Wissensbasierte Auswertung von Chromatogrammen

– Diplomarbeit –

Inv.-Nr.: 200 - 99 D - 094

Christoph Weidling

Mat-Nr. 24057

20. September 1999

Betreuer:	Dr.-Ing. Jürgen Nützel Dr. Thomas Böhme
Verantw. Hochschullehrer:	Univ.-Prof. Dr.-Ing. habil. Wolfgang Fengler
Inventarisierungsnummer:	200 - 99 D - 094

Eidesstattliche Erklärung

Ich versichere, daß ich die vorliegende Diplomarbeit mit dem Titel *Wissensbasierte Auswertung von Chromatogrammen* selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Ilmenau, den 20. September 1999 Christoph Weidling

Inhaltsverzeichnis

1 Einleitung **6**

2 Automatische Auswertung von Chromatogrammen **7**
2.1 Prinzipielle Struktur von Gaschromatogrammen 7
2.1.1 Allgemeine Prinzipien der Chromatographie 7
2.1.2 Besonderheiten bei Gaschromatogrammen 7
2.2 Prinzip des Chromarod-Verfahrens 8
2.3 Ein einfaches Modell der Trennsäule 9
2.4 Beschreibung eines Chromatogrammes 12
2.5 Beschreibung eines Peaks 13
2.6 Prinzipielle Auswertung 13
2.6.1 Detektion der Peaks 14
2.6.2 Integration 15
2.6.3 Fehlerquellen 16
2.7 Überblick über verwandte Arbeiten aus Literatur und Technik . . . 19

3 Auswertung mit Josephine **22**
3.1 Rauschen und Glättung 22
3.1.1 Bestimmen des Rauschwertes 22
3.1.2 Glättung der Daten 23
3.2 Peakdetektion 25
3.2.1 Lotfällung 25
3.2.2 Korrektur der Fußpunkte 27
3.2.3 Ausschluß von „fehlerhaften" Peaks 27
3.3 Erfahrungen aus der Arbeit mit Josephine 28

4 Lernfähige Verfahren **29**
4.1 Idee und Realisierungsmöglichkeiten 29
4.1.1 Der Mensch-Maschine-Dialog 29
4.1.2 Ziele 29
4.1.3 Peakevaluierung 31
4.1.4 Lage der Fußpunkte 32
4.2 Zeitbasierte Auswertung 33

4.3 Anwendung eines Fuzzy-ähnlichen Verfahrens 34
4.3.1 Fuzzy-Sets 34
4.3.2 Struktur des Entscheiders 35
4.4 Lernen von Parametern 36
4.4.1 Struktur des Lernverfahrens 36
4.4.2 Ein Lernschritt 38
4.5 Suche nach dem „besten“ Fußpunkt 40
4.5.1 Berechnung der Kriterien für einen Fußpunkt 40
4.5.2 Berechnung der Bewertung 43
4.5.3 Suche nach der höchsten Bewertung 44
4.5.4 Ein Lernschritt bei der Änderung eines Fußpunktes 45
4.6 Anwendung eines Clusterverfahrens 46
4.6.1 Das Verfahren 46
4.6.2 Erfahrungen mit dem Verfahren 47
4.6.3 Einschätzung und Zusammenfassung des Verfahrens 48

5 Prototypische Implementierung mit Amira 49
5.1 Verwaltung 49
5.1.1 Grobstruktur 49
5.1.2 Struktur eines Rohdatensatzes 50
5.1.3 Struktur einer Peakliste 51
5.1.4 Die Lernfunktion 52
5.2 Beschreibung der Algorithmen zur Generierung einer Peakliste . . . 53
5.2.1 Das Modul zum Finden der Peakkandidaten 55
5.2.2 Das Modul zur Trennung durch Lotfällung 56
5.2.3 Das Modul zur Schulterpeakabtrennung 58
5.2.4 Das Modul zum Sortieren der Peaks 58
5.2.5 Das Modul zum Evaluieren der Peaks 59
5.2.6 Das Modul zum Korrigieren der Fußpunkte 60
5.2.7 Das Modul zum Integrieren der Peaks 60
5.3 Das Lernen von Parametern 60

6 Zusammenfassung und Ausblick 62

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit dem automatischen Auswerten von Chromatogrammen. Insbesondere werden dabei ein lernfähiges Verfahren vorgestellt, welches das Auswerten von Serien von Chromatogrammen unterstützt. Mit einem solchen Verfahren kann dem Analytiker viele lästige, sich ständig wiederholende Arbeit abgenommen werden.

Diese Arbeit entstand im Rahmen eines Drittmittelprojektes der TU Ilmenau mit der ECH Elektrochemie Halle GmbH. Die ECH ist ein Unternehmen, das Gaschromatographen baut und diese Apperaturen mit Auswertesoftware ausstattet. Im Rahmen eines früheren Drittmittelprojektes wurden bereits Verfahren entwickelt, die in Produkten eingesetzt werden. Aufbauend auf den Ergebnissen dieser Arbeit wurde nach verbesserten Verfahren gesucht.

Der wesentliche Inhalt dieser Diplomarbeit ist die Darstellung der dem entwickelnden Verfahren zugrundliegenden Ideen. Im Kapitel 2 werden einige Grundpinzipien der Gaschromatographie erläutert. Im Kapitel 3 wird auf die Software *Josephine* eingegangen. Diese Software ist das Ergebnis des ersten Kooperationsvertrages der oben genannten Partner. Im Kapitel 4 werden die entwickelten Verfahren, deren Ziele und die Motivation für diese Verfahren beschrieben. Schließlich wird im Kapitel 5 auf die prototypische Implementierung der Software *Amira* eingegangen. Diese Software ist das Ergebnis des Kooperationsvertrages und wurde an die ECH Elektrochemie Halle GmbH übergeben.

An dieser Stelle möchte ich mich ganz besonders herzlich bei Dr. Thomas Böhme für die Zusammenarbeit bei diesem Projekt danken. Ein weiteres großes Dankeschön gilt Thomas Richter und Herrn Dr. P. Sivers von der ECH Elektrochemie Halle GmbH für die Unterstützung bei diesem Projekt.

2 Automatische Auswertung von Chromatogrammen

In diesem Kapitel werden grundlegende Prinzipien der Gaschromatographie und der Auswertung von Chromatogrammen beschrieben.

2.1 Prinzipielle Struktur von Gaschromatogrammen

Es wird eine kurze Einführung gegeben, wie ein Gaschromatograph arbeitet. Die Funktionsprinzipien eines solchen Gerätes und das Modell der Trennsäule findet man in ähnlicher Form auch bei anderen chromatographischen Verfahren.

2.1.1 Allgemeine Prinzipien der Chromatographie

Die verschiedenen chromatographischen Verfahren funktionieren nach dem Prinzip der räumlichen Stofftrennung. Durch physikalische oder chemische Kräfte innerhalb einer Trennsäule werden die Komponenten eines Stoffgemisches an der Grenze zwischen einer *mobile Phase* und einer *stationären Phase* getrennt. Das zu untersuchende Stoffgemisch bewegt sich dabei in der mobile Phase entlang der stationären Phase.

Wenn man die Stoffkonzentration der mobilen Phase am Ende der stationären Phase über der Zeit in ein Koordinatensystem aufträgt, erhält man ein Chromatogramm. Bild 2.1 zeigt einen typischen Plot. Zu dem Zeitpunkt, wenn die Stoffkonzentration der mobilen Phase am Ende der stationären Phase sehr hoch ist, entsteht ein Peak. Dabei tritt bei konstanten Umgebungsbedingungen in jedem Plot für jede Komponente eines Gemisches der Peak an der gleichen Stelle auf.

Durch Kalibrierungen kann man aus einem Plot sowohl qualitativ und als auch quantitativ auf die Zusammensetzung des Gemisches schließen.

2.1.2 Besonderheiten bei Gaschromatogrammen

Bei Gaschromatogrammen ist die mobile Phase gasförmig und die stationäre flüssig oder fest. Die Stoffprobe wird zusammen mit einem reaktionsträgen Trägergas (Stickstoff oder Wasserstoff) durch die Trennsäule geleitet.

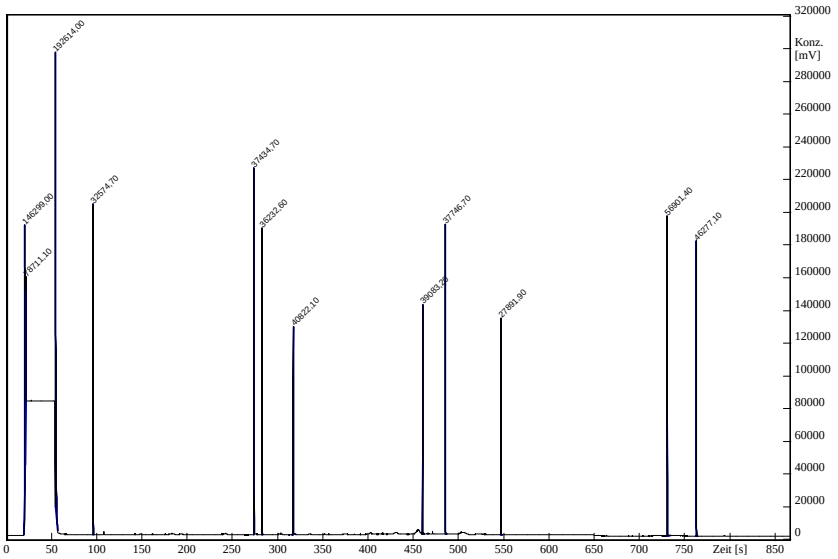


Abbildung 2.1: Gaschromatogramm von Diesel

Beim Verlassen der Trennsäule wird das getrennte Stoffgemisch in einen Detektor, zum Beispiel *Flammenionisationsdetektor* (FID) geleitet. Dabei handelt es sich um eine Kammer, in der sich zwei Elektroden befinden. In dieser Kammer wird das Gas in einer Flamme ionisiert. Wenn nun an die zwei Elektroden eine Spannung angelegt wird, kann man einen Stromfluß messen, der von der Konzentration des Stoffes in der mobilen Phase abhängt.

Die Fläche eines Peaks liefert eine Aussage über die Menge der entsprechenden Komponente des Stoffgemisches. Die Fläche des Peaks verhält sich proportional zur in der Probe enthaltenen Stoffmenge der entsprechenden Komponente.

Wenn man ein Chromatogramm ohne eine Stoffprobe erzeugt, also nur das Trägergas durch die Trennsäule leitet, erhält man ein *Leerchromatogramm*. Der Verlauf der Kurve in einem Leerchromatogramm wird *Basislinie* genannt.

2.2 Prinzip des Chromarod-Verfahrens

Das Chromarod-Verfahren ist ein spezielles chromatographisches Verfahren. Es wurde von Okumura und Kadano 1972 in Japan entwickelt und patentiert. Die ECH hat dieses Verfahren durch eine neuartige Detektionsmöglichkeit weiterentwickelt ([HS98]) und setzt es vorwiegend zur Identifizierung von schwerflüchtigen Kohlenwasserstoffen und Kohlenwasserstoffgemischen ein.

Chromarods sind dünne Quarzstäbe, die mit einer Trennphase beschichtet sind. Das Trennmaterial, zum Beispiel Kieselgel oder Aluminiumoxid, wird mit Hilfe

eines anorganischen Binders aufgetragen. Die Stofftrennung bei diesem Verfahren beruht auf den Wechselwirkungen mit dem Trennmaterial in einem auf der Oberfläche der Trennschicht diffundierenden Lösungsmittel.

Auf ein Ende eines solchen Stäbchens wird eine organische Stoffprobe aufgetragen. Infolge unterschiedlicher Wanderungsgeschwindigkeiten trennen sich die Komponenten des Gemisches voneinander. Nach Abdampfen des Lösungsmittels wird das Stäbchen langsam durch einen Brennofen geschoben. Dabei werden die organischen Verbindungen im Sauerstoffstrom verbrannt.

Infolge des unterschiedlichen Kohlenstoffanteils der Komponenten, die auf der Länge des Stäbchens verteilt sind, ändert sich die Konzentration des CO_2-Anteils im Gasstrom. Dieser Anteil kann mit einem Infrarot-Sensor gemessen werden und erzeugt somit – über einer Zeitachse oder über der Länge des Stäbchens aufgetragen – ein Chromatogramm.

Nach dieser Prozedur wird das Stäbchen gesäubert und kann erneut verwendet werden.

2.3 Ein einfaches Modell der Trennsäule

Wenn man wissen will, wie ein Chromatogramm entsteht, muß man untersuchen, was in der Trennsäule vonstatten geht. Ein einfaches Modell geht davon aus, daß sich die mobile Phase zeitgetaktet bewegt und die Trennsäule in kleine Zellen zerlegt wird. Innerhalb eines Zeittaktes stellt sich schlagartig ein Gleichgewicht ein und die mobile Phase springt ein Element weiter.

Durch diese Vereinfachung ist es möglich, die Zeit mit Hilfe einer Anzahl von Takten darzustellen. Das Modell vereinfacht insofern sehr stark, als daß die Geschwindigkeit, mit der sich die mobile Phase durch die Säule bewegt, nicht berücksichtigt wird. Dabei spielt diese Geschwindigkeit für den Analytiker eine entscheidende Rolle: Sie ist im wesentlichen verantwortlich dafür, wie gut sich die Komponenten voneinader trennen.

Jede Stoffkomponente eines Gemisches verteilt sich aufgrund eines der Komponente entsprechenden Verhältnisses α mit $0 \leq \alpha \leq 1$ auf die stationäre und mobile Phase. Wenn sich also in einer Zelle Z_k der Trennsäule zu einem Zeitpunkt t die Stoffmenge s einer Komponente befindet, verbleiben nach dem Konzentrationsausgleich $s \cdot p$ Anteile in der stationären und $s \cdot q$ Anteile in der mobilen Phase, wenn $p = \alpha$ und $q = 1 - \alpha$.

Dieses Verfahren ist in Bild 2.2 skizziert. Dabei bedeuten S die stationäre und M die mobile Phase. Zu Beginn sind sowohl die mobile als auch die stationäre Phase leer. Zu einem Zeitpunkt t_0 wird die Stoffmenge 1 einer Komponente in die mobile Phase injiziert. Ein dicker horizontaler Strich kennzeichnet das Ende eines Zeittaktes. Die Zeitachse ist rechts dargestellt und verläuft von oben nach unten.

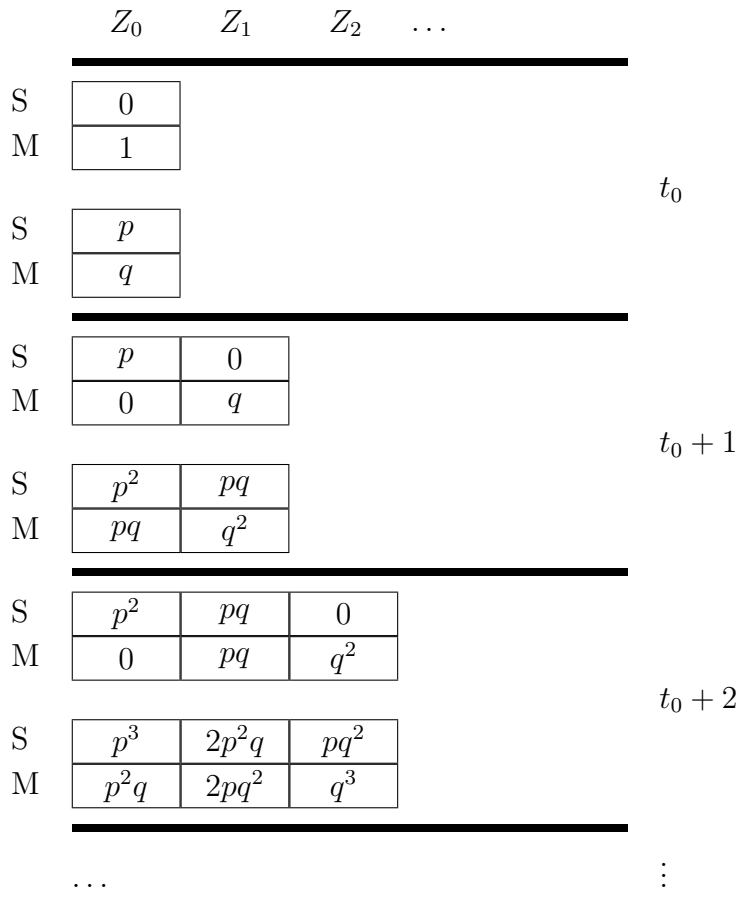


Abbildung 2.2: Ausgleichsvorgänge in der Trennsäule

Nach dem Zeitpunkt $t_0 + n$ erhält man für die Konzentration $c_n(k)$ in der Zelle Z_k:

$$c_n(k) = q\binom{n}{k}p^{n-k}q^k \tag{2.1}$$

Sei nun

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{x^2}{2}}. \tag{2.2}$$

Für große n ist gemäß [BO76] für alle k

$$\lim_{n\to\infty} \frac{\frac{1}{q}c_n(k)}{\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)} = 1, \tag{2.3}$$

wobei $\sigma = \sqrt{npq}$ und $\mu = np$. Damit folgt für große n

$$c_n(k) = \frac{q}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{q}{\sqrt{2\pi}\sigma}e^{-\frac{(k-\mu)^2}{2\sigma^2}} \tag{2.4}$$

Wenn man davon ausgeht, daß ein Zeittakt T Zeiteinheiten dauert und daß die Trennsäule n Einheiten lang ist, kann man k und n in Abhängigkeit von der Zeit t, der Säulenlänge L und der Säulengeschwindigkeit v darstellen.

Bis zum Verlassen der Trennsäule benötigt der Anfang der mobilen Phase[1] $\frac{L}{v}$ Zeit. Da die Säule n Zellen lang ist und die mobile Phase für einen Sprung zu nächsten Zelle T Zeit benötigt, gilt

$$n = \frac{L}{Tv}. \tag{2.5}$$

Wenn man die Verteilung des Stoffes in der Trennsäule betrachtet, erhält man, wie eben gezeigt, eine Glockenkurve über einem Parameter k, der die Nummer der Zelle bezeichnet. Betrachtet man diese Kurve als eine Funktion der Zeit t, erhält man demzufolge auch eine Glockenkurve:

$$H(t) = \frac{q}{b}\varphi\left(\frac{t - t_m}{b}\right) \tag{2.6}$$

Für das Überspringen von k Zellen wird $t = kT$ Zeit benötigt, damit ist $t_m = \mu T = nTp = \frac{L}{v}p$. Wählt man $b = \sqrt{T}\sigma = \sqrt{\frac{L}{v}pq}$, erhält man

$$\begin{aligned} H(t) &= \frac{q}{\sqrt{T}\sigma}\varphi\left(\frac{t - t_m}{\sqrt{T}\sigma}\right) \\ &= \frac{q}{\sqrt{2\pi\frac{L}{v}pq}} e^{\frac{\left(t-\frac{L}{v}p\right)^2}{2\frac{L}{v}pq}} \end{aligned} \tag{2.7}$$

In dieser Formel taucht weder die Anzahl der Zellen noch die Zeit zwischen zwei Sprüngen der mobilen Phase auf. Bedenkt man noch, daß das rechte Ende der mobilen Phase zuerst aufgezeichnet wird, erhält man folgenden Verlauf $p(t)$ des Chromatogrammes[2]:

$$\begin{aligned} p(t) &= H\left(\frac{L}{v} - t\right) \\ &= \frac{q}{b\sqrt{2\pi}} e^{\frac{(t-t_R)^2}{2b^2}}, \end{aligned} \tag{2.8}$$

wobei $t_R = \frac{L}{v} - t_m$ die Retentionszeit des Peaks und q dessen Fläche ist.

[1] Das ist jeweils das rechte Ende der mit „M" markierten Zellen im Bild 2.2

[2] Es wird hier angenommen, daß der Trennvorgang zum Zeitpunkt $t = \frac{L}{v}$ abgeschlossen ist.

