen in zwei verschiedene Gruppen unterteilen kann. Zum einen gibt es sogenannte unabhängige Variablen, die auch **Regressoren** genannt werden. Diese sollten frei von analytischen Messfehlern erfasst worden sein und es sollte kein funktionaler Zusammenhang zwischen diesen Variablen bestehen. Die Regressoren stellen beispielsweise die Bedingungen dar, unter denen eine Messung durchgeführt worden ist (Geräteeinstellungen, Konzentrationen, etc.). Die andere Gruppe von Variablen sind die **Regressanden**. Die Regressanden sind von den Regressoren abhängige Variablen. Es handelt sich hierbei um die analytisch auszuwertenden Messgrößen (beispielsweise Intensitäten, Extinktionen, Spektraldichten etc.). Regressoren werden auch als Einstellgrößen und die Regressanden als Zielgrößen bezeichnet. Bei einer Regressionsanalyse soll der funktionale Zusammenhang zwischen Regressanden und Regressoren ermittelt werden. Besteht zwischen diesen Variablengruppen ein linearer Zusammenhang, so kann dieser mit Hilfe der multiplen linearen Regression ermittelt werden. Hierbei wird davon ausgegangen, dass jeder Regressand y folgendermaßen beschrieben werden kann:

$$f_{(x_1, x_2, \ldots, x_n)} = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

mit

β_0 = Absolutglied,

$\beta_{0,1\ldots n}$ = Regressionskoeffizienten, (4.7-1)

$x_{1,2\ldots n}$ = Regressoren,

$f_{(x_1, x_2, \ldots, x_n)} = y$ = Regressand und,

ε = Messfehler.

Der Regressionsalgorithmus dient zur Abschätzung der Regressionskoeffizienten. Es wird ein Schätzwert für jeden Regressionskoeffizienten gesucht, so dass mit diesen Schätzwerten der in Gleichung (4.7-1) beschriebene Zusammenhang zwischen Regressoren und Regressanden

am besten beschrieben wird. Zu diesem Zweck stellt man die Regressanden als Vektor dar und die zugehörigen Regressoren als erweiterte Regressormatrix. Für einen Datensatz aus m verschiedenen Regressor-Regressanden-Kombinationen auf Basis m verschiedener Messungen gilt:

$$\vec{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \tag{4.7-2}$$

$\vec{\mathbf{y}}$ = Regressandenvektor
$\mathbf{X}$ = Regressormatrix

Die Variablen $y_1, y_2, \ldots, y_m$ sind die Werte des Regressanden aus den Messungen 1 bis m und $x_{i,j}$ ist von der i-ten Messung der j-te Regressor. Für jede Einzelmessung gilt:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n} + \varepsilon_i \quad \text{mit } i = 1, \ldots, m \tag{4.7-3}$$

β_i und ε_i lassen sich zu Koeffizentenvektor und Fehlervektor zusammenfassen:

$$\vec{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}; \quad \vec{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}. \tag{4.7-4}$$

$\vec{\boldsymbol{\beta}}$ = Koeffizientenvektor.
$\vec{\boldsymbol{\varepsilon}}$ = Fehlervektor.

Die Gleichungen für alle Einzelmessungen lassen sich jetzt in Matrixschreibweise darstellen:

$$\vec{\mathbf{y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}. \tag{4.7-5}$$

Wird an Stelle des nicht bekannten Koeffizientenvektors $\vec{\boldsymbol{\beta}}$ einen beliebigen Vektor $\vec{\mathbf{b}}$ mit den Komponenten $b_0, b_1, \ldots b_n$ als Schätzung eingesetzt, so erhält man nach Gleichung (4.7-5) den Regressandenvektor

$$\vec{\mathbf{y}}' = \mathbf{X}\vec{\mathbf{b}}. \tag{4.7-6}$$

Die Schätzung ist um so besser, je geringer die Diskrepanz $\vec{\boldsymbol{\varepsilon}}$ zwischen $\vec{\mathbf{y}}$ und $\vec{\mathbf{y}}'$ ist. Ein Maß dieser Diskrepanz ist die Länge des Differenzvektors oder dessen Quadrat. Letzteres ist

die bequemer handhabbare Residuenquadratsumme zur Schätzung von $\vec{\mathbf{b}}$ (Residue = Differenz zwischen dem Messwert eines Regressanden und seinem Schätzwert):

$$\left\|\vec{\mathbf{y}} - \vec{\mathbf{y}}'\right\|^2 = \sum_{i=1}^{m} \left(y_i - y_i'\right)^2 = \text{Residuenquadratsumme} = Q. \tag{4.7-7}$$

Die Berechnung der optimalen Schätzung von $\vec{\mathbf{b}}$ ist nach Gleichung (4.7-6) und (4.7-7) folgendes Minimierungsproblem:

$$\min\left\|\vec{\mathbf{y}} - \vec{\mathbf{y}}'\right\|^2 = \min_{\vec{\mathbf{b}}}\left\|\vec{\mathbf{y}} - \mathbf{X}\vec{\mathbf{b}}\right\|^2. \tag{4.7-8}$$

Dessen Lösung lautet:

$$\begin{aligned} \vec{\mathbf{b}}^{\text{opt}} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\vec{\mathbf{y}} \\ &\text{mit} \\ \mathbf{X}^{\mathrm{T}} &= \text{transponierte Matrix } \mathbf{X}. \end{aligned} \tag{4.7-9}$$

Die Matrixinversion in Gleichung (4.7-9) kann über den Gauß-Algorithmus (siehe Abschnitt 4.6.3) durchgeführt werden. Allerdings sind nach der Diagonalisierung sehr viele Rücksubstitutionsschritte nötig. Daher ist es sinnvoll Gleichung (4.7-9) so umzustellen, dass sich ein einzelnes lineares Gleichungssystem mit $\vec{\mathbf{b}}$ als Lösung ergibt:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\vec{\mathbf{b}}^{\text{opt}} = \mathbf{X}^{\mathrm{T}}\vec{\mathbf{y}}. \tag{4.7-10}$$

$\vec{\mathbf{b}}$ ist die optimale Schätzung für die Regressionskoeffizienten in Gleichung (4.7-1). Diese Schätzung bezieht sich auf einen Regressanden und heißt **multiple lineare Regression**. Die optimale Schätzung des Regressanden lautet:

$$\begin{aligned} y' &= b_0^{\text{opt}} + b_1^{\text{opt}} x_1 + b_2^{\text{opt}} x_2 + \cdots + b_n^{\text{opt}} x_n \\ &\text{mit} \\ b_{0,1,2,\ldots,\mathrm{n}}^{\text{opt}} &= \text{optimale Schätzung der Regressionskoeffizienten} \\ &= \text{Elemente von } \vec{\mathbf{b}}. \end{aligned} \tag{4.7-11}$$

Ein Spezialfall der multiplen linearen Regression ist die **einfach lineare Regression**, bei der nur ein Regressor vorliegt und Gleichung (4.7-11) zu

$$y' = b_0^{\text{opt}} + b_1^{\text{opt}} x_1 = b_0^{\text{opt}} + b_1^{\text{opt}} x \tag{4.7-12}$$

wird. Die einfache lineare Regression wird in vielen Büchern gesondert abgehandelt, da für die Abschätzung der optimalen Regressionskoeffizienten auch alternativ Formeln zur Verfügung stehen, die ohne Matrixoperationen auskommen. Diese Formeln lassen sich mittels Taschenrechner oder einfacher Tabellenkalkulationen leichter handhaben.