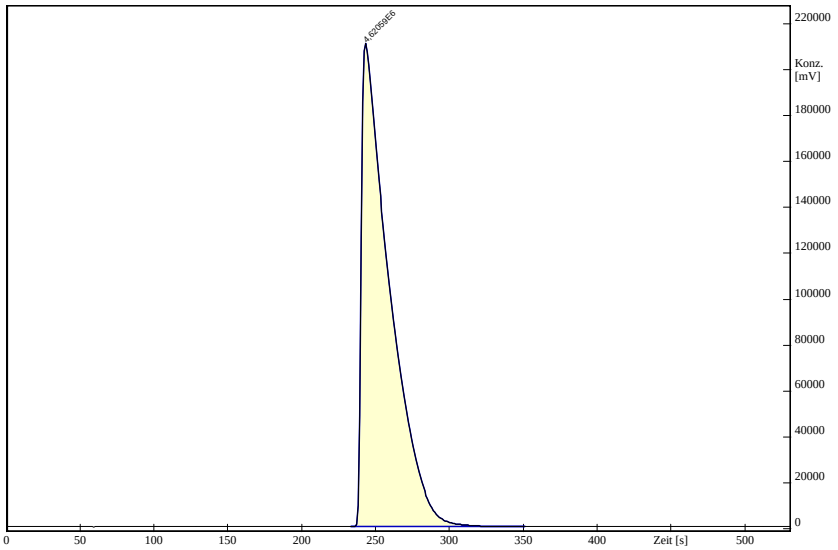


Abbildung 2.3: Ein mit einem Gaschromatographen aufgezeichneter Peak

Betrachtet man einen mit einem Gaschromatographen aufgezeichneten Peak (Bild 2.3), stellt man fest, daß die abfallende Flanke schwächer fällt als die ansteigende Flanke steigt; der Peak besitzt ein *Tailing*. Seltener treten Peaks mit einem *Fronting* auf.

Das Tailing von Peaks kann man mit einer *Exponential modifizierte Gaußfunktion* (EMG) beschreiben (siehe [LEST84]). Die EMG wird durch folgende Gleichung beschrieben:

$$p(t) = \frac{A}{\tau}\, e^{\frac{1}{2}\left(\frac{\sigma_G}{\tau}\right)^2 - \left(\frac{t-t_G}{\tau}\right)} \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \tag{2.9}$$

mit

$$z = \frac{t - t_G}{\sigma_G} - \frac{\sigma_G}{\tau}.$$

2.4 Beschreibung eines Chromatogrammes

Früher wurden die Daten, die während der Arbeit eines Chromatographen entstehen, mittels eines Plotters kontinuierlich auf ein Blatt Papier gezeichnet. Dieser Plot wurde manuell ausgewertet. Heute werden die Daten in den meisten Fällen digital aufgezeichnet. Die Sensordaten müssen dazu digitalisiert werden.

Die Daten liegen dann in Form zweier Folgen bzw. Vektoren reeller Zahlen vor. Die eine Folge $(t_i)_{0 \le i < N}$ enthält die Zeiten, an denen der Datenpunkt aufgenommen wurde, und die andere $(v_i)_{0 \le i < N}$ enthält die zu diesem Zeitpunkt vom Sensor gelieferten Meßwerte.

2.5 Beschreibung eines Peaks

Ein Peak ist in einem solchen Chromatogramm im wesentlichen durch den *linken Fußpunkt*, die *Retentionszeit* und durch den *rechten Fußpunkt* definiert (siehe Bild 2.4). Die Retentionszeit gibt den Zeitpunkt an, an dem die Kurve zwischen den Fußpunkte das Maximum annimmt. Durch die Retentionszeit erkennt der Analytiker, welchem Stoff der Peak entspricht. Der Retentionspunkt ist der Punkt der Kuve am Retentionszeitpunkt.

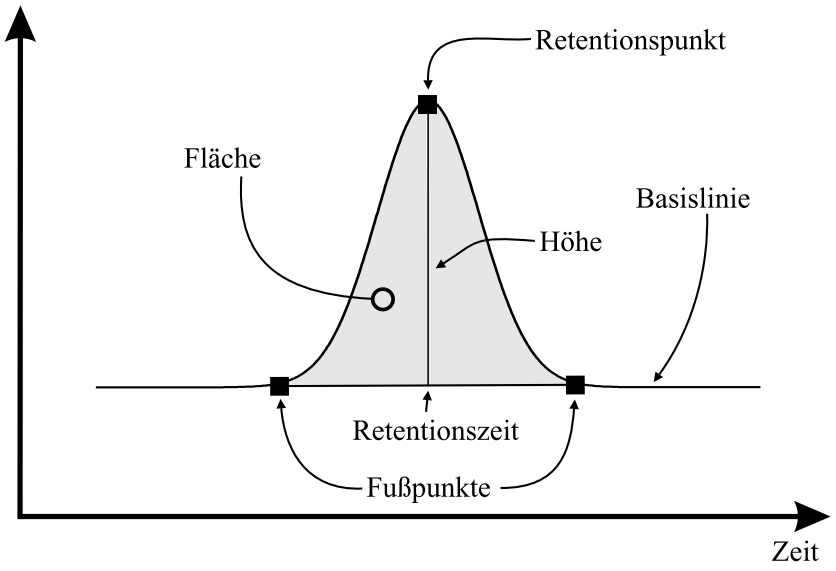


Abbildung 2.4: Wichtige Größen eines Peaks

Um die Stoffmenge der entsprechenden Komponente zu bestimmen, benötigt der Analytiker die Fläche des Peaks. Diese Fläche wird von der Kurve zwischen den Fußpunkten und der Verbindungsstrecke der Fußpunkte begrenzt.

2.6 Prinzipielle Auswertung

Die automatische Auswertung von Chromatogrammen muß zwei Dinge leisten:

- Es müssen die Peaks gefunden werden. Dazu gehört, daß die Fußunkte möglichst präzise gefunden werden.
- Die Peaks müssen integriert werden.

Im folgenden wird die grundsätzliche Herangehensweise an diese beiden Probleme erläutert.

2.6.1 Detektion der Peaks

In einem störungsfrei aufgezeichneten Chromatogramm gehört ein Datenpunkt entweder zu einem Peak oder zur Basislinie. Dies führt zu zwei grundsätzlichen Herangehensweisen der Peakdetektion:

Man sucht die Basislinie. Alles, was nicht zur Basislinie gehört, sind die Peaks. Hat man die Basislinie einmal gefunden, kann man die Kurve der Basislinie vom Chromatogramm subtrahieren. Zur Basisliniendetektion wurden verschiedenen Verfahren in [HER95] untersucht. Im Rahmen von [WEI98] wurden mehrere Verfahren daraus eingeschätzt. Stellvertretend werden hier zwei dieser Verfahren kurz skizziert.

- Ein Verfahren betrachtet den Plot als Fläche. Diese Fläche wird in Pixel zerlegt. Anschließend wird mit einem neuronalen Netz die Basislinie gesucht.
- Ein weiteres Verfahren schätzt eine Basislinie und bewertet jeden Datenpunkt anhand der Entfernung zur geschätzten Linie, ob er zur Basislinie gehört oder nicht. Punkte mit einer schlechten Bewertung werden gestrichen. Dieses Verfahren wiederholt sich so lange, bis eine vorher festgelegte maximale Anzahl von Punkten nicht zur Basislinie gehört.

Diese beiden Verfahren sind sehr aufwendig. Während das erste sehr viel Speicherplatz benötigt, braucht das zweite Verfahren sehr viel Rechenzeit.

Man sucht die Peaks direkt. Hier kann man zwei Herangehensweisen unterscheiden:

- Man sucht die Peaks rekursiv. Dabei nimmt man an, daß der Maximalwert der Folge (v_i) auch der Wert des Retentionspunktes eines Peaks ist. Ausgehend von diesem Maximum sucht man nach links und nach rechts die Fußpunkte dieses Peaks. Anschließend betrachtet man die beiden Teile links vom linken Fußpunkt und rechts vom rechten Fußpunkt als eigenständiges Chromatogramm und analysiert dieses auf die gleiche Weise. Die Rekursion bricht ab, wenn ein die Anzahl der Datenpunkte eine vorgegebene Grenze unterschreitet. Da die dabei entstehenden Teilchromatogramme immer kleiner sind als das große, bricht die Rekursion immer ab.
- Man durchsucht das Chromatogramm (auf der Zeitachse betrachtet) von links nach rechts. Wenn eine vorgegebene Anzahl von unmittelbar aufeinanderfolgenden Datenpunkten ansteigt, das heißt $v_k < v_{k+1} <$

$\cdots < v_{k+m}$, geht man davon aus, daß es sich um eine ansteigende Peakflanke handelt. Die nächste abfallende Flanke ist dann die rechte Peakflanke. Offensichtlich bricht dieses Verfahren am Ende des Chromatogrammes ab.

Das zweite Verfahren arbeitet schneller als das erste. Während man beim zweiten Verfahren nur $O(n)$ Vergleiche benötigt, braucht man für das erste Verfahren $O(n \log n)$ Vergleiche.

2.6.2 Integration

Wie oben bereits erwähnt, gibt die Fläche unter einem Peak Aufschluß darüber, welchen Anteil die entsprechende Komponente am Stoffgemisch hat. Für den Analytiker ist es demzufolge wichtig, eine genaue Angabe über diese Peakfläche zu bekommen.

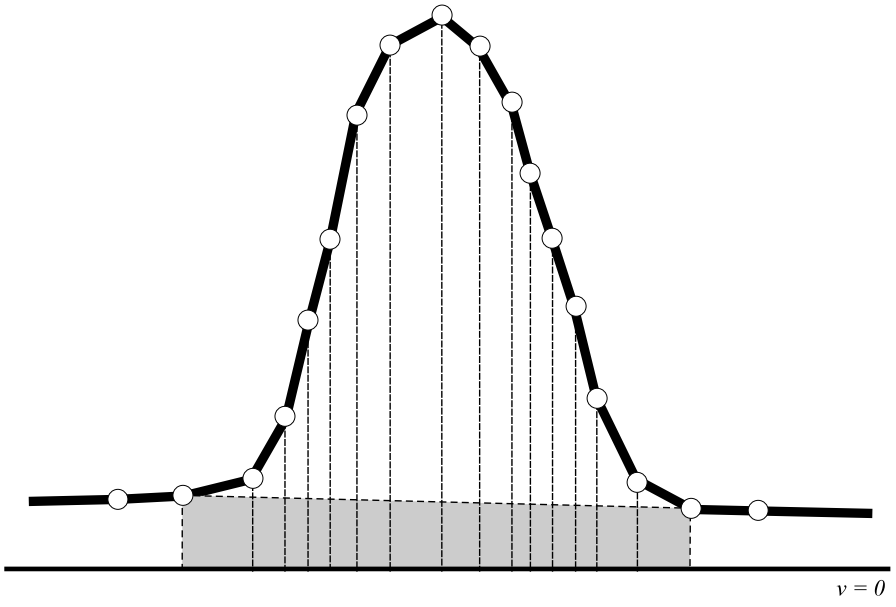


Abbildung 2.5: Flächenbestimmung eines Peaks

Für die Integrationsmethode kann man den Peak einfach als die Vereinigung von Trapezen auffassen. Die Trapeze sind definiert durch jeweils zwei aufeinanderfolgende Datenpunkte $(t_j, v_j)^T$ und $(t_{j+1}, v_{j+1})^T$ sowie durch die Schnittpunkte der Verbindungslinie g der Peakfußpunkte mit den beiden Geraden $t = t_j$ und $t = t_{j+1}$.

Um nicht für jeden Datenpunkt $(t_j, v_j)^T$ den Schnitt mit g berechnen zu müssen, kann man auch zunächst die Trapeze nicht durch g begrenzen, sondern durch die

Gerade $v = 0$. Nachdem man die Summe der Trapezflächen gebildet hat, subtrahiert man das Trapez, das durch die Gerade $v = 0$ und die beiden Peakfußpunkte definiert ist.

In Bild 2.5 ist diese Zerlegung anhand eines Beispiels dargestellt. Die Trapeze sind durch gestrichelte Linien dargestellt. Das zu subtrahierende Trapez ist grau unterlegt.

Schließlich erhält man zur Berechnung der Peakfläche F des Peaks mit den Fußpunkten $(t_l, v_l)^T$ und $(t_r, v_r)^T$ folgende Formel:

$$F = \frac{1}{2}\left(-(v_l + v_r)(t_r - t_l) + \sum_{j=l}^{r-1}(v_{j+1} + v_j)(t_{j+1} - t_j)\right) \tag{2.10}$$

2.6.3 Fehlerquellen

Die eben beschriebenen Verfahren arbeiten sehr gut mit störungsfrei aufgenommen Chromatogrammen. Allerdings arbeiten Chromatographen nicht fehlerfrei. Auch der Analytiker hat einen Einfluß auf die Qualität von Gaschromatogrammen. Folgende Fehlerquellen beieinträchtigen die Güte eines Chromatogrammes:

Rauschen Rauschen wird durch mehrere Faktoren verursacht. Einige dieser Störfaktoren sind Strömungen innerhalb der Trennsäule, die den Trennprozeß unregelmäßig beeinflussen. Außerdem schwankt die Temperatur der Flamme eines FID, so daß die mobile Phase ungleichmäßig ionisiert wird. Schließlich kommen noch Rauschen und Systemfehler bei A/D-Umsetzung hinzu.

Schlecht getrennte Peaks Wenn die mobile Phase zu schnell durch die Trennsäule geführt wird, kann es passieren, daß Komponenten nicht vollständig getrennt werden und sich im Chromatogramm überlappen.

Große Auswirkungen auf die Trenngüte hat die Wahl der Säule. Mit verschieden Säulen können Stoffe unterschiedlich gut in ihre Komponenten zerlegt werden. Obwohl es zum Beispiel möglich ist, die drei Gase Sauerstoff, Stickstoff und Kohlendioxid voneinander zu trennen, gelingt dies nicht mit einer einzigen Säule ([ATB93]).

Schlecht getrennte Peaks treten vor allem bei *Lösungsmittelpeaks* auf. Proben, vor allem feste, werden oft in einem Lösungsmittel aufgelöst. Die entstandene Lösung wird dann als Probe inijziert. Da der Anteil des Lösungsmittels im Vergleich zu den anderen Komponenten sehr groß ist, erzeugt das Lösungsmittel im Chromatogramm einen sehr großen Peak, dessen Fußpunkte sehr weit auseinanderliegen. Wenn sich in der Stoffprobe Komponenten befinden, die eine ähnlich Retentionszeit wie das Lösungsmittel haben, werden diese nicht von dem breiten Lösungsmittelpeak getrennt, son-

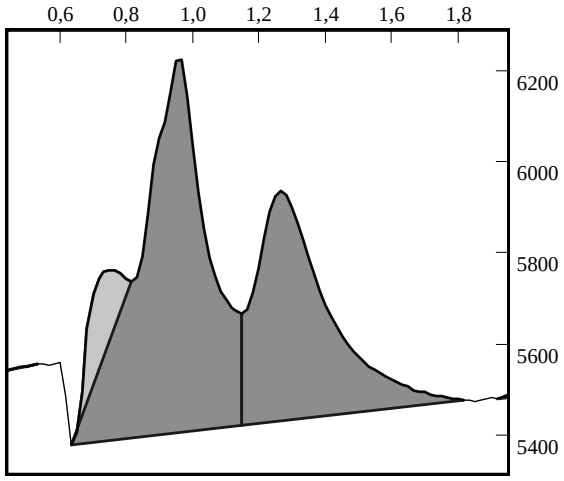


Abbildung 2.6: Durch Lotfällung voneinander getrennte Peaks (dunkel) und ein Schulterpeak (hell)

dern sitzen auf seinen Flanken auf. Man nennt solche aufsitzenden Peaks *Schulterpeaks*. Der Analytiker integriert solche Peaks, indem er sie mit dem *Tangetenverfahren* trennt. Dabei verbindet er die Fußpunkte des Schulterpeaks (siehe Bild 2.6).

Weichen die Peaks nicht derart stark voneinander ab und sind sie nicht richtig getrennt, zeichnet der Analytiker das Lot vom *Talpunkt* (der Punkt, der zwischen den beiden Retentionspunkten das Minimum annimmt) der beiden Peaks zur Absizsse des Koordinatensystems ein. Wo dieses Lot die Verbindung zwischen den äußeren Fußpunkten schneidet, befinden sich der neue linke und rechte Fußpunkt des rechten und linken Peaks (siehe Bild 2.6). Das Verfahren heißt *Lotfällung*.

Der Grund für dieses Vorgehen zeigt sich, wenn man beispielsweise zwei eng beeinanderliegende Gaußkurven addiert (Bild 2.7). Wenn man die beiden Peaks durch Lotfällung voneinander trennt, ist für den Analytiker der Unterschied der Flächen bei der Integration dieser Peaks gegenüber den wahren Flächen der einzelnen Peaks hinnehmbar klein. Allerdings nennt [DYS90] zwei Voraussetzungen für ein solches Vorgehen:

- Die Peaks müssen annäherend die gleiche Höhe besitzen.
- Der Talpunkt liegt nicht oberhalb von 5% der Peakhöhe

Rampen Als Rampen bezeichnet man einen plötzlichen Anstieg der Basislinie. Wenn die Temperatur der Trennsäule plotzlich geändert wird, treten im Chromatogramm Flanken auf, die nicht zu einem Peak gehören.

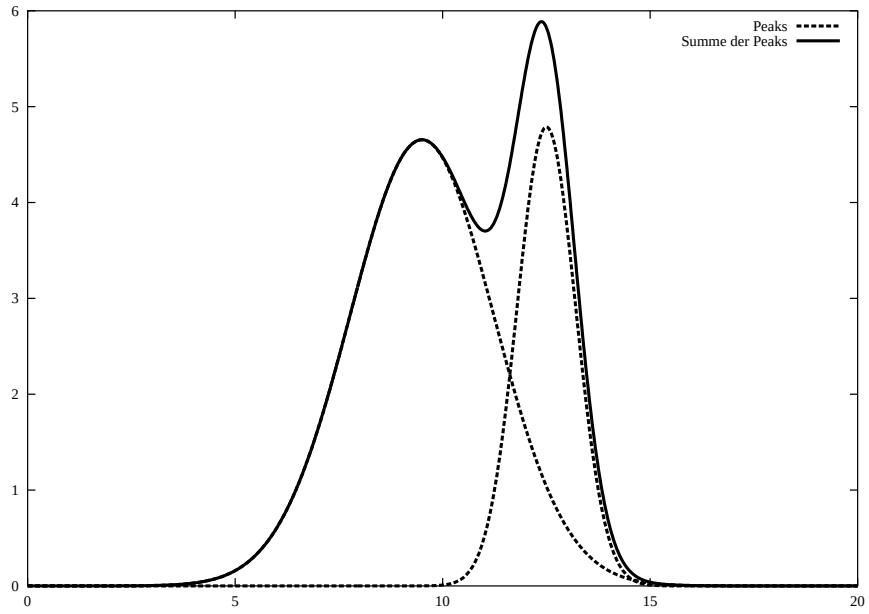


Abbildung 2.7: Zwei sich überlappende Gaußkurven werden addiert

Spikes Als Spikes bezeichnet man Störungen, die wie ein zu schmaler Peak aussehen. Diese können zum Beispiel durch elektrische Störungen auftreten.

Negative Peaks Negative Peaks können auftreten, wenn die Probe in die mobile Phase gegeben wird. Solche Störungen sehen wie umgekippte Peaks aus (siehe Bild 2.8).

Plateaus Wenn der Detektor gesättigt ist oder der A/D-Umsetzer durch ein zu hohes Eingangssignal übersteuert wird, treten Plateaus am oberen Rand des Chromatogrammes auf. Das hat Folgen für die Integration: Der obere Teil des Peaks wird abgeschnitten, und die entsprechende Fläche kann nicht berechnet werden.

Baisliniendrift Bei vielen Chromatogrammen steigt die Basislinie am Ende des Chromatogrammes stetig an. Dafür gibt es vielerlei Gründe. Einerseits kann eine Drift durch Schwankungen im Gasdruck innerhalb der Säule hervorgerufen werden, aber auch ein Temperaturanstieg der Säule oder des Detektors läßt die Basislinie driften. Ein Driften der Basislinie hat insofern Auswirkungen auf die Detektion von Peaks, als daß auf einer driftenden Basislinie die Retentionszeiten leicht verschoben werden. Außerdem wird die Fläche unter einem Peak verfälscht (siehe [DYS90], [WEI98S]). Verfahren zur Korrektur der Basislinie sind patentiert ([POKO85], [ALLA89]).