Sind mehrere Regressanden abzuschätzen, so muss der Vorgang für jeden Regressanden wiederholt werden, so dass für jeden Regressanden ein anderer Satz an Schätzwerten für die Regressionskoeffizienten erhalten wird. Eine solche Regressanden-Mehrfachabschätzung wird als **multivariate lineare Regression** bezeichnet. Hierbei wird Gleichung (4.7-10) zu

$$\begin{aligned}
&\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{B}^{\mathrm{opt}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \\
&\text{wobei} \\
&\mathbf{B}^{\mathrm{opt}} = \text{Lösungsmatrix deren Spalten die Regressionskoeffizienten-vektoren bilden und} \\
&\mathbf{Y} = \text{Regressandenmatrix, deren Spalten die zu den Regressionskoeffizienten gehörenden Regressandenmesswerte beinhalten.}
\end{aligned} \tag{4.7-13}$$

Gleichung (4.7-13) kann ebenfalls über eine direkte Diagonalisierung nach Gauß berechnet werden, wobei $\mathbf{B}^{\mathrm{opt}}$ als Unabhängigenmatrix eines erweiterten bzw. multiplen linearen Gleichungssystem anzusehen ist. Dessen Lösung erfolgt analog zu dem Verfahren, welches schon bei der Matrixinversion angewandt worden ist (vergleiche Abschnitt 4.6.1 und 4.6.3, insbesondere Gleichung (4.6-11) obere Version).

4.7.2 Das multiple Bestimmtheitsmaß

Das multiple Bestimmtheitsmaß ist ein Maß für die Güte der Anpassung der Messwerte durch die Regressionsgleichung (4.7-11). Es wird aus dem Quotienten zweier Quadratsummen gebildet: [22]

$$\begin{aligned}
&R^2 = \frac{Q_R}{Q_y} \\
&\text{mit} \\
&Q_R = \sum_{i=1}^{m}\left((y_i' - \bar{y})^2\right), \\
&Q_y = \sum_{i=1}^{m}\left((y_i - \bar{y})^2\right), \\
&\bar{y} = \text{Mittelwert aller Messwerte des Regressanden } y \text{ und} \\
&R^2 = \text{multiples Bestimmtheitsmaß bzw. R = Korrelationskoeffizient.}
\end{aligned} \tag{4.7-14}$$

Q_R ist die Quadratsumme aus allen Schätzwerten y' abzüglich des Mittelwertes der Messwerte für den Regressor y. Bei Q_y handelt es sich hingegen um die Quadratsumme der Differenzen aus jedem Messwert y und dem Mittelwert aller Messwerte. Der Quotient kann als Verhältnis der durch die Regressionsgerade erklärten Variabilität von y zur Gesamtvariabilität aufgefasst werden.

4.7.3 Das innere Bestimmtheitsmaß

Für die Ausführung einer multivariaten linearen Regression ist es günstig, dass die Regressoren keine Abhängigkeiten untereinander aufweisen. Insbesondere muss eine exakte lineare Abhängigkeit vermieden werden, da ansonsten das der linearen Regression zu Grunde liegende Extremwertproblem mathematisch nicht lösbar ist. Oft kann die Abhängigkeit von Regressoren in einem mehr oder weniger hohen Maße durch einen linearen Zusammenhang beschrieben werden. Dieses Phänomen wird Multikollinearität oder auch Interkorrelation genannt. Bei hoher Interkorrelation treten schwerwiegende Probleme bei der Regressionsrechnung auf. Diese Probleme reichen vom Überlauf der Variablen bei Computerberechungen bis hin zu stark voneinander abweichenden Schätzwerten der Regressionskoeffizienten ($\vec{\mathbf{b}}$) bei verschiedenen Stichproben und Teilansätzen mit reduzierter Anzahl an Regressoren. Das Bestimmtheitsmaß eines Regressors ist ein wichtiges Kriterium zur Quantifizierung der Interkorrelation. Es gibt an, welcher Anteil der Variabilität eines Regressors durch die übrigen Regressoren beschrieben werden kann. Es handelt sich also um das multiple Bestimmtheitsmaß für eine Regressionsrechnung, bei dem die Messwerte des betreffenden Regressors aus der Regressormatrix $\mathbf{X}$ entfernt worden sind und die Regressanden in $\vec{\mathbf{y}}$ ersetzen. Statt über eine erneute Regressionsrechnung lässt es sich auch über

$$
\begin{aligned}
B_i &= 1-\frac{1}{\chi_{i,i}\cdot Q_{x_i}} \\
&\text{mit} \\
Q_{x_i} &= \sum_{j=1}^{m}\left(x_{j,i}-\bar{x}_i\right)^2, \\
\bar{x}_i &= \text{Mittelwert aller Werte des Regressanden } x_i, \\
\chi_{i,i} &= \text{das } i\text{-te Diagonalenelement der Matrix } \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \text{ und} \\
B_i &= \text{inneres Bestimmtheitsmaß für den Regressor } i.
\end{aligned}
\tag{4.7-15}
$$

berechnen.[22] Die für $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ erforderliche Matrixinversion muss für alle Regressoren nur einmal durchgeführt werden. Die Berechung von B_i über die Regression umgeht zwar diese Inversion, muss aber für jeden Regressor wiederholt werden.

4.7.4 Vertrauens- und Vorhersageintervalle

Voraussetzung für die Bestimmung von Vertrauens- und Vorhersageintervallen ist, dass

1. der wahre Wert des Fehlervektors $\vec{\boldsymbol{\varepsilon}}$ der Nullvektor ist und

2. die Streuung der Elemente von $\vec{\boldsymbol{\varepsilon}}$ normalverteilt ist.