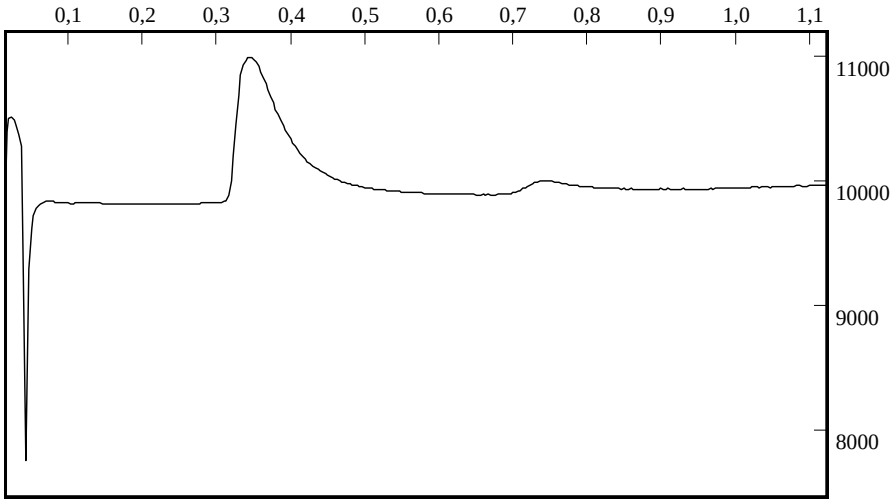


Abbildung 2.8: Ein negativer Peak

Alle diese Störungen müssen gefunden und entsprechend behandelt werden. Die Schwierigkeit dabei ist, daß sich einige Störungen nicht eindeutig von einem Peak unterscheiden.

2.7 Überblick über verwandte Arbeiten aus Literatur und Technik

Wenn man sich nach Literatur zur automatischen Chromatographie umsieht, findet man eine Vielzahl von Produktbeschreibungen ([HP87], [HP93], [ZI89]) und Pressemitteilung von den Herstellern der Chromatographen. Aufgrund der kommerziellen Bedeutung der Gaschromatographie sind genaue Aussagen über die Wirkunsweise marktgängiger Auswerteverfahren kaum zu bekommen. Solche Verfahren werden im allgemeinen von den Herstellern zusammen mit der gaschromatographischen Ausrüstungen im Paket verkauft.

Interessant ist die in [ZI89] beschriebene Sprache, mit der man Rohdaten und Peaks verwalten kann. Die Identifizierung von Peaks kann über verschiedene Retentionsgrößen durch Vergleich mit Tabellen (oder früher ausgewerteten Chromatogrammen) oder interaktiv erfolgen. Alle gängigen Methoden der quantitativen Bestimmung werden unterstützt, einschließlich Regressionsverfahren sowie simulierte Destillation von Rohölen, ebenso die Bestimmung von Kenngrößen zur Beurteilung von Trennsäulen. Die wichtigsten Bearbeitungsschritte können in Verbindung mit einer graphischen Schnittstelle unterstützt durch Fadenkreuz oder Maus ausgeführt werden. Zur Bearbeitung von Chromatogrammserien können beliebige

Folgen von Befehlen zu abrufbaren Prozeduren zusammengestellt werden. Meß- und Auswerteparameter, wie Zeitpunkt der letzten Bearbeitung und der Name des Bearbeiters werden automatisch festgehalten. Für die numerische Bearbeitung von spektroskopischen Rohdaten eignen sich Befehle für Glättungsverfahren, für die Bildung von Abeitungen, für Fouriertransformation und Autokorrelation, für die Unterdrückung von niederfrequentem Rauschen durch Fourieranalyse, für die Auflösungsverbesserung durch Linienfaltung sowie Simulaion theoretischer Linienprofile.

In [ALLA89] wird ein mathematisches Verfahren zur Beseitigung systematischer Fehler bei der Auswertung von Chromatogrammen beschrieben. Die betrachteten Chromatogramme sind Flüssigkeitschromatogramme. Als Sensoren dienen dabei Spektrometer (UV- oder sichtbares Licht). Die beschriebene Methode nutzt hauptsächlich den Umstand, daß der von einem Spektrometer gelieferte Meßwert vektoriell und nicht skalar ist. Sie kann daher nicht direkt auf Gaschromatographen mit Flammenionisationsdetektor oder Wärmeleitfähigkeitssensoren übertragen werden.

Einen Überblick über die Funktionsweise der Auswerteverfahren und über Verfahren zur Kompensation einer Basisliniendrift findet man in Patenten ([TOM92], [HIT92], [POKO85]).

Die bei einem Temperatur- der Strömungsgeschwindigkeitsprogramm bei einer chromatographischen Trennsäule auftretende Basisliniendrift wird in [POKO85] durch Signale von Funktionsgeneratormitteln kompensiert. Die Funktionsgeneratormittel liefern eine Darstellung der Basisliniendrift als analytische Funktion der Temperatur mit trennsäulenspezifischen Parametern. Die Parameter werden in einem Testlauf bestimmt und in die Funktionsgeneratormittel eingegeben. Das Ergebnis ist eine eine Gleichung, die das Verhalten der Basislinie währen der Stabilisierung auf einem neuen Temperaturniveau θ_1 beschreibt:

$$I = ae^{b\left(\theta_1 - \Delta\theta e^{-\frac{t}{\tau}}\right)} + I_0, \tag{2.11}$$

wobei die Säulenkonstanten θ, τ, a, b und I_0 durch einen in dieser Quelle angegeben Algorithmus ermittelt werden.

[HIT92] beschreibt ein Chromatographieanalyseverfahren und ein System, das Analyseverfahren anwendet. Das Verfahren kann die zu erfassenden Bestandteile automatisch auf Basis des Meßergebnisses einer bekannten Probe bestimmen. Das System extrahiert die nötigen Informationen aus dem Chromatogramm, um Peaks zu identifizieren, die den zu erfassenden Bestandteilen entsprechen, und um Breiten von Zeitfenstern der Peaks einzustellen. Dann werden die so eingestellten Zeitfenster auf die Chromatogramme angewendet, die durch die Trennung der unbekannten Proben entstehen, um die Bestandteile zu identifizeiren, die in der unbekannten Probe enthalten sind. Durch Anwenden dieser Erfindung ist es für

einen Bediener nicht notwendig, die Retentionszeit jedes Bestandteils dem System einzugeben.

In [TOM92] wird ein zweistufiges Verfahren zur Erkennung und Auswertung der Peaks in einem Chromatogramm beschrieben. Das Verfahren nutzt ein auf neuronalen Netzen basierendes Verfahren zur Mustererkennung. Es können Mehrfachpeaks erkannt und ausgewertet werden. In einer ersten Stufe werden Daten, die zu einem Peak gehören, ermittelt. Dazu werden Extrema verschiedener Charakteristika (Maxima, Wendepunkte, ...) aus den Rohdaten extrahiert. Ein Mustererkennungsverfahren analysiert die Charakteristika der gefundenen Extrema. Die so gewonnen Informationen über Extrema der Peaks und Peakgruppen werden in eine Liste abgelegt. Anschließend werden die Fußpunkte der entsprechenden Peaks in den Rohdaten identifiziert.

Vielfach findet man Material darüber, wie Peaks auszuwerten sind ([LEST84], [DYS90]), aber nicht, wie man die Auswertung maschinell durchführen kann. Ebensowenig findet man Methoden zur Auswertung von Chromatogrammserien.

3 Auswertung mit Josephine

Im Rahmen eines Drittmittelprojektes der TU Ilmenau mit der ECH Elektrochmie Halle GmbH entstand die Software *Josephine*. Diese Software wertet Gaschromatogramme automatisch aus und stellt so Vorschläge für den Analytiker bereit.

In diesem Kapitel werden einige der dort eingesetzten Algorithmen zur Auswertung kurz beschrieben. Zur detaillierteren Beschreibung sei auf [WEI98] und [WEI98S] verwiesen.

3.1 Rauschen und Glättung

Das Rauschen beeinflußt die Peakdetektion nachhaltig. Deswegen wurde bei der Entwicklung von Josephine großer Wert darauf gelegt, Verfahren

- zum Ermitteln des Rauschwertes, also der Zahl die angibt, wie stark das Rauschen ist, und
- zum Glätten der Daten und damit zur Kompensation des Rauschens

zu implementieren.

3.1.1 Bestimmen des Rauschwertes

Üblicherweise wird das Rauschen durch Streuung der Datenpunkte in einem vom Nutzer markierten Bereich charakterisiert. Die Idee dieser Vorgehensweise ist, daß in einem idealen Chromatogramm die nicht zu einem Peak gehörenden Punkte einen konstanten Ordinatenwert besitzen. Betrachtet man also einen Ausschnitt eines Chromatogrammes, der keinen Peak enthält, so kann die Abweichung des Ordinatenwertes eines Punktes vom Mittelwert der Ordinatenwerte aller Punkte dieses Abschnittes als zufälliger Fehler gedeutet werden. Dementspechend kann die Streuung der Ordinatenwerte in einem peakfreien Abschnitt als Maß für die Stärke des Rauschens gedeutet werden.

Der so erhaltene Wert ist der Wert des Rauschens. Alternativ dazu kann man den Wert des Rauschens direkt eingeben. Das ist dann sinnvoll, wenn man diesen

Wert etwa aus den technischen Unterlagen des Systems kennt und auf Reproduzierbarkeit der Auswertung besonderen Wert legt. Da viele Auswerteparameter vom Rauschen abhängen, ist die Reproduzierbarkeit bei einer automatischen Rauschwertbestimmmung schwierig.

Um den Nutzer zu entlasten, ist alternativ dazu ein Verfahren implementiert worden, um das Rauschen mit einem gleitenden Fenster zu ermitteln: Ein Fenster festgelegter Breite gleitet über die Daten. Dabei wird stets die Streuung der Daten innerhalb des Fensters berechnet. Das Minimum aller dieser Werte ist das Rauschen.

3.1.2 Glättung der Daten

Für stark verrauschte Chromatogramme, wie das in Abbildung 3.1 gezeigte, muß ein Glättungsverfahren verwendet werden. Als ein nützliches Verfahren hat sich das in [SG64] beschriebene erwiesen. Dabei wird ein *gewichteter gleitender Mittelwert* mit dem *Gewichtsvektor* $\mathbf{w}^* \in \mathbb{R}^{2m+1}$ benutzt. Dieser Gewichtsvektor hat die Eigenschaft, daß

$$\mathbf{1}^T \mathbf{w}^* = 1.$$

Dabei ist

$$\mathbf{1} = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}^T \in \mathbb{R}^{2m+1}.$$

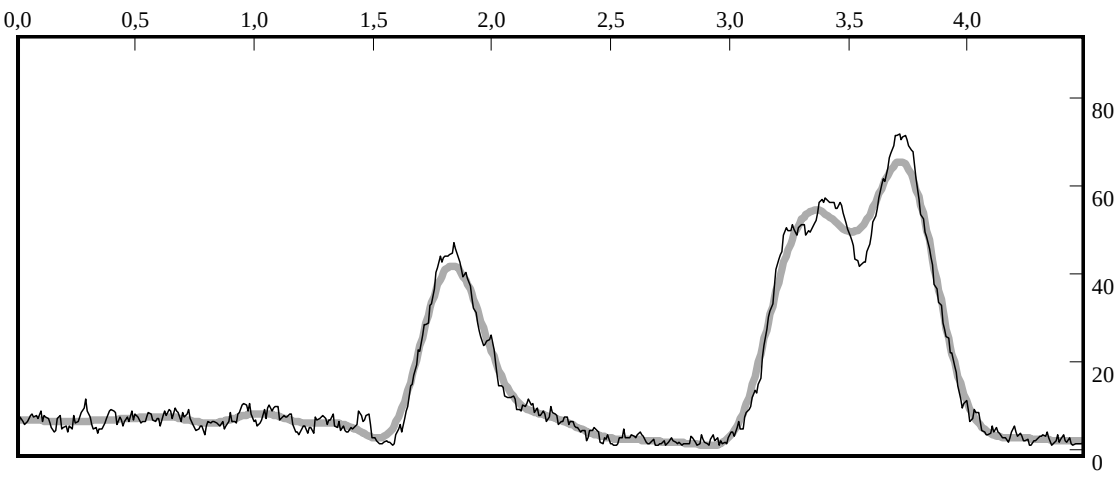


Abbildung 3.1: Ein verrauschtes Chromatogramm und die geglätteten Daten

Bei der Anwendung des Verfahrens gleitet ein $2m+1$ Indizes breites Fenster über die Werte des Chromatogrammes. Der geglättete Wert für einen Index i berechnet sich als Skalarprodukt des Gewichtsvektors mit dem aus den Werten der Indexe im

Datenfenster gebildeten Vektor. Bei l-maliger Anwendung eines Glättungsschrittes ergibt sich folgende Berechnungsvorschrift:

$$s_i^l = \begin{pmatrix} s_{i-m}^{l-1} & s_{i-m+1}^{l-1} & \dots & s_{i+m}^{l-1} \end{pmatrix} \frac{\mathbf{w}}{\mathbf{1}^T\mathbf{w}}, \tag{3.1}$$

wobei $s_j^l = s^0$ für $j < 0$, $s_j^l = s^{N-1}$ für $j \geq n$ und $\mathbf{s}^0 = \mathbf{v}$.

Für ein vorgegebenes $l = l^*$ wird dieses Verfahren abgebrochen und $\mathbf{s} = \mathbf{s}^{l^*}$ gesetzt. $\mathbf{s}$ ist dann der Vektor, der die geglätteten Datenwerte enthält. Je größer dabei l^* gewählt wird, um so mehr werden die Daten geglättet. Bei den meisten Chromarod-Chromatogrammen hat sich $l^* = 20$ als günstiger Wert erwiesen. Den Gewichtsvektor $\mathbf{w}$ kann man aus der Literatur ([SG64]) entnehmen, zum Beispiel für $m = 12$

$$\begin{aligned} \mathbf{w} = (&-253, -138, -33, 62, 147, 222, 287, 322, \\ &387, 422, 447, 462, 467, 462, 447, 422, 387, \\ &322, 287, 222, 147, 62, -33, -138, -253)^T. \end{aligned}$$

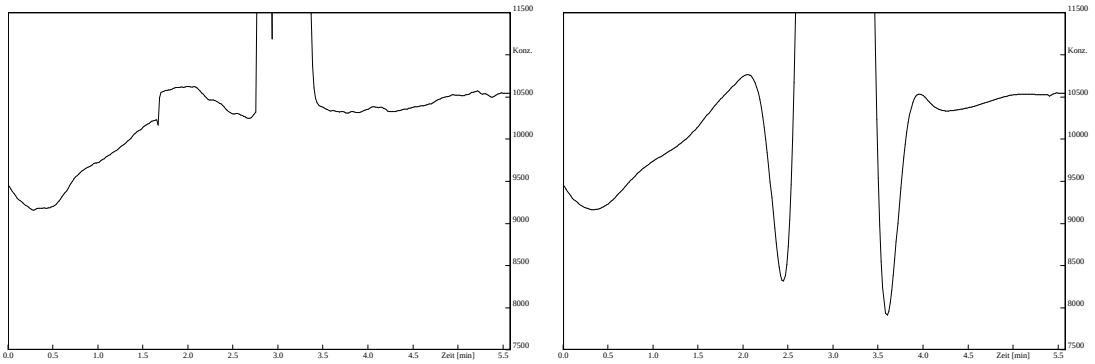


Abbildung 3.2: Ausprägung neuer Peaks nach Glättung

Für die Anwendung des Verfahrens gibt es einige Einschränkungen zu beachten:

- Die Peaks werden in ihrer Höhe und damit in ihrer Fläche verfälscht. Die Lage der Fußpunkte wird hingegen kaum verändert.
- Steile Flanken verursachen ein ,,Überschwingen" der geglätteten Daten. Das hat Auswirkungen auf Peaks mit steilen Flanken, insbesondere auf schmale, hohe Peaks. Wenn die Peaks schmal sind, können neue Peaks entstehen oder kleine Peaks eine stärkere Ausprägung erfahren, weil die Basislinie in der Nähe von großen, schmalen Peaks infolge des Überschwingens nach unten

verschoben wird (Abbildung 3.2). Die Peaks im Chromatogramm müssen also relativ breit sein, das heißt, sie müssen ausreichend viele Datenpunkte besitzen.

3.2 Peakdetektion

Die Peakdetektion arbeitet nach dem in Abschnitt 2.6.1 beschriebenem rekursiven Verfahren. Wenn für ein (Teil-) Chromatogramm das Maximum gefunden wurde, werden dazu passende Fußpunkte gesucht. Bei der Fußpunktsuche wird zunächst vom Maximumspunkt $(t_m, v_m)^T$ ausgehend nach links gegangen. Dabei werden jeweils Differenzen der Form

$$d_j = v_j - v_{j-k}, \quad k \leq j < m$$

betrachtet. Wenn es ein j gibt, so daß $d_j, d_{j-1}, d_{j-2} < \alpha$, dann ist $(t_j, v_j)^T$ der linke Fußpunkt. Das Verfahren bricht entweder ab, wenn ein Fußpunkt diese Bedingungen erfüllt oder wenn der linke Rand des Chromatogrammes erreicht ist. Verläuft die Suche nach dem linken Fußpunkt erfolgreich, wird der rechte Fußpunkt analog datektiert.

Das k gibt die Anzahl der Datenpunkte an, über denen die Differenz aufgespannt ist. Je stärker ein Chromatogramm verrauscht ist, um so größer muß k gewählt werden, damit Störungen keinen Einfluß haben. Der Parameter α ergibt sich aus dem Produkt des Rauschwertes mit einem benutzerdefinierten Faktor.

Wahlweise kann dabei über über den Ausgangsdaten oder über den geglätteten Daten operiert werden. Wenn über den geglätteten Daten gesucht wird, erfolgt die Projektion der Fußpunkte auf die Ausgangsdaten.

Im folgenden werden Spezialfälle behandelt, die während der Peakdetektion auftreten können.

3.2.1 Lotfällung

Es kann durchaus vorkommen, daß während des chromatographischen Trennvorgangs Peaks nicht sauber voneinander getrennt werden. Allerdings ist die Trennung oft genug ausreichend gut, um festzustellen, wo sich die Peaks und deren Retentionszeiten befinden. Solche schlecht getrennten Peaks treten häufig paarweise auf, jedoch können sie auch in größeren Gruppen vorkommen.

Diese schlecht getrennten Peaks erkennt der Analytiker daran, daß sich die inneren Fußpunkte der beiden beteiligten Peaks weit über der (gedachten) Basislinie befinden. Außerdem liegen solche Peaks sehr dicht beeinander.

Für die Integration ist es jedoch unerläßlich, die Flächen voneinander zu trennen. Ein unter den Analytikern beliebtes Verfahren ist die *Lotfällung*. Dabei werden die

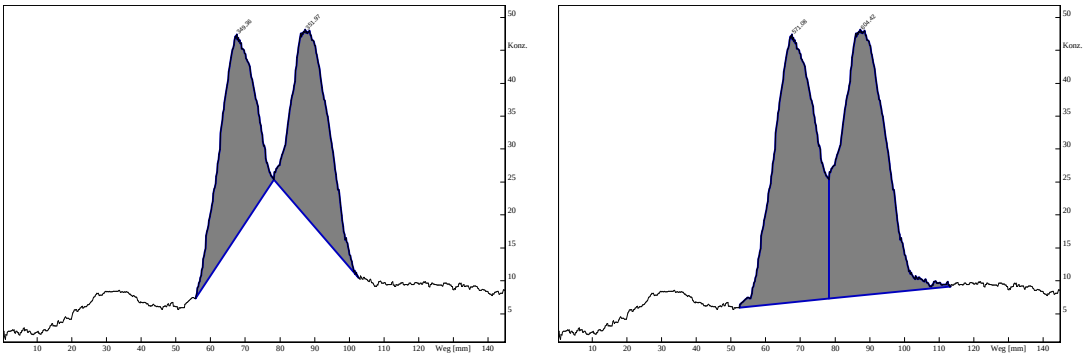


Abbildung 3.3: Zwei nicht getrennte Peaks ohne und mit Lotfällung

Talpunkte zwischen den schlecht getrennten Peaks ermittelt. Talpunkte sind die Punkte zwischen zwei sich überlappenden Peaks mit dem kleinsten Abstand zur Grundlinie. Von den Talpunkten aus wird das Lot auf die Grundlinie gefällt.

Um eine Lotfällung durchzuführen, werden durch Josephine zwei Entscheidungen getroffen:

1. Liegen die beiden zu untersuchenden Peaks dicht genug beeinander?

2. Liegt der Talpunkt zwischen diesen Peaks hoch genug?[1]

Die erste Frage ist leicht zu beantworten. Es wird nur überprüft, ob die Differenz des Indexes des Datenpunktes für den rechten Fußpunkt des linken Peaks und des Indexes des Datenpunktes für den linken Fußpunkt des rechten Peaks um einen vorgegebenen Wert unterschreitet. In Josephine wurde diese Grenze mit 5 angenommen.

Ist das erste Kriterium erfüllt, werden drei Lote auf die Verbindungsgeraden der beiden äußeren Fußpunkte gefällt: Von den beiden Retentionspunkten aus und vom Talpunkt aus. Dabei entstehen drei Strecken. Nun wird das Verhältnis der Länge der Strecke, die durch den Talpunkt geht, durch das Mittel der Streckenlängen, die durch die Retentionspunkte gehen, geteilt. Wenn das Verhältnis einen vom Benutzer eingestellten Wert überschreitet, wird das Lot gefällt.

Durch Lotfällung werden Fußpunkte verändert. Da die veränderten Fußpunkte nun nicht mehr auf dem Polygonzug durch die Datenpunkte liegen, werden durch Lotfällung entstandene Fußpunkt zusammen mit den Fußpunkten für den Peak abgespeichert.

[1] Je höher der Talpunkt liegt, um so größer ist der Fehler bei Integration. Dieser Fehler ist jedoch für den Analytiker akzeptabel.

3.2.2 Korrektur der Fußpunkte

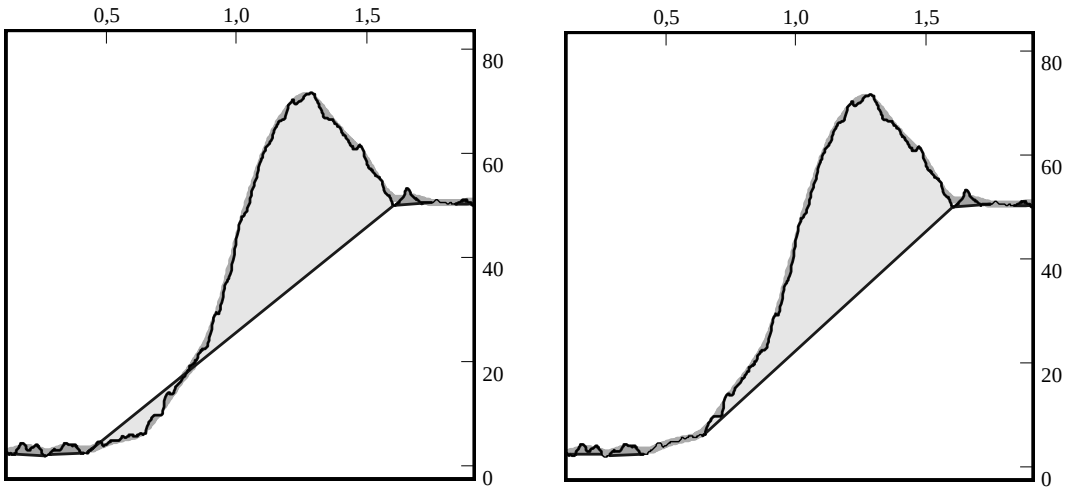


Abbildung 3.4: Negative Flächenanteile und die Korrektur der Fußpunkte

Projiziert man die Fußpunkte auf die originalen Daten, liegen oft Teile der Datenpunkte unterhalb der Verbindungslinie zwischen den Fußpunkten eines Peaks. Diese Teile werden als *negative Anteile* bezeichnet, da bei der Integration des Peaks diese Teile ein negatives Vorzeichen bekommen. Jedoch sind diese Anteile chemisch nicht zu erklären, so daß sie entfernt werden müssen. In Abbildung 3.4 ist eine solche Situation dargestellt. Wenn die Fußpunkte auf die Originaldaten projiziert werden, erhält man als Verbindungslinie zwischen den beiden Fußpunkten die Linie im linken Bild, während die Linie im rechten Bild offensichtlich die bessere Alternative darstellt.

Josephine korrigiert auf Wunsch die Fußpunkte so, daß negative Anteile entfernt werden. Dabei werden sukzessive die Fußpunkte nach innen verschoben, bis die Verbindungsstrecke der Fußpunkte keine Strecke zwischen zwei aufeinanderfolgenden Datenpunkten mehr schneidet.

3.2.3 Ausschluß von ,,fehlerhaften" Peaks

Um zu kleine, zu schmale oder zu breite Peaks auszuschließen, kann der Nutzer die minimale Peakhöhe eingeben. Weiterhin kann er das Verhältnis von zeitlicher Entfernung der Fußpunke zur Höhe einschränken.

Nachdem die Peaks detektiert worden sind, überprüft Josephine, ob die Peaks die vom Nutzer eingegebenen Kriterien erfüllen. Wenn nicht, werden diese Peaks gelöscht.

3.3 Erfahrungen aus der Arbeit mit Josephine

Gaschromatogramme mit geringen Störeinflüssen wertet Josephine akzeptabel aus. Wenn allerdings starke Störungen im Chromatogramm auftreten, funktioniert die Auswertung nicht immer zufriedenstellend. Insbesondere bei flachen Peaks und Plateaus ist die Detektion schwierig, weil in solchen Fällen das Abbruchkriterium für die Fußpunkte bereits für Punkte, welche eigentlich noch einer Peakflanke angehören, erfüllt sein kann. Dies kann dazu führen, daß ein Fußpunkt eines Peaks zu weit oben liegt.

Schlecht getrennte Peaks wertet Josephine grundsätzlich durch Lotfällung aus. Das ist insofern kritisch, als daß Schulterpeaks nicht richtig integriert werden können. Genausowenig können Peaks in der Nähe von Rampen behandelt werden. Wie sehr sich die Basisliniedrift auf die Detektion auswirkt, kann mit Hilfe des folgenden Beispiels verdeutlicht werden:

Man stelle sich ein Chromatogramm mit einem einzelnen Peak vor. Hinter dem Peak steigt die Basislinie über den Peak hinaus an. Das Maximum aller Werte wird am Ende das Chromatogrammes angenommen. Wenn Josephine ein solches Chromatogramm untersucht, vermutet es den einen Peak am Ende des Chromatogrammes. Allerdings gibt es für diesen Peak keinen rechten Fußpunkt, so daß ein solches Chromatogramm nicht ausgewertet werden kann. Da aber solche Chromatogramme in der Praxis nur selten vorkommen, hat das Verfahren dennoch seine Berechtigung.

Die meisten dieser Schwierigkeiten treten am Anfang oder am Ende des Chromatogrammes auf. Um diese Probleme zu umgehen, ist es möglich, ein Fenster zu markieren, in dem ausgewertet wird.

Ein Analytiker nimmt in der Regel Meßreihen gleichartiger Chromatogramme auf. Das heißt, daß er mehrere Chromatogramme mit ähnlichen Peaks auswertet. Mit Josephine muß er dazu jedes dieser Chromatogramme einzeln laden und bearbeiten. Wenn der Analytiker im Chromatogramm zum Beispiel falsch ausgewertete Peaks feststellt, ist es wahrscheinlich, daß diese auch in den anderen Chromatogrammen einer Serie auch zu finden sind. Allerdings hat er keine Möglichkeit, seine Korrekturen für die anderen Chromatogramme zu übernehmen. Insbesondere kann er Auswerteparameter, wie zum Beispiel die Mindesthöhe, nur für das gesamte Chromatogramm, nicht aber für einzelne Bereiche aus einem Chromatogramm festlegen.

Diese Kritikpunkte waren Anlaß dafür, Verfahren zu entwickeln, die solche Auswertungen von Serien unterstützen. Insbesondere war dabei eine wichtige Forderung, Korrekturen nicht für alle Chromatogramme einer Meßreihe zu wiederholen.

4 Lernfähige Verfahren

In diesem Kapitel werden anpassungsfähige Verfahren zur Peakdetektion beschrieben. Das Hauptaugenmerk dieser Verfahren ist darauf gerichtet, den Nutzer bei der Auswertung von Serien von Chromatogrammen zu unterstützen. Dabei soll er in der Lage sein, die Auswertung von Chromatogrammen an bestimmten Stellen – also für bestimmte Peaks – beeinflussen zu können.

4.1 Idee und Realisierungsmöglichkeiten

Zuerst wird betrachtet, wie der Analytiker mit dem Auswerteverfahren kommuniziert, um das Auswerteverhalten des Verfahrens zu ändern. Anschließend werden Ziele für die Verfahren formuliert und es wird das Einschätzen eines Peaks und seiner Fußpunkte aus der Sicht eines Analytikers geschildert.

4.1.1 Der Mensch-Maschine-Dialog

Wenn der Nutzer das Verfahren seinen Bedürfnissen anpassen will, tritt er mit der Maschine in einen Dialog. Der Anwender begutachtet das von der Maschine gelieferte Ergebnis und korrigiert wenn nötig. Diese Korrekturen registriert die Maschine und wiederholt mit den gewonnenen Informationen den Vorgang. Anschließend korrigiert der Nutzer erneut das Ergebnis. Dies wiederholt sich so lange, bis der Nutzer mit dem Ergebnis zufrieden ist.

Für die Auswertung von Chromatogrammen ist dieses Prinzip in Bild 4.1 dargestellt. Die Maschine wertet zunächst ein Chromatogramm aus und erzeugt eine Liste von Peaks. Anschließend begutachtet der Analytiker diese Liste. Nachdem er Korrekturen vorgenommen hat, wird erneut die Liste der Peaks aus den Rohdaten generiert. Die durchgeführten Korrekturen versetzen die Maschine in einen anderen Zustand, so daß man nach einer erneuten Auswertung eine veränderte Peakliste erhält.

4.1.2 Ziele

Es werden im folgenden die Zielstellungen dargestellt, unter welchen das lernfähige Auswerteverfahren entwickelt wurde. Dazu muß zuerst spezifiziert werden, welche

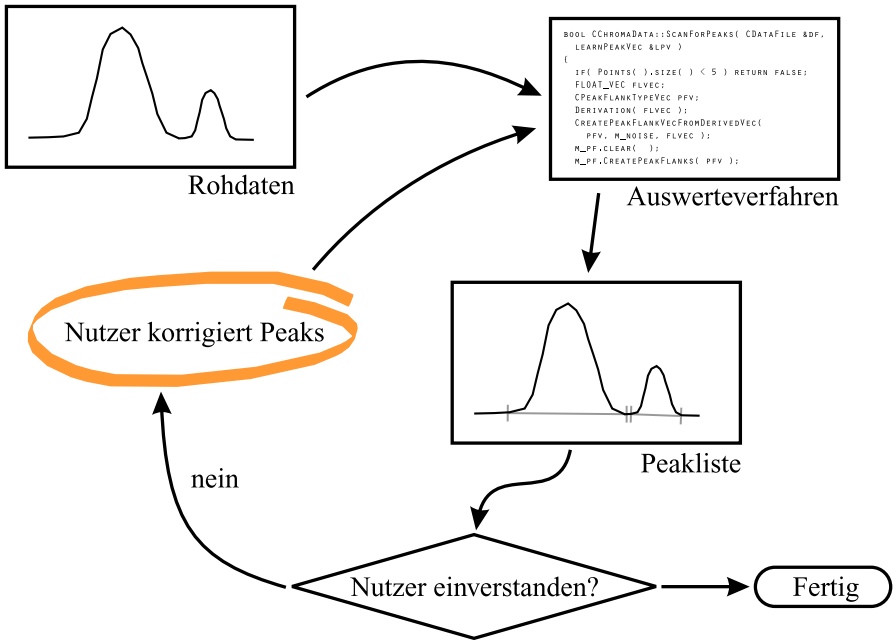


Abbildung 4.1: Prinzip des Mensch-Maschine-Dialogs

Korrekturen durch den Nutzer vorgenommen werden können. Es wird dabei vorausgesetzt, daß bei der Anwendung des Vefahrens dem Nutzer zuerst eine Liste von Kandidaten für Peaks vorgeschlagen wird. Dabei sind die Kandidaten dieser Liste vom Auswerteverfahren als „gültig“ oder „ungültig“ markiert worden, nicht ausreichend getrennte Peaks bereits durch Lotfällung voneinander getrennt oder als Schulterpeak markiert. Diese Liste wird grafisch dargestellt und der Nutzer kann dann

(i) Peaks aus dieser Liste abwählen, das heißt sie als „ungültig“ markieren,

(ii) entscheiden, ob nebeneinander liegende Peaks ausreichend getrennt sind oder nicht,

(iii) im Fall nicht ausreichend getrennter Peaks entscheiden, ob diese Peaks durch Lotfällung getrennt werden oder einer dieser Peaks ein Schulterpeak ist, und

(iv) die Fußpunkte von Peaks verschieben.

Aus den genannten Zielstellungen ergeben sich unter anderem folgende Schlußfolgerungen für die Realisierung des Verfahrens:

Bei den unter (i) - (iii) genannten Korrekturmöglichkeiten handelt es sich um Ja-Nein-Entscheidungen: Bleibt der Peak als gültig markiert oder nicht? Sind die

Peaks ausreichend getrennt oder nicht? Sollen nicht ausreichend getrennte Peaks durch Lotfällung voneinander getrennt oder einer dieser Peaks als Schulterpeak markiert werden? Dem Nutzer muß außerdem die Möglichkeit gegeben werden, vom Verfahren gemachte Korrekturen wieder rückgängig machen zu können. Er muß also die Möglichkeiten haben,

- Peakkandidaten aus dieser Liste als „gültig“ markieren,
- die Trennung zweier Peaks aufzuheben,
- die Art der Trennung (Lotfällung, Schulterpeak) umzukehren.

Durch den unter (iv) beschriebenen Korrektureingriff wird die Auswahl eines Fußpunktes aus einer Menge von Punkten beeinflußt. Dieser Auswahl liegt ein Optimierungsverfahren mit mehreren Zielfunktionen zugrunde. Die unterschiedliche Natur dieser beiden Probleme rechtfertigt den Einsatz unterschiedlicher Lernverfahren.

Aus der Natur eines lernfähigen Verfahrens ergibt sich, daß während des Lernvorgangs Peakkandidaten der Liste verändert werden können, welche der Nutzer im Ausgangszustand belassen möchte. Um solche unerwünschten Effekte weitestgehend zu unterdrücken, wird gefordert, daß sich (lokale) Korrekturen nicht auf die Auswertung des gesamten Chromatogramms auswirken.

4.1.3 Peakevaluierung

Es gilt nun, Kriterien zu finden, mit deren Hilfe man einschätzen kann, ob ein Peak gültig („gut“) ist oder nicht.

Viele Applikationen gehen davon aus, daß man die Form eines Peaks einschätzen muß. Wenn man eine bestimmte Klasse von Chromatogrammen betrachtet, ist dieser Ansatz erfolgversprechend. Bei vielen chromatographischen Verfahren entspricht die Form eines Peaks der einer Glockenkurve oder EMG-Funktion.

Das zu entwickelnde Verfahren soll sich aber nicht auf typische Formen beschränken. Es soll zum Beispiel auch Peaks erkennen, bei denen der Chromatograph im gesättigten Bereich betrieben wurde. Nach vielen Gesprächen mit Analytikern haben sich dabei folgende wesentliche Merkmale herauskristallisiert:

Die Höhe Entscheidend dafür, ob ein Peak gültig ist oder nicht, ist die Höhe. Wenn man die Höhe im Verhältnis zum Rauschen betrachtet, kann man einschätzen, wie gut ein Peak detektiert worden ist. Ein Peak muß mindestens die drei- bis vierfache Höhe des Rauschwertes besitzen.

Das Verhältnis von Fläche zum Quadrat der Höhe Das Verhältnis dieser beiden Wete gibt an, ob der Peak „wohlgeformt“ ist. Dieses Kriterium kann

dazu benutzt werden, um zum einen Spikes zu erkennen, zum anderen aber auch, um Erhebungen, die keine Peaks sind, zu detektieren.

Das Verhältnis von Unsymmetrie zum Quadrat der Höhe Das Verhältnis dieser beiden Wete gibt an, wie unsymmetrisch der Peak ist. Dabei ist die Unsymmetrie die Differenz der Teilflächen rechts und links von der Retentionszeit. Ist dieser Wert positiv, hat der Peak ein Tailing auf der rechten Seite, ist er negativ, dann hat er ein Fronting. Wenn ein Peak allzu unsymmetrisch ist, deutet dies darauf hin, daß er aufgrund von Detektionsfehlern verfälscht worden ist.

Die Breite auf halber Peakhöhe Wenn man die Breite des Peaks auf der halben Höhe betrachtet, kann man ebenso Aussagen über die Form eines Peaks treffen. Alledings ist dieser Wert bei schlecht aufgezeichneten oder schmalen Peaks nur sehr ungenau zu ermitteln. Deswegen wird dieses Kriterium im weiteren Verlauf nicht mehr berücksichtet.

Es bieten sich zwei unterschiedliche Methoden an, um zu entscheiden, ob ein Peak gültig ist oder nicht. Zum einen kann man auf Einhaltung aller diese Eigenschaften überprüfen und den Peak dann und nur dann für gut befinden, wenn er alle Eigenschaften genügend gut erfüllt. Der andere Ansatz geht davon aus, daß der Peak auch Kriterien nicht gut erfüllen kann, wenn er den anderen Kriterien um so besser genügt.

4.1.4 Lage der Fußpunkte

Josephine ist in der Lage, Fußpunkte zu korrigieren, um negative Flächenanteile des Peaks zu vermeiden. Das zu entwickelnde Verfahren soll darüber hinaus die Fußpunkte aufgrund weiterer Kriterien korrigieren können.

Zum einen betrifft das Fälle, in denen ein Fußpunkt zu weit nach außen gezogen worden ist, etwa weil das Tailing des Peaks stark ausgeprägt ist. Das ist vom Analytiker nicht immer erwünscht. Ein ähnliches Problem ergibt sich, wenn ein Peak in der Nähe einer Rampe befindet. Dann wird ein „dummer“ Algorithmus einen Fußpunkt unterhalb und den anderen oberhalb der Rampe setzen, was aber ein vollkommen falsches Bild bei der Integration liefert.

Ein Chemiker nennt für einen Fußpunkt folgende Kriterien:

- Der Fußunkt soll in einem lokalen Minimum der Daten liegen oder wenigstens in der Nähe eines solchen.
- Der Fußunkt soll auf oder in der Nähe der Basislinie liegen.
- Der Fußunkt soll so liegen, daß keine negativen Flächenanteile entstehen.

- Beide Fußpunkte des Peaks respektive die beiden äußeren Fußpunkte einer Gruppe sollen auf gleichem Niveau liegen.

Man sieht schnell ein, daß sich die Kriterien untereinander widersprechen können. Wenn sich etwa ein Peak an einer ansteigenden Rampe befindet, kann sich der linke Fußpunkt entweder in der Nähe eines lokalen Minimums oder auf gleicher Höhe mit dem anderen Fußpunkt befinden. Andere Kriterien können manchmal gar nicht erfüllt werden. Zum Beispiel kann der rechte Fußpunkt eines Peaks auf einer ansteigenden Basislinie nicht in einem lokalen Minimum liegen.

Wie man sieht, hat man es hier wiederum mit Kriterien zu tun, die nicht einfach mit „ist erfüllt“ oder „ist nicht erfüllt“ eingeschätzt werden dürfen. Dies führt zu einem Algorithmus, der mit unscharfen Aussagen umgehen kann.

4.2 Zeitbasierte Auswertung

Die Erfahrungen mit Josephine haben gezeigt, daß die Parameter für die Kritieren sowohl der Peaks als auch der Fußpunkte zeitlich variieren. Dafür seien drei Gründe als Beispiel genannt:

- Die Basislinie driftet nur am Ende eines Chromatogrammes nach oben. Driftet die Basislinie in einem Chromatogramm, können sich die Fußpunkte der Peaks nicht auf gleicher Höhe befinden, während im vorderen Teil des Chromatogrammes die Forderung nach gleicher Höhe der Fußpunkte berechtigt ist.

- Lösungsmittelpeaks treten in einem Chromatogramm meist nur im vorderen Teil auf. Wenn der Analytiker diese Peaks nicht mit in der Peakliste haben möchte, kann er für diesen Bereich alle Peaks, die eine bestimmte Höhe oder eine bestimmte Fläche überschreiten, als ungültig markieren.

- Bestimmte Fehler, wie zum Beispiel negative Peaks, treten nur an bestimmten Stellen im Chromatogramm auf.

Aus diesen Gründen sollen die Parameter für die Kriterien der Peaks und der Fußpunkte zeitabhängig ermittelt werden. Als wichtige Größe zur Ermittlung der zeitabhängigen Peakparameter beietet sich die Retentionszeit an. Wenn beispielsweise die Mindesthöhe eines Peaks durch die Funktion $h(t)$ beschrieben ist, dann ist für den Peak p mit der Retentionszeit t_R die Mindesthöhe $h(t_R)$. Aufgrund eines zeitlichen Versatzes bei der Aufzeichnung von Chromatogrammen ist es wichtig, daß diese Funktionen stetig sind.