Die Messung darf demnach nicht durch systematische Fehler beeinflusst worden sein. Die Residuen bzw. die Elemente des Fehlervektors müssen einen statistischen Fehler widerspiegeln, der mit Hilfe einer Gauß-Verteilungskurve beschrieben werden kann, sobald unendlich viele Messungen vorliegen. Nur unter diesen Voraussetzungen gilt für das Vertrauensintervall der Schätzung des wahren Wertes eines Regressanden: [22]

$$y'_k \pm t_{m-n-1;\,p} \cdot \sqrt{\frac{Q}{m-n-1} \cdot \left(\vec{\mathbf{x}}_k^{\mathrm{T}} \cdot \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \cdot \vec{\mathbf{x}}_k\right)}$$

mit

y'_k = Schätzwert des Regressanden y auf Basis der Regressorwerte der k - ten Messung, (4.7-16)

$\vec{\mathbf{x}}_{\mathrm{k}}^{\mathrm{T}}$ = $\begin{pmatrix}1 & x_{k,1} & x_{k,2} & \cdots & x_{k,2}\end{pmatrix}$ = k - te Zeile von $\mathbf{X}$,

$\vec{\mathbf{x}}_k$ = Regressandenvektor der k - ten Messung ($\vec{\mathbf{x}}_{\mathrm{k}}^{\mathrm{T}}$ ist $\vec{\mathbf{x}}_k$ transponiert)

$t_{m-n-1;\,p}$ = und Student - Faktor, zweiseitige Fragestellung, für das

= Wahrscheinlichkeitsniveau p in %.

Der wahre Wert des Regressors liegt für $\vec{\mathbf{x}}_k$ innerhalb dieses Intervalls. Es ist auch möglich statt eines schon für die Regressionsrechnung verwendeten Satzes von Regressorwerten einen neuen Satz in die Gleichung einzusetzen. Es sollte jedoch darauf geachtet werden, dass nur Regressorwerte verwendet werden, die im Gültigkeitsbereich des linearen Modells liegen. Setzt man eine Wertekombination ein, die den zur Regressionsrechnung verwendeten Werten nicht ähnelt, kann diese Voraussetzung unter Umständen nicht mehr erfüllt werden.

Bei einer erneuten Messung liegt der Wert für den Regressanden mit einer Wahrscheinlichkeit von p innerhalb dieses Intervalls: [22]

$$y' \pm t_{m-n-1;\,p} \cdot \sqrt{\frac{Q}{m-n-1} \cdot \left(\frac{1}{w} + \vec{\mathbf{x}}^{\mathrm{T}} \cdot \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \cdot \vec{\mathbf{x}}\right)}$$

mit

y' = Schätzwert des Regressanden y auf Basis der Regressorwerte, die durch den Regressorvektor $\vec{\mathbf{x}}$ repräsentiert werden,

$\vec{\mathbf{x}}$ = $\begin{pmatrix}1\\x_1\\x_2\\\vdots\\x_n\end{pmatrix}$ = Regressorvektor, (4.7-17)

$\vec{\mathbf{x}}^{\mathrm{T}}$ = transponierter Regressorvektor und

w = Anzahl an Wiederholmessungen zur Mittelung des Regressorwertes.

Auch hier muss darauf geachtet werden, dass nur Werte in die Regressorvariablen eingesetzt werden, für die das lineare Modell gültig ist.

4.7.5 Variablenreduktion in der multiplen linearen Regression

Bei der Auswertung von Messreihen mit Hilfe der multivariaten Regression stellt sich häufig die Frage, welche Regressoren überhaupt zur Abschätzung der Regressionskoeffizienten herangezogen werden sollen. Ziel einer Variablenreduktion ist es, aus dem **Totalansatz** mit der maximalen Anzahl an Regressoren denjenigen Teilansatz (**Subset**) zu ermitteln, welcher eine optimale Abschätzung der Regressionskoeffizienten erlaubt. Eine solche Abschätzung kann als um so besser angesehen werden, je kleiner die Residuenvarianz ist. Bei der Residuenvarianz handelt es sich um den Quotienten aus Fehlerquadratsumme und Anzahl der Freiheitsgrade:

$$\mathrm{VAR}_{(\vec{\mathbf{b}})} = \frac{Q_{(\vec{\mathbf{b}})}}{n-m-1} \tag{4.7-18}$$

mit

$\mathrm{VAR}_{(\vec{\mathbf{b}})}$ = Residuenvarianz bezogen auf den über $\vec{\mathbf{b}}$ definierten Regressionskoeffzientensatz.

$Q_{(\vec{\mathbf{b}})}$ = Residuenquadratsumme errechnet mit dem über $\vec{\mathbf{b}}$ definierten Regressionskoeffizientensatz.

Eine Optimierung nur anhand der Residuenvarianz wird um so schwieriger, je größer die Interkorrelation ist. Es existieren in diesem Fall oft mehrere Subsets mit ähnlicher Varianz. In einem solchen Fall müssen neben Ergebnissen aus statistischen Berechnungen weitere Informationen verwendet werden. Zum einen sollte man den sich mit jedem zusätzlichen Regressor erhöhenden Rechenaufwand in den Auswahlprozess einbeziehen. Es sollte ein Optimum zwischen Zahl und Aussagekraft der Regressoren gefunden werden. Zudem sollten Regressoren eliminiert werden, die mit hoher Wahrscheinlichkeit nicht zum wahren Modell gehören, wobei auch fachspezifische Informationen in Betracht zu ziehen sind. [22]

4.8 Die Quantifizierung von chromatographischen Signalen

Die Quantifizierung chromatographischer Signale kann mit Hilfe eines Computers in drei Schritten vorgenommen werden:

1. Bestimmung der Höhe der Basislinie,

2. Ermittlung der Grenzen, innerhalb derer sich ein chromatographisches Signal befindet und

3. Berechnung der relevanten Daten für das chromatographische Signal wie Fläche, Höhe, Halbwertsbreite (Breite auf halber Höhe) oder auch ein Maß für die Deformation (Fronting oder Tailing).

Für die Bestimmung der Höhe der Basislinie ist es günstig, wenn die Basislinie keinen Drift aufweist. In diesem Fall kann angenommen werden, dass das Null-Signal immer dem gleichen Wert entspricht. Liegt das Chromatogramm in Form eines Daten-Arrays vor, so muss lediglich der Mittelwert aus den Elementen eines signalfreien Bereiches berechnet werden. Dieser Mittelwert wird auch Blindwert genannt.

Bei nicht driftender Basislinie kann ein Schwellenwert herangezogen werden, um zu bestimmen, wann ein Signal beginnt oder endet. Hierbei vergleicht man jedes Element des Datenarrays mit der Summe aus Blind- und Schwellenwert. Wird dieser Wert überschritten, so beginnt ein Signal, wird er unterschritten, so endet ein Signal (siehe Abbildung 4.8-1). Überlappen Signale, so muss die Signalerkennung anhand eines Schwellenwertes um eine Minimumdiagnose ergänzt werden. Tritt ein lokales Minimum auf, so wird dieses als Ende des ersten und Anfang eines zweiten Signals gewertet (siehe Abbildung 4.8-2).