4.3 Anwendung eines Fuzzy-ähnlichen Verfahrens

In *Josephine* wurde erfolgreich folgender Entscheidungsablauf verwendet, um einen Peak auf Qualität zu überprüfen:

1. Ist der Peak hoch genug?

2. Liegt das Verhältnis von Höhe zu Breite innerhalb eines vorgegebenen Bereiches?

Wenn beide Fragen bejaht werden, ist der Peak gültig.

Dieses Prinzip soll nun in einer erweiterten Form verwendet werden. Manchmal ist ein Peak sehr hoch, hat aber eine seltsame Form, weil er etwa durch Sättigung abgeschnitten wurde. Dieser Peak soll aber trotzdem als gültig erkannt werden. Ebenso soll ein Peak erkannt werden, dessen Gestalt zwar der definierten Form entspricht, aber der das Peak-Höhen-Kriterium nicht exakt erfüllt.

4.3.1 Fuzzy-Sets

Wenn man mit Hilfe der Fuzzy-Logic Entscheider entwickelt, ordnet man einer linguistischen Variable ein Fuzzy-Set zu. Für das Verhältnis von Peakhöhe zum Systemrauschen könnte man beispielsweise die linguistischen Variablen „Peak ist klein“, „Peak ist mittelhoch“ und „Peak ist hoch“ einführen. Jeder linguistischen Variable kann dann ein Wert (Fuzzy-Wert) zwischen 0 und 1 zugewiesen werden. Dieser Wert gibt an, inwieweit die Aussage der linguistischen Variable erfüllt ist (Bild 4.2).

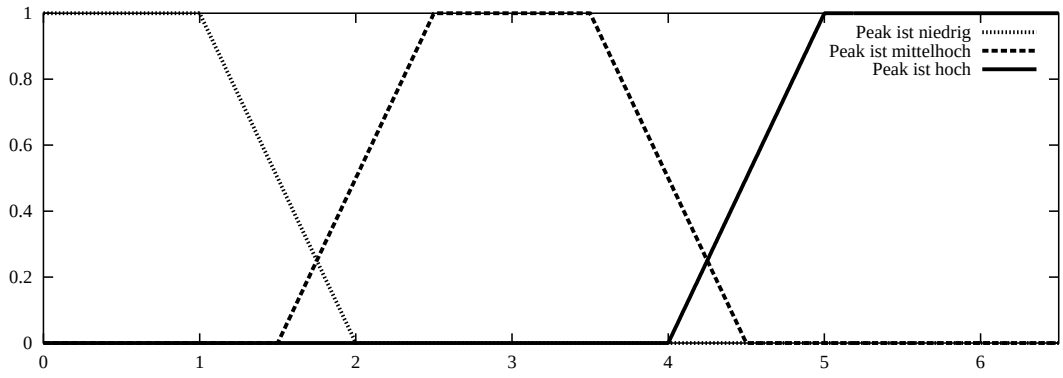


Abbildung 4.2: Fuzzy-Sets für die linguistischen Variablen „Peak ist klein“, „Peak ist mittelhoch“ und „Peak ist hoch“

Hat man für andere Kriterien weitere linguistische Variablen eingeführt, kann man Regeln aufstellen, wie „Wenn der Peak hoch ist und das Verhältnis von Fläche zu Höhenquadrat mittel, dann ist der Peak gültig."

Dem zu entwickelnden Verfahren soll dieses Prinzip zugrunde gelegt werden. Es wird aber für jedes Kriterium nur eine linguistische Variable eingeführt. Diese Variable gibt an, wie gut das Kriterium erfüllt ist. Beispielsweise entspricht „Peakhöhe ist gut" dem „Peak ist hoch" aus Bild 4.2. Die Bewertung eines Peaks kann dann auf eine Regel reduziert werden.

4.3.2 Struktur des Entscheiders

Der Entscheider ist eine Erweiterung des im Abschnitt 3.2.3 auf Seite 27 vorgestellten Verfahrens, mit dem Josephine zu breite oder zu schmale Peaks aus der Liste der Peaks entfernen kann.

In den Entscheider sollen die beiden dort genannten Kritierien einfließen. Die Bewertung der Aussage „Peak ist hoch" ist schon im letzten Abschnitt erklärt worden. Die Aussage „Das Verhältnis von Höhe zu Breite liegt innerhalb eines vorgegebenen Bereiches" soll in einer veränderten Form umgesetzt werden.

Da die Breite eines Peaks schwer zu bestimmen ist, soll dieses Kriterium durch „Das Verhältnis von Peakfläche zum Höhenquadrat ist gut" ersetzt werden, was eine ähnliche Aussagekraft hat. Dieses Kriterium soll so bewertet werden, daß hohe und niedrige Werte des Verhältnisses eine schlechte und mittlere eine gute Bewertung erhalten. Somit erhält man für die Bewertungsfunktion ein Trapez (Bild 4.3).

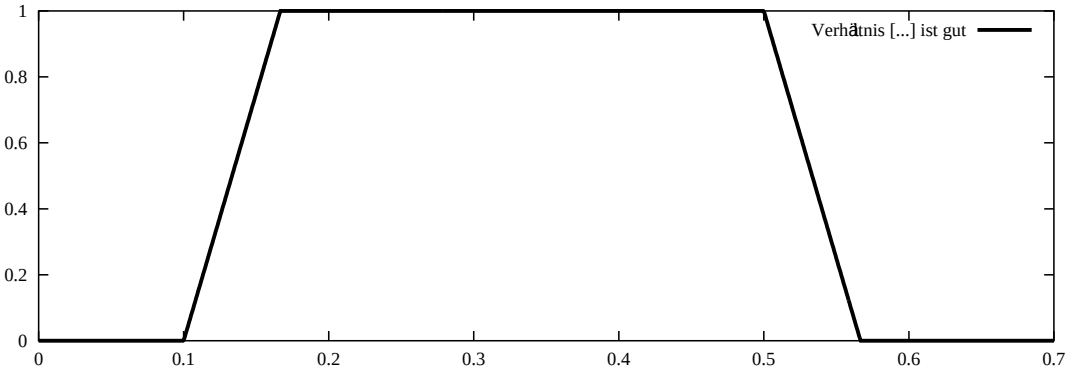


Abbildung 4.3: Bewertung des Kriteriums „Das Verhältnis von Peakfläche zum Höhenquadrat ist gut"

Während man beim Höhenkriterium vom Analytiker eine klare Aussage über

den Verlauf der Funktion erhalten kann, ist es an dieser Stelle schwieriger, so daß man diese Werte empirisch ermitteln muß.

Die Bewertungen der beiden Funktionen sollen nun zu einer Aussage über den Peak verknüpft werden. Dafür eignet sich der als *Fuzzy-AND* bezeichnete Operator

$$\mu_{\text{and}}(a, b, p) = p \min(a, b) + \frac{1-p}{2}(a + b) \tag{4.1}$$

Dabei sind a und b die Bewertungen zweier Aussagen. Für den Parameter p gilt $0 \leq p \leq 1$. Ist $p = 1$, ist das Ergebnis die kleinere der beiden Bewertungen, eine schlecht bewertet Aussage wirkt somit limitierend. Ist hingegen $p = 0$, ist das Ergebnis das arithmetische Mittel beider Bewertungen, damit wirkt eine gut bewertete Aussage kompensierend gegenüber einer schlecht bewerteten Aussage. Als günstiger Wert für den Parameter hat sich $p = 0.1$ herausgestellt.

4.4 Lernen von Parametern

Verfahren zur Mustererkennung kann man durch einen überwachten Lernvorgang trainieren. Dabei benutzt man Neuronale Netze, denen ein Eingangsmuster präsentiert wird. Das Netz ordnet diesem Eingangsmuster eine Ausgangsaktivierung zu. Der „Lehrer“ muß dann entscheiden, wie sehr die Ausgangsaktivierung vom gewünschten Resultat abweicht. Durch diese Entscheidung werden Parameter, die den Algorithmus steuern, verändert.

Die Kriterien zur Bewertung der Peaks besitzen verschiedene Parameter, welche die Bewertung der entsprechenden Aussagen beeinflussen, nämlich die Eckpunkte der Rampen[1] und Trapeze. Diese Parameter werden während eines Lernvorganges angepaßt.

Dieser Prozeß ist ein überwachter Lernvorgang. Während das Verfahren Lösungen anbietet, muß der Nutzer entscheiden ob diese Lösungen richtig sind oder nicht.

4.4.1 Struktur des Lernverfahrens

In die Entscheidung über die Qualität eines Peaks fließen mehrere Kriterien ein. Da die Parameter, um die Kriterien für einen gegebenen Peak auszuwerten, unabhängig voneinander gelernt werden[2], wird die Struktur am Beispiel der minimalen Peakhöhe erläutert. Die Struktur der anderen Funktionen ist dann analog

[1] Der Begriff *Rampe* wird mit zwei Bedeutungen benutzt: Einmal bezeichnet *Rampe* eine Störung in einem Chromatogramm und zum anderen ist mit *Rampe* der Anstieg der Bewertungsfunktion gemeint

[2] Diese Entscheidung wurde getroffen, um die Struktur des Verfahrens einfach zu halten und hat sich in den Tests bewährt.

aufgebaut. Zur Beschreibung einer Rampe sei noch einmal auf den Kurvenverlauf der Bewertung der Aussage „Peak ist hoch“ aus Bild 4.2 verwiesen.

Peaks sollen zeitabhängig ausgewertet werden. Somit ist die Bewertungsfunktion $\mu_{\mathrm{h_{min}}}(t,h)$ sowohl von der Zeit t als auch von der Peakhöhe h abhängig.

Betrachtet man $\mu_{\mathrm{h_{min}}}(t,h)$ für ein festes t, erhält man eine Bewertungsfunktion wie in Bild 4.2, läßt man hingegen h fest, erhält man eine *Lernfunktion*. Es gilt

$$\mu_{\mathrm{h_{min}}}(t,h) = \begin{cases} 0 & \text{falls } h \leq h^0(t), \\ 1 & \text{falls } h \geq h^1(t), \\ \frac{h-h^0(t)}{h^1(t)-h^0(t)} & \text{sonst.} \end{cases} \tag{4.2}$$

$h^0(t)$ und $h^1(t)$ sind dabei die Stellen der Bewertungsfunktion zum Zeitpunkt t, an denen sich die untere und obere Ecke der Rampe befinden. $h^0(t)$ gibt somit den größten Wert der Höhe an, dessen Bewertung eine 0 ergibt. $h^1(t)$ gibt den kleinsten Wert der Höhe an, dessen Bewertung eine 1 ergibt. Es ist klar, daß für jedes t gelten muß $h^0(t) \leq h^1(t)$. Die Differenz $h^1(t) - h^0(t)$ ist ein Maß dafür, wie scharf die Bewertung der Aussage „Die Höhe des Peaks ist gut.“ zum Zeitpunkt t ist.

„Lernen“ heißt dann nichts anderes, als die Funktionen $h^0(t)$ und $h^1(t)$ zu verändern. Diese Funktionen sollen bei einem Lernvorgang lokal veränderbar sein können. Das bedeutet, das sich Änderungen nur auf einen zeitlich begrenzten Bereich auswirken sollen.

Solche Funktionen benutzt man häufig im CAGD-Bereich[3] ([FAR94]). Am populärsten sind wohl die NURBS-Kurven[4], mit denen ein Designer nahezu jeden erdenklichen Kurvenverlauf konstruieren kann. Um eine Kurve zu verändern, benutzt er sogenannte *Kontrollpunkte*. Verschiebt der Designer einen Kontrollpunkt, wird die Kurvenverlauf um diesen Kontrollpunkt herum ebenfalls in Richtung der neuen Position des Kontrollpunktes verschoben. Eine solche Kurve ist im wesentlichen[5] durch die Lage der Kontrollpunkte definiert. Die Kontrollpunkte bilden einen Polygon, das ganz grob den Kurvenverlauf beschreibt.

Dieses Prinzip soll hier angewandt werden: Zur Beschreibung der Funktionen werden Polygone verwendet. Das Prinzip soll anhand der Struktur der Funktion $h^1(t)$ beschrieben werden:

- Man wähle eine Zahl $c \in \mathbb{R}$ mit $c > 0$. Man nehme eine Folge $(h^1_j)_{j\geq 0}$ reeller Zahlen, wobei $h^1_j = h^1_{\min}$ für $j \geq 0$. $h^1_{\min}$ ist die vorher festgelegte minimale Peakhöhe.

[3] *Computer aided geometric design* (Computergestützter geometrischer Entwurf)

[4] *Nun uniform rational B-Splines*

[5] Ein Designer kann außerdem weitere Parameter der Kontrollpunkte angeben, die den Kurvenverlauf zwar lokal beeinflussen, den wesentlichen Verlauf der Kurve aber nicht ändern. Siehe hierzu die angegebene Literatur.

- Die Folgenglieder betrachte man auf einer Zeitachse verteilt, und zwar so, daß h_j^1 über dem Zeitpunkt $c \cdot j$ liegt. Damit sind $(c \cdot j, h_j^1)^T$ Knoten eines Polygones.
- $h^1(t)$ berechnet sich nach folgender Vorschrift:[6]

$$h^1(t) = \begin{cases} h_0^1, & \text{falls t ¡ 0} \\ \left(h_{j+1}^1 - h_j^1\right) \frac{t-j}{c} & \text{sonst, wobei } j = \left\lfloor \frac{t}{c} \right\rfloor \end{cases} \tag{4.3}$$

Weil zu Beginn alle Folgenglieder mit dem Wert $h_{\min}^1$ initialisiert worden sind, ist $h^1(t) = h_{\min}^1$ für alle t. Die Knoten für die Funktion $h^0(t)$ kann man entweder auch abspeichern oder man ermittelt sie aus den Knoten der Funktion $h^1(t)$. Die zweite Variante erscheint günstiger, weil $h^0(t)$ von $h^1(t)$ abhängt.

Es ist also zu klären, wie steil die Rampe der Bewertungsfunktion gewählt werden soll. Diese Frage ist äquivalent zur Frage, wie groß $d_j^h = h_j^1 - h_j^0$ zu wählen ist. Dazu seien zwei Möglichkeiten genannt:

- Man setzt $d_j^h = d^h = \text{const}$. In diesem Fall braucht man nur die Knoten für $h^1(t)$ zu speichern. Für $d^h = 0$ erhält man den Spezialfall eines Sprunges in der Kurve: Ist die Höhe eines Peaks mit der Retentionszeit t kleiner als $h^0(t)$, ist die Bewertung der Peakhöhe 0, ansonsten 1.
- Man führt über jeden Knoten j von h_j^1 eine Statistik, wie sehr er bisher verändert worden ist. Von allen Werten, die h_j^1 jemals angenommen hat, wird die Standardabweichung σ_j^h berechet. Man setzt $d_j^h = \sigma_j^h$. Um die Statistik zu führen, müssen für jeden Knoten weitere Werte (Summe aller bisher angenommen Werte, Summe der Quadrate aller bisher angenommen Werte, Anzahl aller angenommenen Werte) gespeichert werden.

Damit ist die Struktur der beiden Funktionen $h^0(t)$ und $h^1(t)$ erklärt. Obwohl es möglich wäre, für jede Lernfunktion die entsprechenden Knoten unterschiedlich dicht über der Zeitachse zu verteilen, soll diese Verteilung für alle Lernfunktionen konstant sein. Es stellt sich nun die Frage, wie die Größe c zu wählen ist. Je kleiner c ist, um so feiner wird das Polygon. Es dürfte aber auf der Hand liegen, daß es nicht sinnvoll ist, c nicht kleiner zu wählen als zeitliche Auflösung des Chromatogrammes.

4.4.2 Ein Lernschritt

Es wird ein Lernschritt anhand der Struktur der Funktion $h^1(t)$, die eben beschrieben wurde, erläutert. Für die anderen Bewertungsunktionen funktioniert der Lernschritt analog.

[6] $h^1(t)$ wird hier aus technischen Gründen auch für negative t definiert.

Angenommen, der Analytiker möchte einen als ungültig markierten Peak mit der Retentionszeit t, dem linken Fußpunkt beim Zeitpunkt t_l und dem rechten Fußpunkt beim Zeitpunkt t_r als gültig markieren. Die Höhe dieses Peaks sei h.

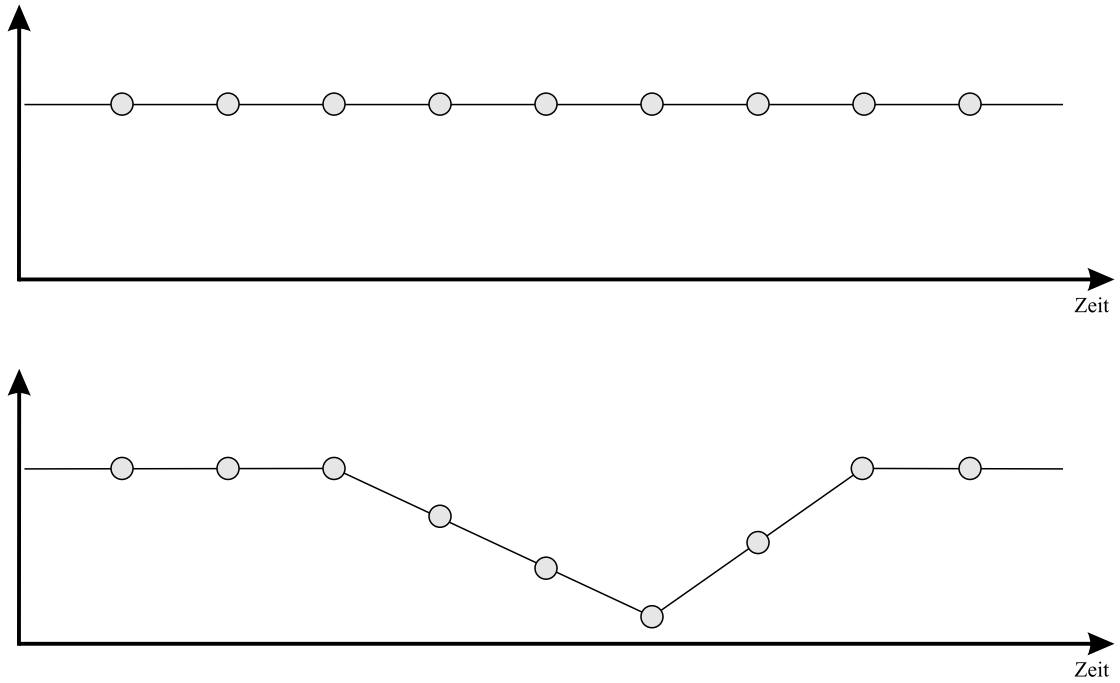


Abbildung 4.4: Beispiel des Verlaufs einer Lernfunktion vor (oben) und nach einem Lernschritt (unten)

Dann müssen einige Knoten von $h^1(t)$ verschoben werden. Dabei wird folgende Regel angewendet:

$$h_j^1 ::= \lambda_j h_j^1 + (1 - \lambda_j)h. \tag{4.4}$$

Behandelt man die Knoten für $h^0(t)$ unabhängig von den Knoten für $h^1(t)$, müssen diese natürlich in gleicher Weise verändert werden. Berechntet man die Knoten für $h^0(t)$ wie im letzten Abschnitt beschrieben aus den Knoten für $h^1(t)$, hat man dieses Problem nicht.

Es werden drei Ansätze vorgestellt, wie man λ_j ermitteln kann:

- Man gibt sich einen *Lernradius* $r > 0$ vor. Für alle j, für die $c \cdot j \notin [t-r, t+r]$, ist $\lambda_j = 1$. Ansonsten ist $\lambda_j = \frac{|c \cdot j - t|}{r}$.

- Man benutzt zwei Lernradien $r_l = t - t_l$ und $r_r = t_r - t$. Für alle j, für die

$c \cdot j \notin [t - r_l, t + r_r]$, ist $\lambda_j = 1$. Ansonsten gilt folgende Vorschrift:

$$\lambda_j = \begin{cases} \frac{c \cdot j - t}{r_r}, & \text{falls } c \cdot j \geq t, \\ \frac{t - c \cdot j}{r_l}, & \text{falls } c \cdot j < t \end{cases}$$

Dieses Prinzip ist in Bild 4.4 verdeutlicht.