Für die Signalflächenbestimmung werden alle Elemente des Datenarrays, die sich innerhalb der Grenzen des Signals befinden, unter Subtraktion des Blindwertes (Blindwertbereinigung) aufsummiert. Wenn ein sehr kleines Signal bestimmt werden soll, welches von einem sehr großen Signal überlappt wird, so kann eine Basislinienkorrektur nicht mehr anhand eines einzelnen Blindwertes erfolgen. Stattdessen muss eine Tangente an das kleinere Signal angelegt werden um diese zu subtrahieren (siehe Abbildung 4.8-3). [23]

Die Signalhöhe entspricht dem blindwertbereinigten Signalmaximum. Zur Bestimmung der Halbwertsbreite ermittelt man für die ansteigende und die abfallende Signalflanke jeweils die zwei benachbarten Arrayelemente, welche jeweils einen blindwertbereinigten Messwert unter und über der halben Signalhöhe enthalten. Jetzt muss jedem Arrayelement eine Retentionszeit zugeordnet werden. Wenn durch die ausgewählten Punktepaare jeweils eine Gerade gezogen wird, so kann die Retentionszeit bei Über- und Unterschreitung der halben Signalhöhe in-

terpoliert werden. Abbildung 4.8-4 verdeutlicht die Vorgehensweise. Der Betrag des Quotienten aus den Steigungen der Geraden ist ein gutes Maß für die Abweichung des chromatographischen Signals von der idealen Gauß-Form. Dieser Wert entspricht meistens dem Betrag des Quotienten aus maximaler und minimaler Steigung, welcher ebenfalls ein Maß für die Signaldeformation bezüglich der Gaußkurve ist. Manchmal ist es sinnvoll den Offset eines Signals mitzubestimmen. Der Offset ist die Abweichung des Signalsockels von der Höhe der Basislinie. Zu diesem Zweck wird bis zu einer bestimmten Anzahl von Arrayelementen vor dem Signalanfang das absolute Minimum bestimmt und dieser Datenpunkt eventuell geglättet. Der so erhaltene Wert ist der Offset.

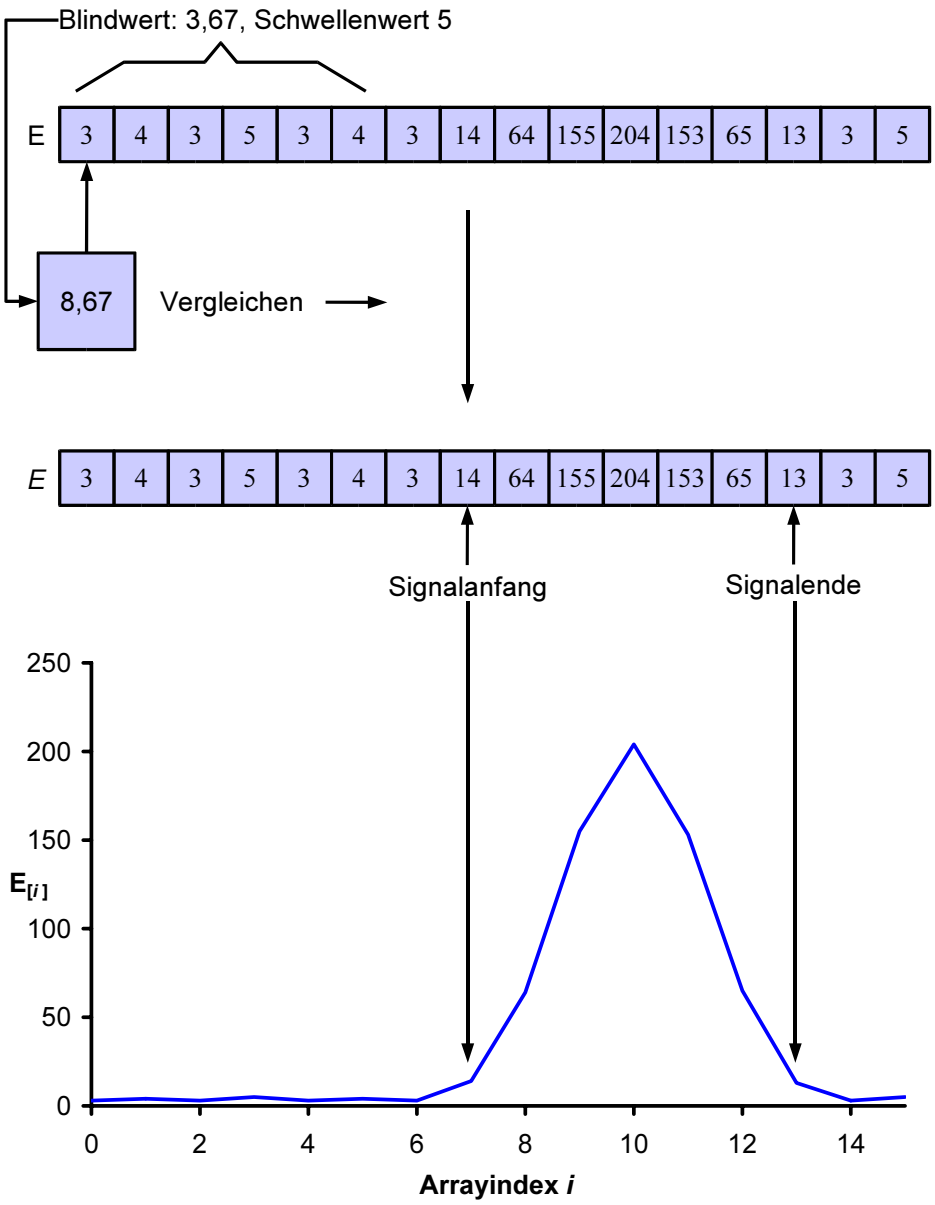

Abbildung 4.8-1: Blockschema des Signalerkennunsalgorithmus mit 5 als Schwellenwert