- Man benutzt 4 Punkte t_l^0, t_l^1, t_r^1 und t_r^0 auf der Zeitachse, wobei $t_l^0 \leq t_l^1 \leq t \leq t_r^1 \leq t_r^0$. Durch diese 4 Punkte ist es möglich, in einem trapezförmigen Bereich um t herum zu lernen, wenn man für λ_j folgende Vorschrift anwendet:

$$\lambda_j = \begin{cases} 1 - \frac{c \cdot j - t_l^0}{t_l^1 - t_l^0}, & \text{falls } t_l^0 < c \cdot j < t_l^1, \\ 0, & \text{falls } t_l^1 \leq c \cdot j \leq t_r^1, \\ \frac{c \cdot j - t_r^1}{t_r^0 - t_r^1}, & \text{falls } t_r^1 < c \cdot j < t_r^0, \\ 1, & \text{sonst.} \end{cases}$$

4.5 Suche nach dem „besten“ Fußpunkt

Die unter Abschnitt 4.1.4 genannten Kriterien für die Fußpunkte sollen nun bewertet werden. Wie erwähnt, können sich einige der Kriterien untereinander widersprechen. Aus diesem Grund sollen die Kriterien gewichtet werden. Ein Kriterium mit einem größeren Gewicht ist höher zu bewerten als eines mit einem kleinen Gewicht.

4.5.1 Berechnung der Kriterien für einen Fußpunkt

Bevor die Kriterien ausgewertet werden können, müssen sie zunächst berechnet werden. Diese berechneten Werte sollen x^0, x^1, x^2 und x^3 heißen. Die Berechnung geschieht mit 4 Funktionen $F_p^0(i)$, $F_p^1(i)$, $F_p^2(i)$ und $F_p^3(i)$. Dabei ist i der Index des Datenpunktes und p der Peak, für den die Werte berechnet werden.

Für den Fußpunkt mit den Koordinaten $(t_i, v_i)^T$ soll das nun bezüglich des Peaks p mit den Fußpunkten $(t_l, v_l)^T$ und $(t_r, v_r)^T$ geschehen. Wenn sich der Peak in einer Gruppe befindet, dann betrachte man die Peakgruppe p mit den äußeren Fußpunkten $(t_l, v_l)^T$ und $(t_r, v_r)^T$. Dann ist entweder $i = l$ oder $i = r$.

Berechnung, ob sich der Punkt in einem lokalen Minimum befindet

Zunächst wird für den Bereich des Chromatogrammes, der vom Peak überdeckt wird, der maximale Wert $\Delta_{\max}$ des Beträge der diskreten Ableitungen ermittelt.

Anschließend wird die diskrete Ableitung

$$\delta_j = \frac{v_{j+1} - v_j}{t_{j+1} - t_j}$$

für $j = i - 1$ und $j = i$ ermittelt[7]. Nun ist $\min(-\delta_{j-1}, \delta_j)$ positiv, wenn der Punkt j ein lokales Minimum ist. Um negative Werte für x^0 zu vermeiden, ist

$$x^0 = F_p^0(i) = 1 + \frac{\min(-\delta_{i-1}, \delta_i)}{\Delta_{\max}}. \tag{4.5}$$

Berechnung der Nähe zur Basislinie

Um zu berechnen, wie nah sich der Fußpunkt an der Basislinie befindet, kann man überprüfen, wie nah sich der Fußpunkt am Minimum aller Daten $v_{\min}$ befindet. Damit die Größe normiert vorliegt, wird durch die vertikale Ausdehnung des Chromatogrammes dividiert. Das Maß für die Nähe zur Basislinie sei x^1.

$$x^1 = F_p^1(i) = 1 - \frac{v_i - v_{\min}}{v_{\max} - v_{\min}} \tag{4.6}$$

Alternativ zum Minimum und zur Ausdehnung des gesamten Chromatogrammes kann man das Minimum und die Ausdehnung eines Teilbereiches ermitteln. Dieser Teilbereich kann zum Beispiel nach links und nach rechts durch benachbarte Peaks (oder die Grenzen des Chromatogrammes, sofern keine Nachbarpeaks vorhanden sind) begrenzt werden.

Berechnung der negativen Flächenanteile

Zur Berechnung der negativen Flächenanteile wird der Peak wie im Abschnitt 2.6.2 auf Seite 15 beschrieben in Trapeze zerlegt. Von den vier Seiten eines solchen Trapezes liegen zwei auf Loten, eine auf der Verbindungslinie der Fußpunkte und eine ist die Verbindungsstrecke zwischen zwei aufeinanderfolgenden Datenpunkten.

Die Endpunkte der Verbindungsstrecke der aufeinanderfolgenden Datenpunkte haben die Koordinaten $(t_j, v_j)^T$ und $(t_{j+1}, v_{j+1})^T$. Die Enden der Seite, die auf der Verbindung der Fußpunkte liegt, haben dann die Koordinaten $(t_j, b_j)^T$ und $(t_{j+1}, b_{j+1})^T$, wobei man b_j und b_{j+1} leicht mit Hilfe der Geradengleichung der Geraden b durch die Fußpunkte berechnen kann. Gesucht ist nun der negative Flächenanteil F_j^{neg} dieses Trapezes.

Sei nun $d_j = v_j - b_j$ und $d_{j+1} = v_{j+1} - b_{j+1}$. Dann gibt es drei Fälle, wie die beiden Strecken zueinander liegen können (siehe Bild 4.5):

[7] $\delta_{-1} = \delta_{N-1} = 0$

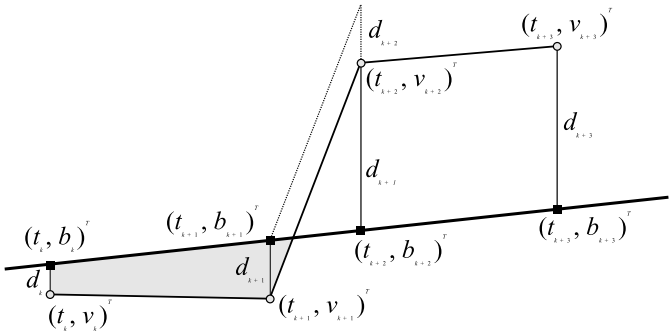


Abbildung 4.5: Drei Fälle der Lage der Strecken des Datenpolygons zur Verbindungslinie der Fußpunkte und die Berechnung der negativen Flächenanteile bei Schnitt des Datenpolygons mit der Verbindungslinie der Fußpunkte

- Beide Datenpunkte liegen unterhalb oder auf b. Dann ist $d_j < 0$ und $d_{j+1} < 0$. Der negative Anteil des Teilstückes ist das Trapez selber und damit

$$F_j^{\text{neg}} = -\frac{1}{2}(t_{j+1} - t_j)(d_{j+1} + d_j). \tag{4.7}$$

- Die Datenpunkte liegen auf verschiedenen Seiten von b. Dann ist entweder $d_j < 0$ und $d_{j+1} > 0$ oder $d_j > 0$ und $d_{j+1} < 0$. Sei nun

$$f = \frac{|d_j|}{|d_j| + |d_{j+1}|}$$

und

$$F = \frac{1}{2}(|d_j| + |d_{j+1}|)(t_{j+1} - t_j).$$

Dann erhält man unter Betrachtung ähnlicher Dreiecke

$$F_j^{\text{neg}} = \begin{cases} f^2 F, & \text{falls } d_j < 0 \text{ und } d_{j+1} > 0 \\ (1-f)^2 F & \text{sonst.} \end{cases} \tag{4.8}$$

- Beide Datenpunkte liegen oberhalb oder auf b. Dann ist $d_j > 0$ und $d_{j+1} > 0$. Das Teilstück hat keinen negativen Anteil:

$$F_j^{\text{neg}} = 0. \tag{4.9}$$

Der gesamte Anteil der negativen Fläche am Peak berechnet sich dann aus der Summe der Anteile der Teilstücken. Damit die Größe genormt ist, wird sie durch die Fläche F des Peaks dividiert. Damit erhält man schließlich

$$x^2 = F_p^2(i) = \frac{\sum_{j=l}^{r-1} F_j^{\text{neg}}}{F}. \tag{4.10}$$

Berechnung, ob sich beide Fußpunkte auf gleichem Niveau befinden

Es muß unterschieden werden, ob es sich um den linken oder rechten Fußpunkt handelt. Wenn h die Höhe des Peaks bezeichnet, dann ist

$$x^3 = F_p^3(i) = \begin{cases} \max(0, 1 - \frac{|v_i - v_r|}{h}), & \text{falls } i = l, \\ \max(0, 1 - \frac{|v_i - v_l|}{h}), & \text{falls } i = r. \end{cases} \tag{4.11}$$

Das Festsetzen der unteren Schranke auf 0 ist deswegen wichtig, weil bei einem Peak an einer steil ansteigenden Basislinie der vertikale Abstand der Fußpunkte durchaus größer sein kann als die Höhe.

4.5.2 Berechnung der Bewertung

Nachdem die Werte der Kriterien berechnet worden sind, müssen diese bewertet werden.

Die Wichtungen der Kriterien

Da die Gewichte der Kriterien gelernt werden sollen, müssen sie ähnlich wie die Parameter der Bewertungsfunktionen für einen Peak behandelt werden. Deswegen gibt es für jedes Kriterium – und damit für jedes Gewicht – eine Lernfunktion. Diese Lernfunktionen heißen $w^0(t)$, $w^1(t)$, $w^2(t)$ und $w^3(t)$.

Berechnen der Bewertung

Um die Bewertung für einen Fußpunkt $(t_i, v_i)^T$ eines Peaks p zu berechnen, werden zunächst, wie im letzten Abschnitt beschrieben, die Werte der Kriterien x^0, x^1, …, x^k mit Hilfe der Funktionen $F_p^0(i)$, $F_p^1(i)$, …, $F_p^k(i)$ berechnet[8]. Anschließend werden die Gewichte für den Zeitpunkt t_i mit Hilfe der Lernfunktionen für die Gewichte ermittelt. Diese Gewichte werden so normiert, daß ihre Summe 1 beträgt.

Die einzelnen Bewertungen der Kriterien werden mit den dazugehörenden Gewichten multipliziert und anschließend summiert. Damit ist die Bewertung s_i der Funktionen des Fußpunktes $(t_i, v_i)^T$

$$s_i = \frac{\mathbf{x}^T \mathbf{w}}{\mathbf{1}^T \mathbf{w}}, \tag{4.12}$$

[8] Bisher werden die oben genannten 4 Kriterien benutzt, damit ist $k = 4$. Das Prinzip ist aber für jede Anzahl von Kriterien zu verwenden.

wobei

$$\mathbf{x} = \begin{pmatrix} x^0 \\ x^1 \\ \vdots \\ x^k \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w^0(t) \\ w^1(t) \\ \vdots \\ w^k(t) \end{pmatrix} \quad \text{und} \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

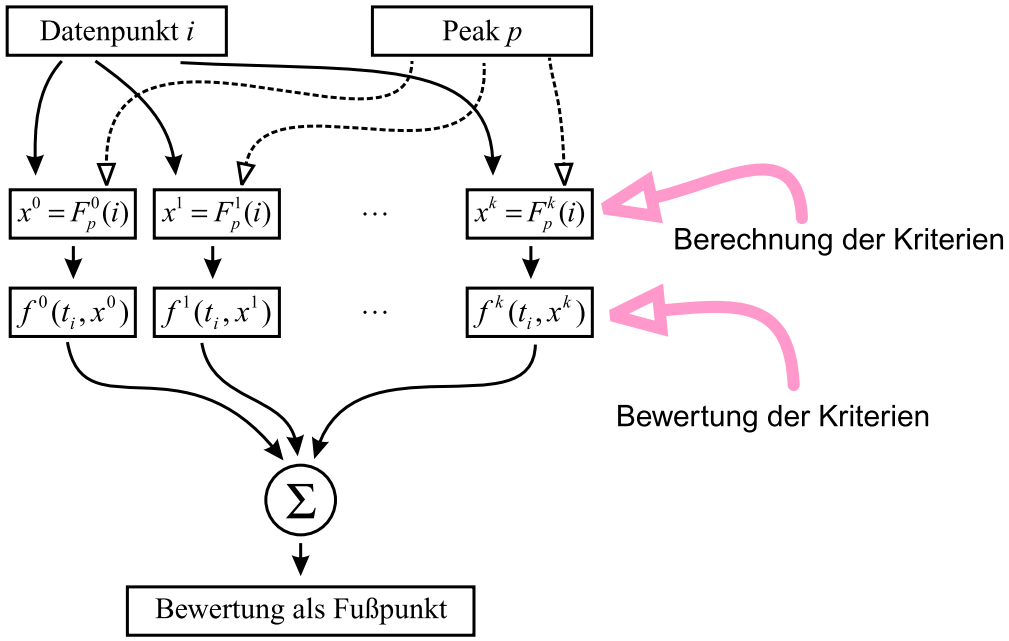


Abbildung 4.6: Bestimmung der Bewertung des Punktes i als Fußpunkt des Peaks p

In Bild 4.6 ist noch einmal die Vorgehensweise beim Bewerten des Fußpunktes $(t_i, v_i)^T$ dargestellt. Die Funktionen $F_p^j(t_i)$ berechnen das Kriterium j und die Funktionen $f^j(t_i, x^j) = x^j w^j(t_i)$ wichten diese Bewertungen entsprechend der Lernfunktionen $w^j(t)$. Am Ende werden diese Bewertungen aufsummiert, woraus sich die Bewertung des Fußpunktes ergibt.

4.5.3 Suche nach der höchsten Bewertung

Um nun denjenigen Fußpunkt mit der höchsten Bewertung für einen Peak p zu finden, muß man zunächst alle in Frage kommenden Datenpunkte ermitteln. Wenn man davon ausgeht, daß sich der Datenpunkt mit der besten Bewertung für den linken Fußpunkt im dem von der linken Peakhälfte abgedeckten Bereich befindet, kann man für jeden Datenpunkt $(t_i, v_i)^T$ folgendermaßen vorgehen:

1. Lege den linken Fußpunkt auf den Datenpunkt $(t_i, v_i)^T$.

2. Berechne die Bewertung s_i für den veränderten Peak.

3. Der Datenpunkt mit dem Index i, für den s_i das Maximum annimmt, ist der neue linke Fußpunkt des Peaks p.

Analog verfährt man, um den besten rechten Fußpunkt zu finden. Es stellt sich die Frage, welcher Fußpunkt bei der Fußpunktkorrektur zuerst bewegt werden sollte. Als geeignet hat sich dabei gezeigt, daß man den Fußpunkt mit dem größten vertikalen Spielraum in einem begrenzten Intervall (zum Beispiel innerhalb von 5 Datenpunkten nach links und nach rechts vom jeweiligen Fußpunkt) nehmen sollte.

4.5.4 Ein Lernschritt bei der Änderung eines Fußpunktes

Wenn der Nutzer mit einem Fußpunkt nicht einverstanden ist, soll er ihn versetzen können. Das kann beispielsweise geschehen, indem er ihn mit der Maus markiert und dann verschiebt. Angenommen, der Nutzer verschiebt den Fußpunkt $(t_i, v_i)^T$ des Peaks p nach $(t_{i'}, v_{i'})^T$.

Um diese Änderung für das Lernverfahren zu registrieren, kann man folgenden Lernschritt ausführen:

1. Berechne die Bewertungen $x^0, x^1, \ldots, x^k$ für den Fußpunkt $(t_i, v_i)^T$ und die Bewertungen $y^0, y^1, \ldots, y^k$ für den neuen Fußpunkt $(t_{i'}, v_{i'})^T$.

2. Wenn $y^l - x^l > 0$, hebe alle Gewichte w_j^l der Lernfunktion $w^l(t)$, für die $c \cdot j \in [t_i, t_{i'}]$, um den Wert 1 an. Ist $y^l - x^l < 0$, senke diese Gewichte um 1 ab. Verfahre so mit allen $l \in \{0, 1, \ldots, k-1\}$.

Es kann durchaus passieren, daß die Veränderung der Gewichte nach einer einmaligen Verschiebung des Fußpunktes noch nicht die gewünschten Resultate liefert. Allerdings werden nach mehrmaliger Anwendung dieses Schrittes die Gewichte für die Kriterien, die bei dem neuen Fußpunkt besser bewertet werden als bei dem alten Fußpunkt, so stark angehoben worden sein, daß die anderen Kriterien keine Rolle mehr spielen und der Fußpunkt dann an die vom Nutzer gewünschte Stelle gesetzt wird.

Es kann nur ein solcher Punkt als neuer Fußpunkt gelernt werden, der in wenigstens einem Kriterium besser ist als der der alte Fußpunkt. Unter der Annahme, daß zur Bewertung der Fußpunkte geeignete Kriterien herangezogen wurden, schränkt dies aber die praktische Anwendbarkeit des Verfahrens nicht ein.

4.6 Anwendung eines Clusterverfahrens

Es soll abschließend noch ein anderer Zugang zu einem lernfähigen Auswerteverfahren vorgestellt werden. Ähnlich wie bei der oben beschriebenen Vorgehensweise wird davon ausgegangen, daß zu Beginn der Auswertung eines Chromatogrammes zuerst eine Liste mit Peakkandidaten ermittelt wird. Den Peaks in dieser Liste wird dann jeweils ein Bewertungsvektor zugeordnet. Dieser könnte etwa als Komponenten die Höhe und das Verhältnis von Fläche zu Höhenquadrat haben.

Diese Bewertungsvektoren werden dann als Punkte in einem euklidischen Raum gedeutet. Die Peakkandidaten werden durch einen Experten klassifiziert, und diese Klassifizierung wird dann auf die entsprechenden Punkte übertragen.

Die zugrundeliegende Idee ist, daß Punkte, die ein und der selben Klasse angehören, auch räumlich nahe beieinander liegen. Wenn man die Punktmengen vieler ähnlicher Chromatogramme gemeinsam betrachtet, würde es sich anbieten, diese Punktwolke durch ein geeignetes Verfahren in Cluster zu zerlegen und dann die Cluster anstelle der einzelnen Punkte zu klassifizieren.

Die Auswertung eines Chromatogrammes könnte dann so vorgenommen werden, daß die den Peakkandidaten entsprechenden Punkte dieses Chromatogrammes jeweils dem nächstgelegenen Cluster zugeordnet werden und damit klassifiziert werden.

Der Erfolg des Verfahrens hängt wesentlich von der Wahl der in den Bewertungvektor eingehenden Charakteristika ab. Die mit dieser Auswahl verbundenen Schwierigkeiten führten dann auch zu der Entscheidung, das Verfahren nicht zur Anwendung zu bringen.

4.6.1 Das Verfahren

In [GDI93] ist ein Verfahren angegeben, das eine Menge von Punkte in eine vorgegebene Anzahl von Clustern aufteilt. Dieses Verfahren wird hier benutzt.

Gegeben seien eine Menge X von m Punkten $x^0, x^1, \ldots, x^{m-1} \in \mathbb{R}^n$. Ferner seien k Cluster gegeben, die durch ihre Zentroide $c_0, c_1, \ldots, c_{k-1} \in \mathbb{R}^n$ definiert sind. Dabei soll $1 \leq k \ll m$ sein.

Dabei ist das m die Anzahl der Kriterien, die zur Einschätzung des Peaks benutzt werden. Die Komponenten des Vektors x^j enthalten die Bewertung der Kriterien des Peaks j.

Die Zentroide werden zunächst initialisiert. Entweder werden sie mit zufälligen Werten belegt oder sie werden mit Elementen aus X initialisiert. Anschließend wird jeder Punkt einem Cluster zugeordnet. Üblicherweise wird dazu ein Abstandsmaß (euklidischer Abstand, Manhatten-Abstand) benutzt: Der Punkt wird demjenigen Cluster zugeordnet, zu dem er den kleinsten Abstand hat. Existieren deren mehrere, kann man sich ein Cluster aussuchen.

Danach wird aus allen Punkten, die zu einem Cluster gehören, der Schwerpunkt gebildet. Das ist dann der neue Zentroid. Diese Berechnung wird für alle Cluster durchgeführt. Nachdem das geschehen ist, werden die Punkte durch den Abstand erneut auf die Cluster verteilt. Dieses wiederholt sich so lange, bis das Abbruchkriterium

$$\max_i |c_i^{\text{alt}} - c_i^{\text{neu}}| < \epsilon$$

erfüllt ist. Dabei sind c_i^{alt} und c_i^{neu} der alte und der neue i-te Clustermittelpunkt und ϵ eine vorgegebene Schranke.