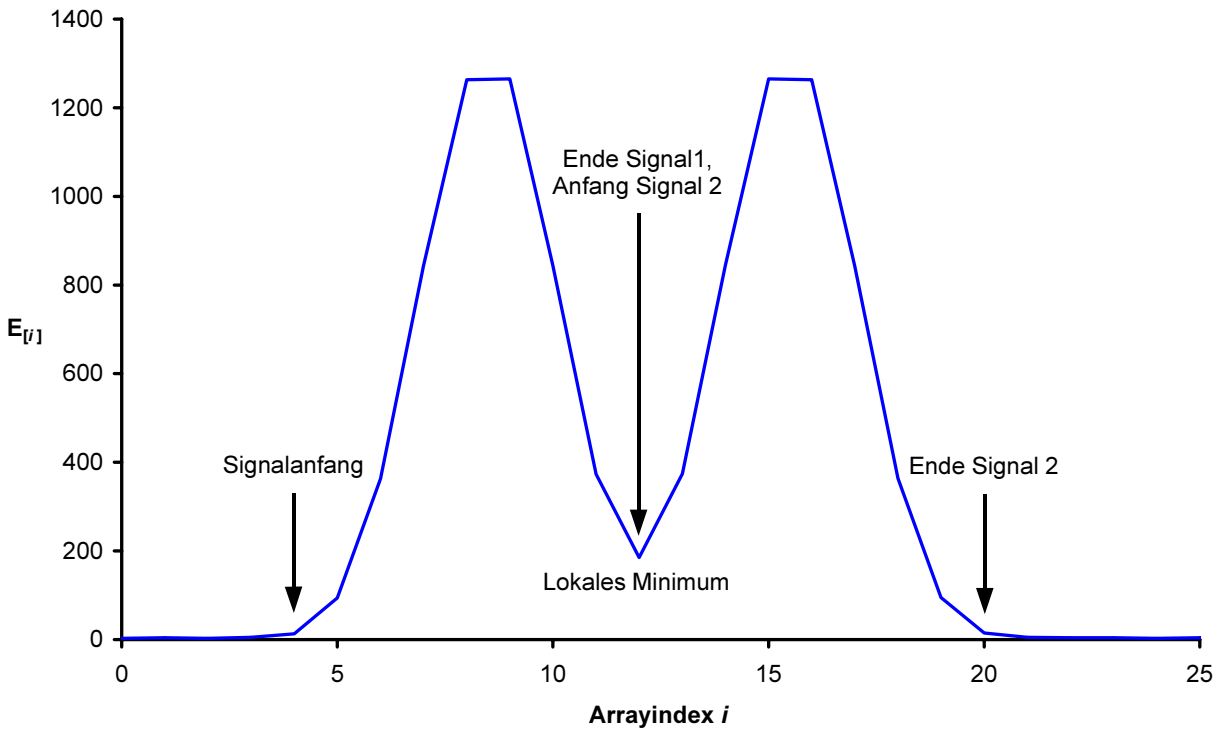

Abbildung 4.8-2: Signalerkennung bei nicht ganz aufgelösten Chromatogrammen

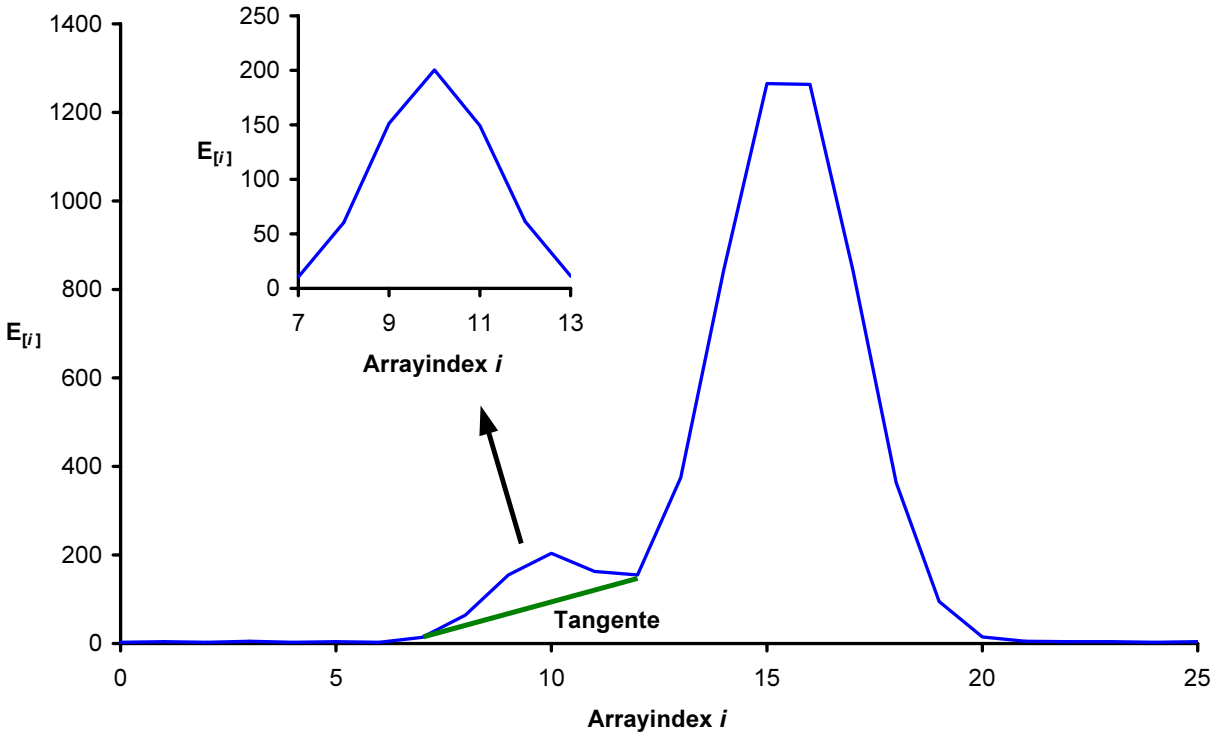

Abbildung 4.8-3: Basislinienkorrektur bei zwei überlappenden unterschiedlich hohen Signalen durch Subtraktion einer Tangente