Die Vorteile des Verfahrens sind:

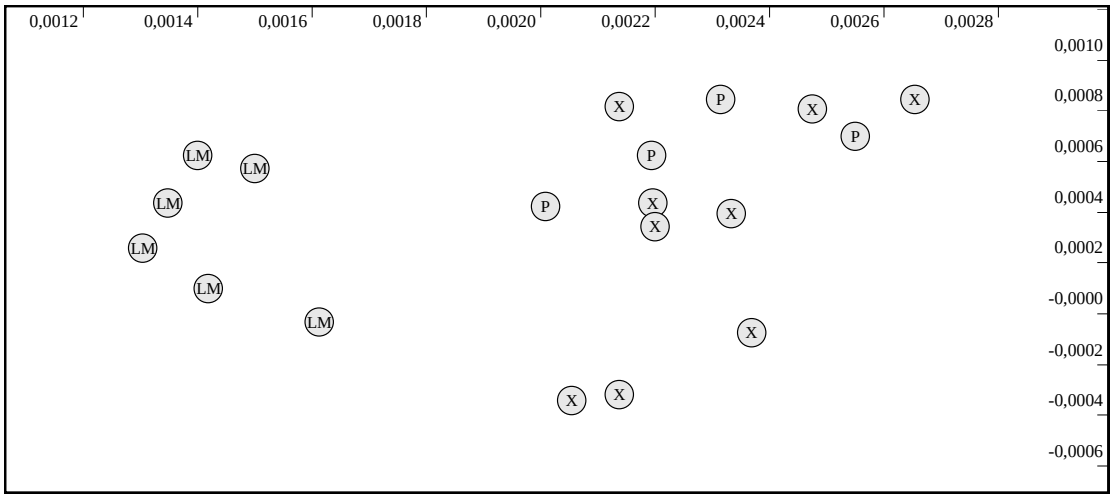
- Es ist einfach zu implementieren. Es ist leicht verständlich und hinreichend oft getestet.
- Es findet die Cluster meistens schon bei einer kleinen Anzahl von präsentierten Datenpunkten.
- Es ist egal, in welcher Reihenfolge die Datenpunkte dem Algorithmus präsentiert werden.

Dem stehen aber nicht zu unterschätzende Nachteile gegenüber:

- Die Anzahl der zu bildenden Cluster muß von vornherein bekannt sein. Das ist insofern schwerwiegend, als daß man sich zum einen Gadanken machen muß, wieviele Arten von Peaks man haben möchte.
- Der Erfolg des Verfahrens kann davon abhängen, wie die Zentroide zu Beginn initialisiert werden.
- Hat man es mit vollkommen unterschiedlichen Größen zu tun, ist es schwierig, den Abstand zweier Punkte voneinander zu interpretieren. Insbesondere kann es vorkommen, daß manche Dimensionen kaum Einfluß auf den Abstand haben. Wenn beispielsweise die Werte der ersten der Punkte zwischen 0 und 1 liegen und die Werte der zweiten Komponente zwischen 0 und 10^6, spielt die erste Komponente nur eine untergeordnete Rolle. Diesen Effekten kann man bedingt entgegenwirken, wenn man komponentenweise die Standardabweichung der Datenpunkte berechnet und damit die Komponenten normiert.

4.6.2 Erfahrungen mit dem Verfahren

Man sieht schnell ein, daß es keinen Sinn macht, zum Beispiel die Höhe eines Peaks als eine Dimension des Raumes zu benutzen. Denn ein Peak kann nahezu jede Höhe annehmen, so daß es keine vernünftige Aufteilung in Cluster geben kann.



P Einzelstehender Peak, **X** Fehlerhafter Peak, **LM** Mittlerer Peak eine Peakgruppe

Abbildung 4.7: Das Verhältnis von Unsymmetrie zum Quadrat der Peakhöhe aufgetragen über das Verhältnis von Peakfläche zum Quadrat der Peakhöhe von Peaks aus einer manuell ausgewerteten Serie von Chromatogrammen.

Deswegen muß man Parameter benutzen, die unabhängig von der Größe des Peaks sind. Man könnte etwa einen zweidimensionalen Raum aufspannen, dessen Dimensionen das Verhältnis von Peakfläche zum Quadrat der Höhe und das Verhältnis von Unsymmetrie zum Quadrat der Höhe sind. Dann erhält man für eine per Hand ausgewertet Serie beispielsweise Bild 4.7.

Man sieht an diesem Bild, daß sich schon bei einer manuell ausgewerteten Serie bezüglich der Clusterbildung Schwierigkeiten einstellen. Zwar sind die Peaks des Typs LM klar von den anderen separiert auszumachen, allerdings liegen Peaks der Typen X und P durcheinander und völlig unsepariert voneinander in der rechten Bildhälfte herum.

Das ist ein Beispiel dafür, daß es – jedenfalls mit den genannten Peakparametern – nicht möglich ist, eine per Hand ausgewertete Serie in Bereiche einzuteilen.

4.6.3 Einschätzung und Zusammenfassung des Verfahrens

Aufgrund des eben gezeigten Beispiels, welches nur eines von vielen ähnlichen ist, wurde das Verfahren zur Peakevaluierung verworfen. Weitere gravierende Nachteile sind, daß man keine absoluten Werte, wie Höhe oder Fläche, als Komponenten benutzen darf und daß bekannt sein muß, wieviele Cluster man braucht.

5 Prototypische Implementierung mit Amira

Die im letzten Kapitel beschriebenen Verfahren wurden in der prototypischen Software *Amira* implementiert. Amira umfaßt zum einen die Algorithmen zur Auswertung als auch eine Schnittstelle zur interaktiven Auswertung.

In diesem Kapitel wird die Struktur von Amira erläutert und anschließend wird auf die Umsetzung der Algorithmen eingegangen. Die Schnittstelle zur Interaktion zwischen Nutzer und Software ist nicht Gegenstand dieser Arbeit und wird hier nicht weiter beschrieben.

5.1 Verwaltung

In diesem Abschnitt wird beschrieben, wie die Daten in Amira verwaltet werden. Dabei wird zunächst die grobe Struktur beschrieben und anschließend die Details.

Wenn im folgenden von *Vektoren* die Rede ist, ist eine spezielle Form von Feldern gemeint. Da die Software in C++ geschrieben ist, wurde reger Gebrauch von der *Standard C++ Template Library* (STL) gemacht. Diese Bibliothek stellt eine Datenstruktur `vector` zur Verfügung, mit der man Felder dynamisch verwalten kann. Im Gegensatz dazu sind die Felder, die C++ zur Verfügung stellt, nicht dynamisch änderbar. Um diesem Unterschied Rechnung zu tragen, wird der Begriff *Vektor* verwendet. Einen sehr guten Überblick über die Strukturen der STL und deren Anwendung bietet [ECK99].

5.1.1 Grobstruktur

Um eine Serie von Chromatogrammen auswerten zu können, werden folgende Daten benötigt:

- Eine Liste von Rohdatensätzen
- Eine Liste von Peaklisten
- Eine Lernfunktion[1]

[1] Im letzten Kapitel wurde erklärt, das man mehrere Lernfunktionen benötigt. Die Erklärung, warum hier nur eine solche Funktion aufgelistet ist, folgt später.

Sowohl die Peaklisten als auch die Rohdatensätze sind Elemente von Vektoren. Diese drei genannten Strukturen sind im Bild 5.1 dargestellt: Das Auswerteverfahren nimmt aus der Liste der Rohdatensätze ein vom Nutzer ausgesuchtes Element und wertet die Rohdaten dieses Elementes aus. Die Peaks, die das Ergebnis dieses Vorganges sind, werden in der entsprechenden Peakliste abgespeichert. Zur Auswertung wird die Lernfunktion benutzt.

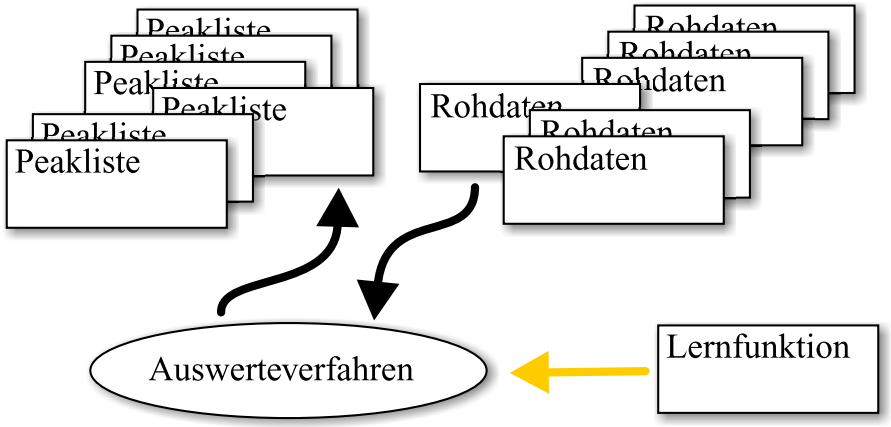


Abbildung 5.1: Zusammenspiel zwischen Auswerteverfahren, Peaklisten, Chromatogrammen und Lernfunktion

Die Zuordnung der Peakliste zum entsprechenden Rohdatensatz erfolgt über einen Schlüssel, der zusammen mit der Peakliste abgespeichert wird und auf den entsprechenden Rohdatensatz verweist. Damit steht die Möglichkeit offen, einem Rohdatensatz mehrere Peaklisten zuzuordnen. Das macht dann Sinn, wenn man einen Datensatz mit verschiedenen Parametern (Rauschwert, Lernfunktion, ...) auswerten und einen direkten Vergleich der Ergebnisse haben möchte. In Amira wurde diese Möglichkeit der Mehrfachzuordnung jedoch nicht genutzt.

5.1.2 Struktur eines Rohdatensatzes

Ein Rohdatensatz ist ein Vektor mit Elementen des Typs `FL_POINT`. Der Typ `FL_POINT` repräsentiert einen Datenpunkt und enthält neben den Koordinaten auch die Ableitung des Rohdatensatzes an der Stelle dieses Datenpunktes sowie den Wert der geglätteten Daten an dieser Stelle.

Der Rohdatensatz enthält außerdem einen Schlüssel, der beim Anlegen des Datensatzes generiert wird. Der Schlüssel ist für alle jemals geladenen Sätze eindeutig.

5.1.3 Struktur einer Peakliste

Eine Peakliste ist ein Vektor[2] mit Elementen des Types `PEAK`. Ein `PEAK` enthält Felder, in denen folgende Werte abgespeichert sind:

- Koordinaten (t^l, v^l) und (t^r, v^r) der beiden Fußpunkte sowie deren Indizes l und r der entsprechenden Datenpunkte im zugehörigen Chromatogramm,
- Koordinaten (t^m, v^m) des Retentionspunktes sowie der Index m des entsprechenden Datenpunktes im zugehörigen Chromatogramm,
- Die Indizes der Datenpunkte für die Fußpunkte des Peaks, so wie sie beim ersten Detektieren der Peaks aus dem Rohdatensatz gefunden werden,
- Fläche, Höhe,
- Typ des Peaks,
- Flag, ob der Peak auf einer rechten Flanke aufsitzt (nur gültig bei Schulterpeaks),
- Flag, durch das der Peak als unveränderbar markiert ist.

Die Redundanz, die durch das Speichern sowohl der Indizes der Datenpunkte für Retentionspunkt und Fußpunkte als auch der Punkte selber entsteht, ist deswegen von Vorteil, weil einige Verfahren die Zeiten der Fußpunkte und des Retentionspunktes benötigen, andere wiederum benötigen die Indizes der zugehörigen Datenpunkte. Man kann zwar von der Zeitangabe eines Punktes auf dessen Index schließen und umgekehrt, aber durch die Einführung der Redundanz spart man einen Schritt, der den Index aus der Zeit oder die Zeit aus dem Index ermittelt.

Der Typ des Peaks ist als Zeichenkette abgespeichert. Zur Zeit werden folgende Typen unterstützt:

Einzelstehende gültige Peaks Der Typ ist P.

Schulterpeaks Der Typ ist S.

Peaks, die zu einer Peakgruppe gehören Der Typ ist L. Bei Peaks dieses Typs existieren außerdem Untergruppen, nämlich L, M und R für einen linken, inneren und rechten Peak der Peakgruppe. Für einen inneren Peak einer Gruppe ist der Typ also LM.

[2] Der Begriff *Peakliste* ist im Zusammenhang mit einem Vektor irreführend, weil eine Liste eine volkommen andere Datenstruktur ist als ein Vektor. Da sich dieser Begriff jedoch etabliert hat, wird er hier weiterhin benutzt.

Ungültige Peaks Der Typ ist X.

Sieht man einmal von Schulterpeaks ab erhält man für die Peaks p_0, p_1, ... einer Peakliste folgende Ungleichung:

$$p_0.t^l < p_0.t^m < p_0.t^r \leq p_1.t^l < p_1.t^m < p_1.t^r \leq p_2.t^l \ldots \tag{5.1}$$

Mit dem Einführen von Schulterpeaks wird diese Ordnung aber aufgehoben, da ein Schulterpeak immer einen anderen Peak überlappt. Aus diesem Grund werden Schulterpeaks am Ende der Peakliste abgespeichert, um wenigstens für einen Teil der Peakliste die obige Ordnung zu sichern.

Eine Peakliste enthält weiterhin eine Kopie des Schlüssels des zugehörigen Rohdatensatzes.

5.1.4 Die Lernfunktion

Eine Lernfunktion ist ein Vektor mit Elementen des Typs `learnPeak`. Sie enthält einen Parameter c. Dieser Parameter gibt an, wie groß die zeitliche Entfernung zweier aufeinanderfolgende Elemente des Vektors ist. Das Element mit dem Index j liegt dabei über dem Zeitpunkt $\frac{j}{c}$.

Da die Verteilung der Punkte auf der Zeitachse für jedes Kontrollpolygon identisch sein soll, wurden alle Kontrollpunkte für ein und denselben Zeitpunkt in einem `learnPeak` derart zusammengefaßt, daß man anstelle der Punkte $(t_j, w_j^0)^T$, $(t_j, w_j^1)^T$, ..., $(t_j, w_j^{m-1})^T$ für die Lernfunktionen $w^0(t)$, $w^1(t)$, ... $w^{m-1}(t)$ nun einen einzigen Punkt $(t_j, w_j^0, w_j^1, \ldots, w_j^{m-1})^T \in \mathbb{R}^{m+1}$ verwalten muß.

Da die Komponente eines Kontrollpunktes, welche die Zeit enthält, anhand des Indexes des Punktes errechnet werden kann, müssen nur noch Punkte der Form $(w_j^0, w_j^1, \ldots, w_j^{m-1})^T \in \mathbb{R}^m$ verwaltet werden.

Deswegen ist `learnPeak` seinerseits ein Vektor mit Elementen des Typs `learnPeakItem`. Der Typ `learnPeakItem` enthält folgende Elemente:

- Eine untere (`min`) und eine obere (`max`) Schranke für das entsprechende Kriterium. Das sind der kleinste Wert und der größte Wert, für den das Kriterium mit 1 bewertet wird. Diese Werte spielen bei den Lernfunktionen für Fußpunkte keine Rolle. Wenn für ein Kriterium, wie zum Beispiel die Peakhöhe, keine obere Grenze existiert, wird die obere Schranke nicht beachtet.

- Den Wert `value` für das Gewicht. Dieser Wert wird nur für Lernfunktionen, die den Verlauf eines Gewichtes beschreiben, benutzt.

- Ein Feld `dev`, das die Standardabweichung aller bisher von diesem Kriterium zu der dem zugehörigen `learnPeak` entsprechenden Zeit angenommenen Werte enthält.

Um die Punkte des Polygons, das die Lernfunktion eines Gewichtes beschreibt, zu bekommen, kann man direkt das Feld `value` der entsprechenden Komponente der Kontrollpunkte vom Typ `learnPeak` benutzen.

Um die Eckpunkte der Lernfunktionen zum Evaluieren der Peaks zu erhalten (siehe Bild 4.2 auf Seite 34), bedarf es noch eines geringen Aufwandes. Dies soll an dem folgenden Beispiel erklärt werden:

Es wird der Verlauf (also die Eckpukte) der Bewertungsfunktion für das Verhältnis von Peakfläche zum Quadrat der Peakhöhe zum Zeitpunkt t gesucht. Die Elemente vom Typ `learnPeak` seien mit L_0, L_1, ... bezeichnet. Das Minimum und das Maximum der Werte für das Verhältnis von Peakfläche zum Quadrat der Peakhöhe zum Zeitpunkt t sind in Komponente mit der Nummer 1 des `learnPeak` L_j mit dem Index $j = \lfloor \frac{t}{c} \rfloor$ abgespeichert. Damit erhält man als obere Eckpunkte des Trapezes der Bewertungsfunktion zum Zeitpunkt t die Punkte $(L_j^1.\texttt{min}, 1)^T$ und $(L_j^1.\texttt{max}, 1)^T$. Dann sind die unteren Punkte des Trapezes die folgenden: $(L_j^1.\texttt{min} - L_j^1.\texttt{dev}, 0)^T$ und $(L_j^1.\texttt{max} + L_j^1.\texttt{dev}, 0)^T$

5.2 Beschreibung der Algorithmen zur Generierung einer Peakliste

In diesem Abschnitt wird beschrieben, wie aus den Rohdaten die Peakliste unter Einbeziehung der Lernfunktion generiert wird. Dabei werden mehrere voneinander unabhängige Module betrachtet:

Das Modul zum Finden der Peakkandidaten Dieses Modul findet zunächst ansteigende und abfallende Flanken in einem Rohdatensatz. Anschließend werden aus den Flanken die Peakkandidaten generiert.

Das Modul zur Trennung durch Lotfällung Dieses Modul überprüft, ob Peakkandidaten eng genug zusammenliegen und wenn ja, ob dann der Talpunkt hoch genug liegt, um eine Lotfällung durchzuführen. Gegebenenfalls werden Peaks durch Lotfällung voneinander getrennt.

Modul zur Integration Dieses Modul berechnet die Flächen der Peakkandidaten.

Modul zur Evaluierung der Kandidaten Dieses Modul überprüft jeden Peak hinsichtlich seiner Höhe und seines Verhältnisses von Fläche zum Höhenqudrat.

Modul zur Abtrennung von Schulterpeaks Dieses Modul überprüft Peakgruppen, ob sie Schulterpeaks enthalten und trennt diese gegebenenfalls ab.

Modul zur Korrektur der Fußpunkte Dieses Modul sucht die „besten“ Fußpunkte für die Peaks.

Modul zum Sortieren der Peaks Dieses Module ordnet die Peaks in der Liste so an, daß sich Schulterpeaks am Ende befinden. Ansonsten werden die Peaks nach der Retentionszeit sortiert. Neben dieser Sortierung leistet dieses Modul außerdem:

- Es wird für jeden Peakkandidaten der Retentionspunkt bestimmt.
- Die Koordinaten der Fußpunkte werden aus den Indizes der Fußpunkte berechnet. Dabei spielt es eine Rolle, ob der Peak einer Gruppe angehört.

Dieses Modul muß immer dann aufgerufen werden, wenn an den Fußpunkten oder am Typ eines Peaks etwas verändert worden ist.

Wie diese Module geschaltet sind, geht aus Bild 5.2 hervor.

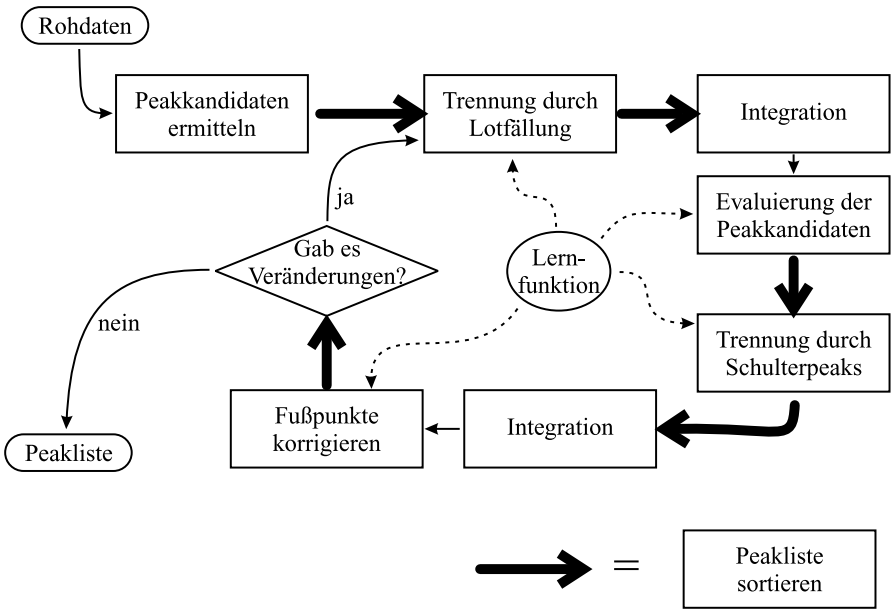


Abbildung 5.2: Ablauf der Auswertung eines Chromatogrammes

Zur nachfolgenden Beschreibung der einzelnen Module seien folgende Vereinbarungen getroffen:

- Mit $\mathfrak{C}$ sei ein Rohdatensatz mit N Datenunkten $(t_i, v_i)_t^T 0 \leq i < N$ bezeichnet. s_i bezeichnet den geglätteten Wert an der Stelle t_i. $\mathfrak{C}$.`noise` bezeichnet den Rauschwert des Chromatogrammes $\mathfrak{C}$.