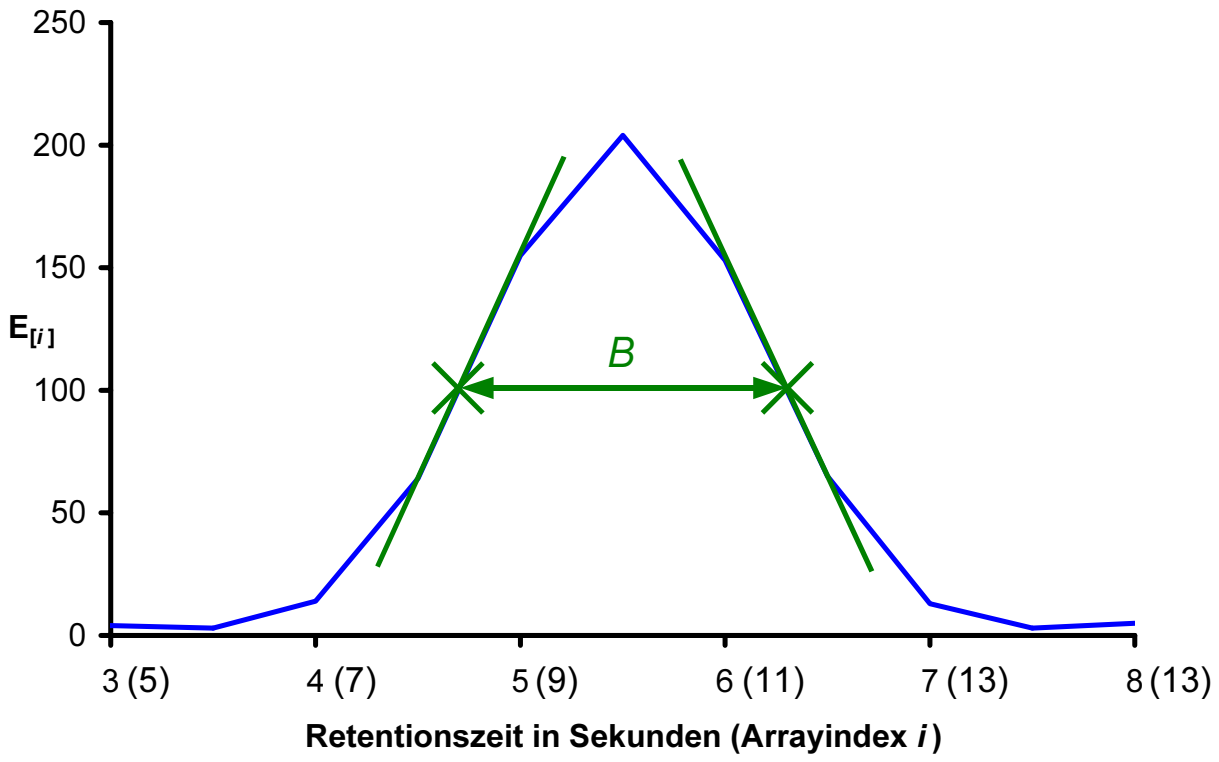

Abbildung 4.8-4: Interpolation der Halbwertsbreite B durch Anlegen zweier Geraden durch die Datenpunkte, die der halben Signalhöhe am nächsten kommen

4.9 Chromatographische Mehrkanalmessungen

4.9.1 Das Angleichen von Basislinien

Statt nur ein chromatographisches Signal aufzuzeichnen, können mit einem Computer auch mehrere chromatographische Signale gleichzeitig erfasst werden. Dieses Vorgehen ist beispielsweise beim Einsatz eines Polychromators zur Messwerterfassung sinnvoll. Man erhält über eine solche Mehrkanalmessung mehrere Einzelchromatogramme, die z. B. bei verschiedenen Wellenlängen aufgezeichnet worden sind. Solche Einzelchromatogramme können erst dann gut miteinander verglichen werden, wenn ihre Basislinien alle auf der Nulllinie liegen würden. Zu diesem Zweck muss analog zum vorherigen Abschnitt für jedes Einzelchromatogramm die Höhe der Basislinie bestimmt werden. Diese Blindwerte werden von jedem Arrayelement des zugehörigen Einzelchromatogrammes abgezogen. Man erhält auf diese Weise blindwertbereinigte Einzelchromatogramme mit angeglichenen Basislinien (siehe Abbildung 4.9-1).

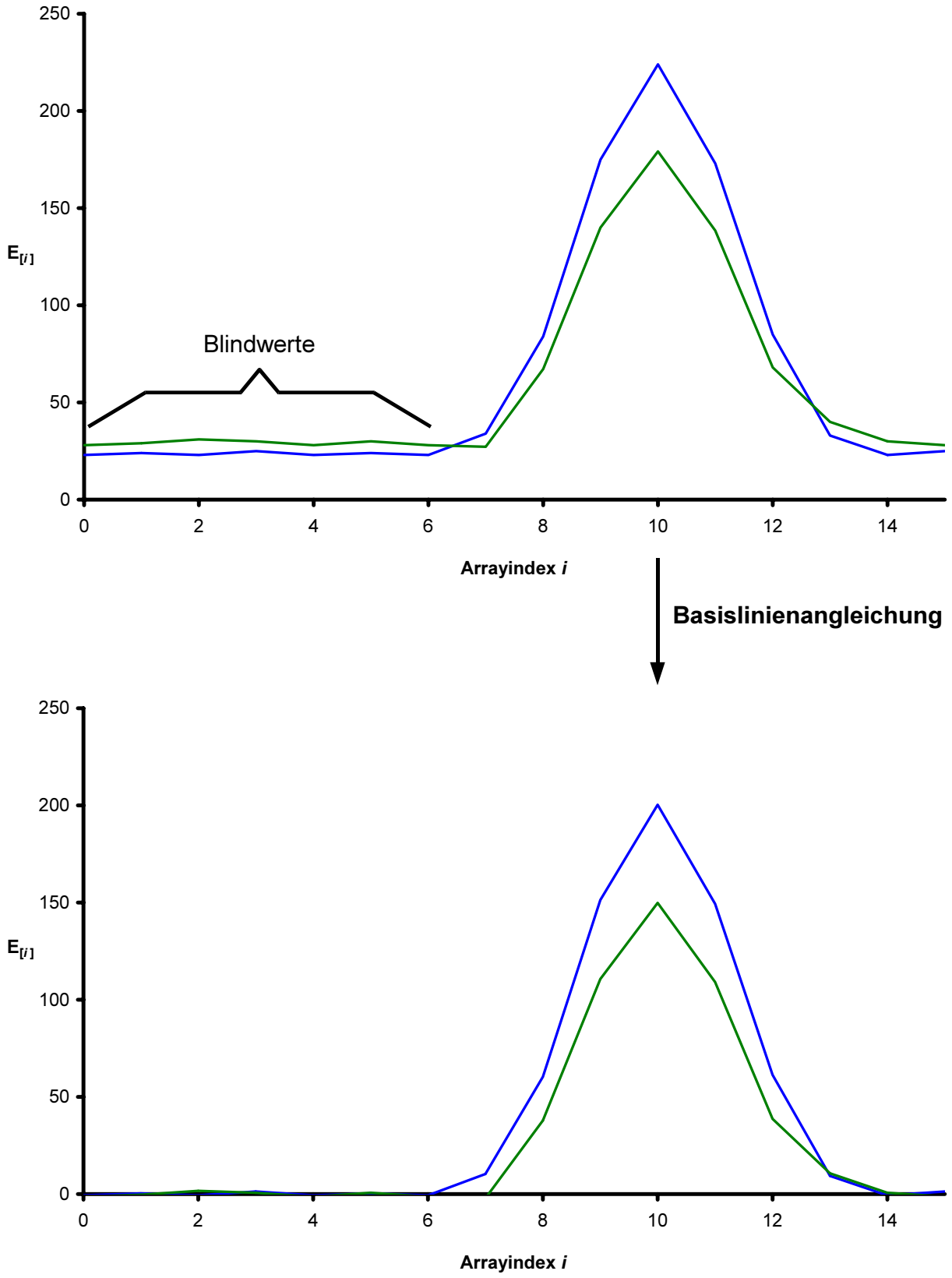

Abbildung 4.9-1: Basislinie bei parallel aufgezeichneten Chromatogrammen angleichen

4.9.2 Driftkorrektur

Immer wieder kommt es vor, dass die Basislinien von Chromatogrammen nicht konstant bleiben. Bei Mehrkanalmessungen mit Hilfe eines Polychromators können Probleme mit der Basislinie zumindest dann beseitigt werden, wenn sie alle Messkanäle gleichermaßen und gleichzeitig betreffen. Hierbei kommt es zu einer Parallelverschiebung der Basislinie aller Chromatogramme. Dies ist bei einigen elektronischen Problemen der Fall. Abbildung 4.9-2 zeigt, wie sich unter diesen Umständen die Basislinie bei einer Mehrkanalmessung innerhalb von 200 Sekunden auf Grund von elektronischen Störungen ändert. Die Kompensation der Basislinienschwankungen kann anhand eines zusätzlich aufgezeichneten Einzelchromatogramms erfolgen. Dieses Driftkorrektureinzelchromatogramm wird in einem Spektralbereich gemessen, wo keine Signale zu erwarten sind. Falls die Basislinienschwankungen nur träge erfolgen, lohnt sich eine Glättung des Driftkorrektureinzelchromatogramms. Die Glättung ist sinnvoll, damit das Rauschen bei der nachfolgenden Basislinienkorrektur nicht verstärkt wird. Anschließend wird aus allen Messwerten des Driftkorrektureinzelchromatogramms der Mittelwert gebildet. Jeder im Driftkorrektureinzelchromatogramm enthaltenen Einzelmesswert wird um diesen Mittelwert verringert (siehe Abbildung 4.9-3). Danach wird von jedem Einzelmesswert in jedem Einzelchromatogramm der entsprechende Messwert im Driftkorrektureinzelchromatogramm subtrahiert. Jeder Datenpunkt eines Einzelchromatogramms wird also folgender mathematischer Prozedur unterworfen:

$$S_{[i]}^{\mathrm{K}} = S_{[i]} - \left(S_{[i]}^{\mathrm{D}} - \overline{S}^{\mathrm{D}}\right)$$

$S_{[i]}^{\mathrm{K}}$ = die im i - ten Datenpunkt eines Einzelchromatogrammes gespeicherte Spektraldichte nach der Basislinienkorrektur,

$S_{[i]}$ = die im i - ten Datenpunkt eines Einzelchromatogrammes gespeicherte Spektraldichte vor der Basislinienkorrektur, (4.9-1)

$S_{[i]}^{\mathrm{D}}$ = die im i - ten Datenpunkt des Driftkorrektureinzelchromatogrammes gespeicherte Spektraldichte nach einer Bewegtsegmentglättung und

$\overline{S}^{\mathrm{D}}$ = Mittelwert über alle im Driftkorrektureinzelchromatogramm gespeicherten Spektraldichten.

Das Ergebnis einer solchen Driftkorrektur zeigt Abbildung 4.9-4.

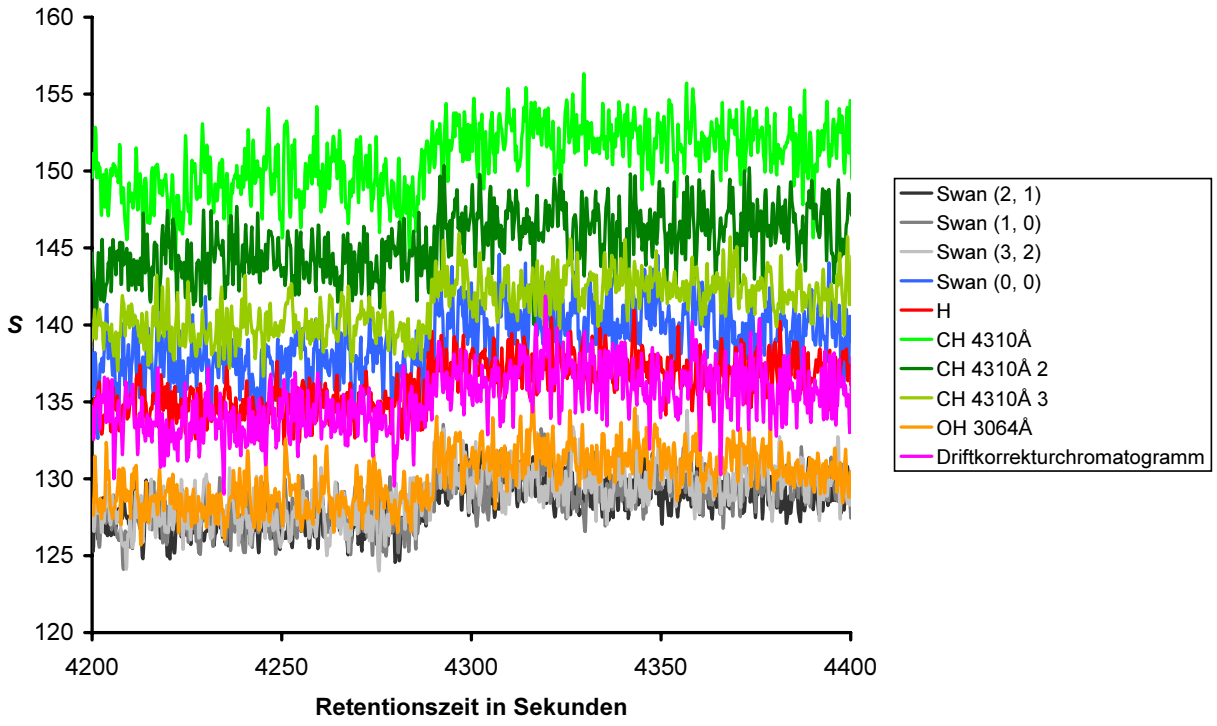

Abbildung 4.9-2: Basislinien für eine Reihe von parallel aufgezeichneten Einzelchromatogrammen

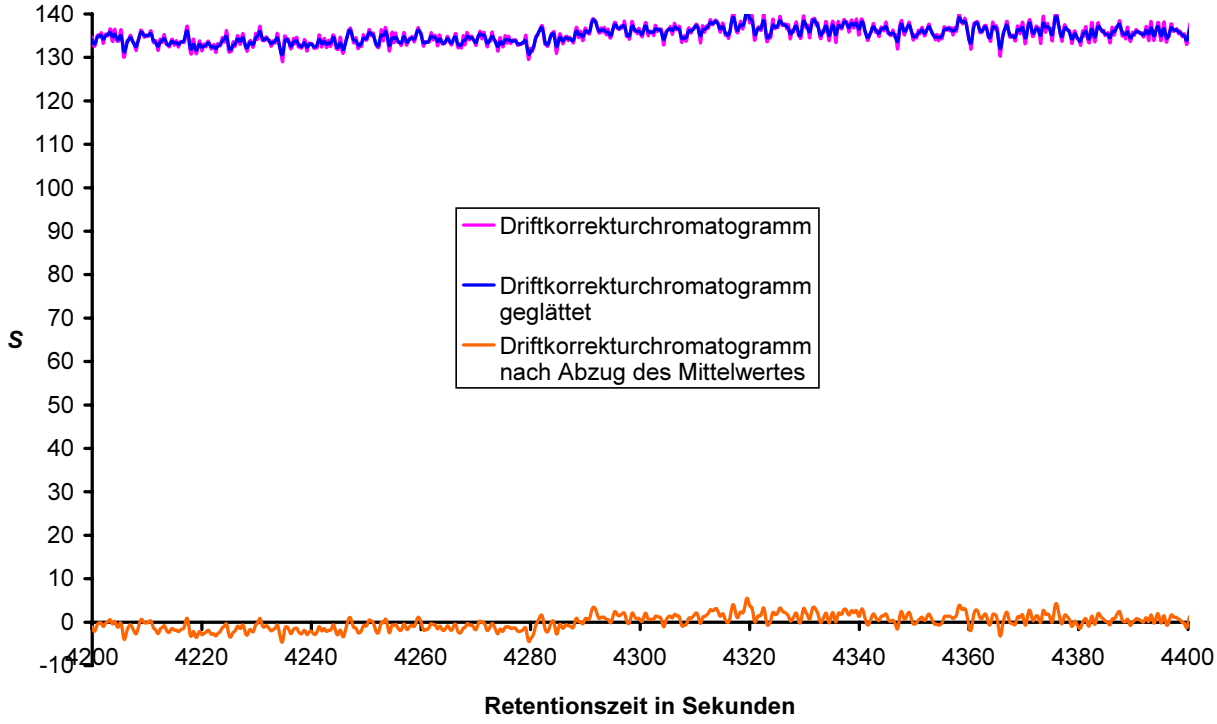

Abbildung 4.9-3: Das Einzelchromatogramm zum Driftausgleich vor und nach der Glättung mit einem 13 Elemente großen Filter auf Basis einer symmetrischen Binominalverteilung und nach Glättung und Abzug des Mittelwertes über alle seine Datenpunkte

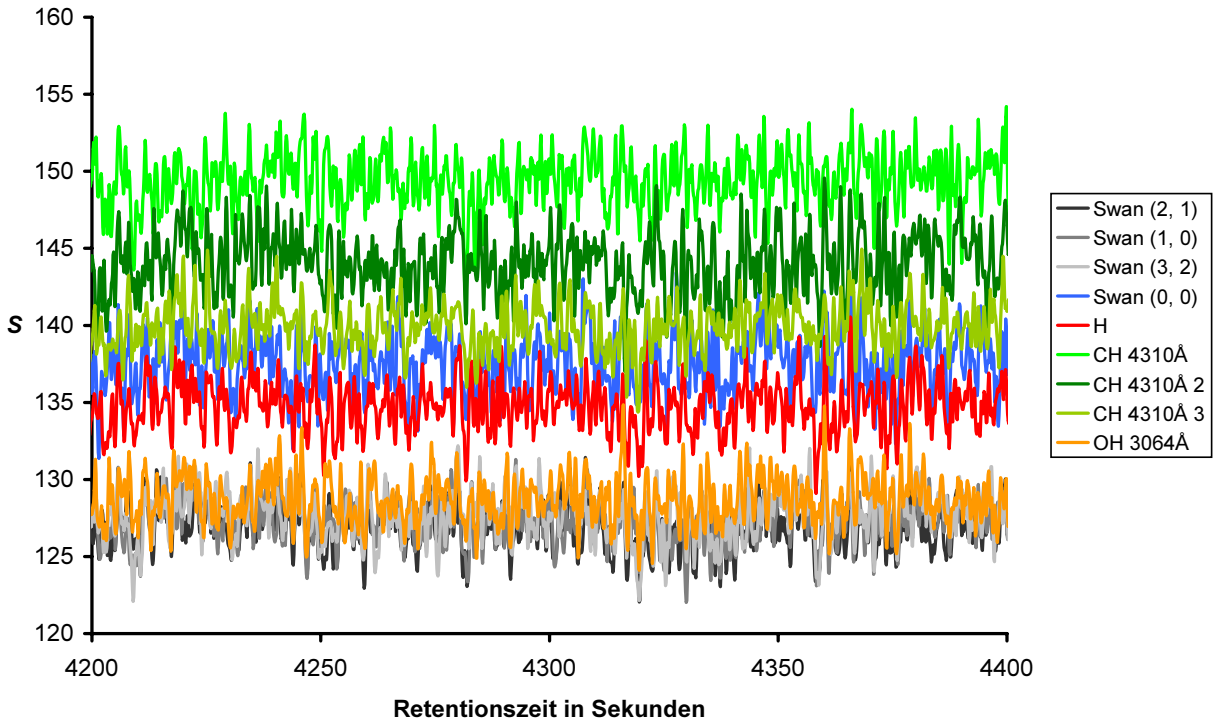

Abbildung 4.9-4: Basislinien für eine Reihe von parallel aufgezeichneten Einzelchromatogrammen nach einer Driftkorrektur

4.9.3 Das Vorgehen bei Memory-Effekten

Memory-Effekte treten in einem Detektionssystem immer dann auf, wenn Substanzreste den Detektor durch Anlagerungsreaktionen langsamer verlassen als sie in den Detektor gelangen. Verbleiben Substanzreste sehr lange in der Detektor-Apparatur, so ist eine Blindwertbestimmung nur nach sehr langen Wartezeiten möglich. In einem solchen Fall ist es sinnvoll die Blindwerte in einer vor den eigentlichen Messungen vorangestellten Kalibrationsmessung zu bestimmen.

Bei Mehrkanalmessung können unter der Voraussetzung, dass das Driftkorrekturchromatogramm (siehe vorherigen Abschnitt) nicht durch Memoryeffekte beeinflusst worden ist, systematische Fehler beim Angleichen der Basislinien vermieden werden (siehe Abschnitt 4.9.1). Zu diesem Zweck werden bei der Kalibrationsmessung die Differenzen zwischen dem Blindwert des Driftkorrekturchromatogramm und den Blindwerten aller von Memoryeffekten betroffenen Messkanäle gespeichert. Bei dem Angleichen der Basislinie werden in diesem Fall die Blindwerte nicht aus den aktuell gemessenen Einzelchromatogrammen bestimmt, sondern aus dem Blindwert des aktuell gemessenen Driftkorrekturchromatogramms und den bei der Kalibration gespeicherten Blindwertdifferenzen berechnet.

Memory-Effekte erschweren auch die Quantifizierung chromatographischer Signale. Die Bestimmung der Blindwerte muss über die Kalibrationsmessung erfolgen. Ein Memory-

Effekt kann aber dazu führen, dass der Schwellenwert über große Bereiche oder im gesamten Chromatogramm überschritten ist, ohne dass ein chromatographisches Signal vorliegt. Die Bestimmung der Signalgrenzen nach dem in Abschnitt 4.8 vorgestellten Algorithmus ist in diesem Fall nicht möglich. Bei Mehrkanalmessungen kann es hilfreich sein die Signalgrenzen anhand eines nicht von Memoryeffekten betroffenen Einzelchromatogrammes zu ermitteln und das Ergebnis auf die von Memoryeffekten beeinflussten Einzelchromatogramme anzuwenden. Die Blindwertbereinigung sollte durch das Anlegen einer Tangente erfolgen. Nur wenn die Gesamtmenge an Substanzrest bestimmt werden soll, darf mit dem über die Kalibrationsmessung ermittelten Blindwert gearbeitet werden.