- Ein Peak p habe die Fußpunkte $(p.t^l, p.v^l)$ und $(p.t^r, p.v^r)$ sowie den Retentionspunkt $(p.t^m, p.v^m)$. Der Typ des Peaks sei mit $p.\texttt{typ}$ bezeichnet. Die Indizes dieser Punkte seien entsprechend mit $p.l$, $p.m$ und $p.r$ bezeichnet. $p.f$ ist die Fläche des Peaks und $p.h$ dessen Höhe. $p.l^*$ und $p.r^*$ bezeichnen die Indizes der Fußpunkte, wie sie beim ersten Detektieren des Peaks ermittelt worden sind.

- Ist $(x_i)_{i\in\mathbb{N}}$ eine Folge reeller Zahlen, dann ist $\Delta x_i = x_{i+1} - x_i$. Dieser Wert wird als *Vorwärtsdifferenz* bezeichnet.

- Die Lernfunktion wird mit L bezeichnet. Die Kontrollpunkte der Lernfunktion seien mit L_0, L_1, ... bezeichnet. Die verschiedenen Komponenten der Kontrollpunkte seien L_0^k, L_1^k, ... Weiterhin ist $L^k(t)$ diejenige Funktion, die durch die k-te Komponente beschrieben wird.

  $L_j^2.\texttt{min}$ bedeutet dann zum Beispiel denjenigen Wert, für den das Kriterium mit der Nummer 2 zum Zeitpunkt cj gerade noch mit 1 bewertet wird. $L^2(t)$ bedeutet denjenigen Wert, für das das Kriterium mit der Nummer 2 zum Zeitpunkt t gerade noch mit 1 bewertet wird. Wenn $\frac{t}{c}$ nicht ganzzahlig ist, wird der Funktionwert mit Hilfe der zu t benachbarten Punkte interpoliert.

5.2.1 Das Modul zum Finden der Peakkandidaten

Um die Peakkandidaten zu finden, eignen sich prinzipiell die unter Abschnitt 2.6.1 auf Seite 14 vorgestellten Verfahren zur direkten Peaksuche. Dabei hatte sich unter Abschnitt 3.3 gezeigt, daß das rekursive Verfahren Schwächen hat. Aus diesem Grund ist in Amira folgendes Vorgehen implementiert:

1. Bilde eine Folge $(d_i)_{0\le i<N-1}$ mit

$$d_i = \begin{cases} -1, & \text{falls } \Delta s_i < -\mathfrak{C}.\texttt{noise}, \\ 1, & \text{falls } \Delta s_i > \mathfrak{C}.\texttt{noise}, \\ 0 & \text{sonst.} \end{cases} \tag{5.2}$$

2. Erzeuge die Folge der Flanken nach Algorithmus 1. Dabei sei eine Flanke f durch ihren Startindex $f.s$, ihre Länge $f.l$ und ihren Typ $f.t$ gekennzeichtet. Es gibt zwei Typen von Flanken, nämlich 1 für eine ansteigende und -1 für eine abfallende.

3. Erzeuge die Peakkandidaten nach Algorithmus 2. Dieser Algorithmus sucht in der Liste der Flanken die erste ansteigende Flanke. Dazu wird eine abfallende Flanke gesucht. Wenn zwei solche Flanken gefunden worden sind, markieren der Anfang der ansteigenden und das Ende der abfallenden Flanke die Fußpunkte.

Algorithmus 1 Erzeugen der Flanken

Voraussetzung: Folge $(d_i)_{0 \leq i < N-1}$ mit den Vorzeichen der Vorwärtsdifferenzen

```
i ← 0, m ← 0
while i < N − 1 do
  if d_i ≠ 0 then
    k = i + 1
    while k < N − 1 and d_k ≠ d_i do
      k ← k + 1
    end while
    if k − i ≥ 3 then
      f_m.s ← i, f_m.l ← k − i, f_m.t ← d_i
      m ← m + 1
    end if
    i ← i + k
  else
    i ← i + 1
  end if
end while
return (f_j)_{0≤j<m}
```

Wenn man sich Algorithmus 2 betrachtet, stellt man fest, daß die Flanken sehr breit sind: Wenn mehrere ansteigende Flanken aufeinander folgen, wird immer das linke Ende der linken Flanke als linker Fußpunkt eines Peakkandidaten betrachtet. Gleiches gilt für die rechte Flanke. Das kann dazu führen, daß die Fußpunkte viel zu weit von der Retentionszeit entfernt liegen. Als folge dessen treten meist deformierte Peaks mit einem hohen negativen Flächenanteil auf. Das ist eine große Schwäche des Verfahrens gegenüber dem rekursiven Ansatz. Aus diesem Grund müssen die Fußpunkte der Peakkandidaten nachkorrigiert werden.

Im folgenden werden die Begriffe „Peak“ und „Peakkandidat“ gleichgesetzt und der Begriff „Peak“ für beide verwendet, weil beide Typen von den meisten Modulen gleich behandelt werden.

5.2.2 Das Modul zur Trennung durch Lotfällung

Für die Überprüfung, ob zwischen zwei Peaks ein Lot gefällt werden soll, wird das Verfahren aus Abschnitt 3.2.1 modifiziert. Ein Peak p wird immer mit seinem linken Nachbarn a überprüft. Liegen a und p dicht beeinander, also weniger als 5 Datenpunkte voneiander entfernt, und sind sie nicht schon durch Lotfällung voneinander getrennt, wird die Höhe des Talpunktes $(t_v, v_v)^T$ über der Verbindungslinie der äußeren Fußpunkte der beteiligten Peaks berechnet. Anschließend

Algorithmus 2 Erzeugen der Liste der Peakkandidaten

Voraussetzung: Folge $(f_j)_{0 \leq j < m}$ mit der Folge der Flanken

```
j ← 0, k ← 0
while j < m do
  if f.t = 1 then
    i ← j + 1
    while f_i.t ≠ −1 and i < m do
      i ← i + 1
    end while
    if i = m then
      break
    end if
    while f_i.t = −1 and i < m do
      i ← i + 1
    end while
    i ← i − 1
    p_k.l ← f_j.s, p_k.r ← f_i.s + f_i + r
    p.typ = P
    k ← k + 1
    j ← i + 1
  else
    j ← j + 1
  end if
end while
return (p_j)_{0≤j<k}
```

wird dieser Wert durch den Durchschnitt der Höhen der beteiligten Peaks dividiert. Dieser Wert sei h_v. Anschließend wird dieser Wert mit $L^2(t_v)$ verglichen. Liegt er darüber, wird das Lot gefällt.

Ob die zwei Peaks a und p bereits voneinander getrennt sind, erkennt man daran, daß entweder p vom Typ LM oder LR ist.

Wenn das Lot gefällt wird, wird der Typ des Peaks auf L gesetzt. Ist der alte Typ von p schon L gewesen, deutet das darauf hin, daß p einer anderen Gruppe angehört. In diesem Fall wird der Untertyp M gesetzt, ansonsten wird der Untertyp R gesetzt. Entsprechend wird a behandelt.

5.2.3 Das Modul zur Schulterpeakabtrennung

Voraussetzung für eine Schulterpeakabtrennung ist, daß die zu untersuchenden Peaks vorher durch Lot voneinander getrennt worden sind. Zunächst wird getestet, ob die Peaks a und b mit $a.m < b.m$ durch Lot voneinander getrennt worden sind. Dieser Test ist der gleiche wie der im letzten Abschnitt genannte.

Anschließend werden die Flächen beider Peaks verglichen. Wenn $a.f < b.f$, wird $v = \frac{b.f}{a.f}$ berechnet, ansonsten $v = \frac{a.f}{b.f}$.

Angenommen, es gilt $a.f < b.f$. Ist nun $v > L^3(a.t_m)$, ist a ein Schulterpeak von b. Damit wird der linke Fußpunkt von b mit dem linken Fußpunkt von a gleichgesetzt. a erhält den Typ S. Der Typ von b wird P, wenn a vorher den Typ LL hatte, weil in diesem Fall die Peakgruppe aufgelöst worden ist. Beim anschließenden Sortieren wird a an das Ende der Liste befördert.

Wenn $a.f \geq b.f$, wird getestet, ob b ein Schulterpeak ist. Der weitere Ablauf ist dann analog.

5.2.4 Das Modul zum Sortieren der Peaks

Die erste Aufgabe dieses Moduls ist es, die Peaks nach folgender Ordnung zu sortieren. Wenn a und b zwei Peaks sind, dann werden der Reihe nach folgende Punkte ausgewertet:

1. Wenn $a.\texttt{typ} \neq S$ und $b.\texttt{typ} = S$, dann ist $a < b$.
2. Wenn $a.\texttt{typ} = S$ und $b.\texttt{typ} \neq S$, dann ist $a \not< b$.
3. $a < b$ genau dann, wenn $a.m < b.m$.

Nachdem die Peaks sortiert worden sind, wird überprüft, ob alle Peakgruppen zusammenhängend sind. Betrachtet man die Peaks, die keine Schulterpeaks sind, der Retentionszeit nach sortiert, muß für einen Peak p folgendes gelten:

1. Ist p vom Typ LM dann sind der linke und rechte Nachbarpeak beide vom Typ L.
2. Ist p vom Typ LL rechte Nachbarpeak vom Typ L.
3. Ist p vom Typ LR linke Nachbarpeak vom Typ L.

Das Überprüfen der Peakliste auf Einhaltung dieser Bedingungen und gegebenenfalls eine Korrektur ist dann wichtig, wenn zum Beispiel eine Peakgruppe aufgespalten werden soll oder zwei Peakgruppen verschmolzen werden sollen.

Nachdem sichergestellt worden ist, daß die Peakliste bezüglich der Gruppen konsistent ist, wird der Retentionspunkt für einen Peak gesucht. Dabei ist

$$p.m = k \quad \text{mit } v_k = \max\{v_{p.l+1}, v_{p.l+2}, \ldots, v_{p.r-1}\} \tag{5.3}$$

Anschließend erhält man $(p.t^m, p.v^m) = (t_{p.m}, v_{p.m})$. werden die Fußpunkte anhand der Indizes $p.l$ und $p.r$ berechnet. Handelt es sich bei einem Peak p um einen einzelstehenden Peak oder um einen Schulterpeak, dann ist $(p.t^l, p.v^l) = (t_{p.l}, v_{p.l})$ und $(p.t^r, p.v^r) = (t_{p.r}, v_{p.r})$. Bei einem Peak einer Peakgruppe werden zunächst die die Gruppe begrenzenden Fußpunkte $(t_l, v_l)^T$ und $(t_r, v_r)^T$ ermittelt. Damit werden dann die Fußpunkte von p folgendermaßen berechnet:

$$\begin{pmatrix} p.t^l \\ p.v^l \end{pmatrix} = \begin{pmatrix} t_{p.l} \\ v_l + (t_{p.l} - t_l)\frac{v_r - v_l}{t_r - t_l} \end{pmatrix} \tag{5.4}$$

und

$$\begin{pmatrix} p.t^r \\ p.v^r \end{pmatrix} = \begin{pmatrix} t_{p.r} \\ v_l + (t_{p.r} - t_l)\frac{v_r - v_l}{t_r - t_l} \end{pmatrix}. \tag{5.5}$$

5.2.5 Das Modul zum Evaluieren der Peaks

Dieses Modul überprüft jeden Peak p der Liste, ob er gültig ist. Dabei werden zunächst die Eckpunkte

$$\begin{pmatrix} h^1 \\ 1 \end{pmatrix} = \begin{pmatrix} L^0(p.t_m).\texttt{min} \\ 1 \end{pmatrix} \tag{5.6}$$

und

$$\begin{pmatrix} h^0 \\ 0 \end{pmatrix} = \begin{pmatrix} L^0(p.t_m).\texttt{min} - L^0(p.t_m).\texttt{dev} \\ 0 \end{pmatrix} \tag{5.7}$$

der Rampe zum Evaluieren der Höhe ermittelt. Damit wird der Wert

$$\mu_{h_{\min}} = \mu^h\left(p.t^m, \frac{p.h}{\mathfrak{C}.\texttt{noise}}\right) \tag{5.8}$$

nach Gleichung (4.2) berechnet. Analog wird die Bewertung μ_{form} für das Verhältnis von Peakfläche zu Höhenquadrat mit Hilfe der Funktion $L^1(t)$ berechnet. Damit wird mit Hilfe von Gleichung (4.1) die Gesamtbewertung μ_p für den Peak berechnet:

$$\mu_p = \mu_{\text{and}}(\mu_{h_{\min}}, \mu_{\text{form}}, 0.1) \tag{5.9}$$

Wenn $\mu_p < 0.5$, dann wird der Typ von p auf X gesetzt als Zeichen dafür, daß er ungültig ist. Sollten dadurch Peakgruppen aufgespalten werden, wird durch das Modul, das für das Sortieren zuständig ist, entsprechende Korrekturen vorgenommen.

5.2.6 Das Modul zum Korrigieren der Fußpunkte

Dieses Modul arbeitet genauso wie unter Abschnit 4.5 auf Seite 40 ff. beschrieben. Die entsprechenden Gewichte für einen Fußpunkt $(t_i, v_i)^T$ erhält sind $L^k(t)$.`value` für $k = 4, 5, 6, 7$.

Der Bereiche, in dem die Fußpunkt gesucht werden, werden für den linken Fußpunkt durch $p.l^*$ und $p.m-1$ begrenzt. Für den rechten Fußpunkt ist dieser Bereich analog durch $p.m + 1$ und $p.r^*$ festgelegt.

5.2.7 Das Modul zum Integrieren der Peaks

Das Integrationsmodul arbeitet nach der im Abschnitt 2.6.2 auf Seite 15 beschriebenen Methode mit einer Erweiterung: Wenn ein Peak p einer Lotfällungsgruppe angehört, liegen die Fußpunkte in der Regel nicht auf dem Polygonzug durch die Datenpunkte. Aus diesem Grund ändert sich die Flächenformel geringfügig:

$$p.f = \frac{1}{2}\left(-(p.v^l + p.v^r)(p.t^r - p.t^l) + \sum_{j=p.l}^{p.r-1}(v_{j+1} + v_j)(t_{j+1} - t_j)\right). \tag{5.10}$$

5.3 Das Lernen von Parametern

Das Lernen von Parametern hängt stark mit der Implementierung der Schnittstelle zwischen Software und Nutzer zusammen. In Amira wurde dabei eine rein grafische Schnittstelle implementiert: Der Nutzer kann ausschließlich durch Bedienen der Maus Peaks als gültig oder ungültig markieren und Fußpunkte verschieben. Nach einer Korrektur werden die Chromatogramme neu ausgewertet. Damit entspricht der Ablauf der Dialoges dem im Bild 4.1 auf Seite 30 dargestellten Prinzip. Dem Nutzer steht außerdem die Möglichkeit offen, Peaks als „fest“ zu markieren. In diesem Fall wird der Peak durch das Auswerteverfahren nicht mehr beeinflußt.

Das Lernen von Parametern als Reaktion auf Änderungen durch den Anwender wurde in Amira wie unter Abschnitt 4.4.2 beschrieben implementiert. Allerdings werden auf Wunsch des Vertragspartners die Kriterien zur Evaluierung der Peaks alle gleichzeitig gelernt: Markiert der Analytiker einen bis dahin als ungültig markierten Peak als gültig, werden sowohl die Knoten der Lernfunktion für die Höhe und die Knoten der Lernfunktion für das Verhältnis von Peakfläche zum Höhenquadrat verschoben.

Gravierender ist aber, daß beim umgekehrten Fall, nämlich daß der Nutzer einen als gültig markierten Peak für ungültig erklärt, ebenso verfahren wird. Damit kann nicht unterschieden werden, ob der Nutzer den Peak abgewählt hat, weil er zu breit, zu schmal oder zu niedrig ist.

Um den Bereich der Veränderungen innerhalb der Lernfunktionen festzulegen, stehen alle drei unter Abschnitt 4.4.2 genannten Möglichkeiten, den Lernradius festzulegen, zur Verfügung. Es hat sich als günstig erwiesen, beim Anlernen eines Kriteriums das letzte der drei Verfahren zu benutzen. Dabei werden die dort genannten Punkte t_l^0, t_l^1, t_r^1 und t_r^0 folgendermaßen festgelegt:

Lernen von Peakparametern des Peaks p $t_l^0 = p.t^l$, $t_l^1 = t_r^1 = p.t^m$ und $t_r^0 = p.t^r$

Lernen eines Trennverfahrens für die Peals a und b, wobei a und b benachbart sind und $a.m$ ¡ $b.m$. Dann werden die 4 Punkte folgendermaßen bestimmt: $t_l^0 = a.t^l$, $t_l^1 = a.t^m$, $t_r^1 = b.t^m$ und $t_r^0 = b.t^r$

Um eine Fußpunktverschiebung zu lernen, wird die im Abschnit 4.5.4 beschriebenen Methode benutzt. Dabei werden die Werte der entsprechenden Komponente des Kontrollpolygons der am Lernvorgang beteiligten Kontrollrpunkte um den festen Wert 1 in nach oben oder nach unten verschoben. Dadurch kann man davon ausgehen, daß sich Gewichte auch nach einer mehrmaligen Korrektur nicht übermäßig groß sind.

Insbesondere bei der Korrektur der Fußpunkte ist man sich noch nicht ganz klar darüber, ob nicht eine Verschiebung der Gewichte in Abhängigkeit von der Stärke der Änderung der Kriterien angebracht ist und wie man in einem solchen Fall die neuen Gewichte bestimmen kann. An dieser Stelle ist noch Forschungsarbeit zu leisten.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde gezeigt, wie man das Auswerten von Chromatogrammserien unterstützen kann, indem man Auswerteparamter zeitabhängig lernt. An dieser Stelle gibt es aber noch viel mehr Möglichkeiten, den Benutzer zu unterstützen. Zum Beispiel eignen sich die vorgestellten Lernverfahren nur bedingt für Serien, in denen die einzelnen Chromatogramme einen starken zeitlichen Versatz zeigen.

Bei vielen Chromatogrammserien gibt es darüber hinaus einen Versatz der Peaks, der eine Funktion der Zeit ist. Das liegt daran, daß sich die Umgebungsbedingungen (Luftdruck, Temperatur, ...) bei mehreren Aufzeichnungen ändern können; immerhin kann das Aufzeichnen eines einzigen Gaschromatogrammes mehrere Minuten dauern. Um das Auswerteverfahren weiter zu verbessern, könnte man anhand hinreichend stark ausgeprägter Peaks oder auch bestimmter wiederkehrender Fehler diesen Versatz korrigieren. Das würde auf die Anwendung von Mustererkennungsverfahren hinauslaufen. Insbesondere würde man durch eine solche Versatzkorrektur in der Lage sein, in den verschiedenen Chromatogrammen einer Serie die Peaks einander zuordnen zu können.

Weiterhin könnte man die Erkennung eines Peaks zu einem Zeitpunkt t verbessern, indem man Informationen über Komponenten, die Peaks zum Zeitpunkt t erzeugt, schon während des Auswertevorgangs einbezieht.

Insofern ist die Suche nach neuen Verfahren auf diesem Gebiet bei weitem noch nicht abgeschlossen.

Literaturverzeichnis

[ALLA89] Richard Lacey, Falo Alto, *Baseline Correction for Chromatography*, US Patent Number 4,802,102

[ATB93] Analysentechnik Beringer, *ATB's Analysenttechnische Informationen, Band 1*, im Eigenverlag der ATB Analysentechnik Beringer Vertriebsgesellschaft mbH, Mainz-Kastel, 2. Erweiterte Auflage November 1993.

[BO76] A. A. Borowkow, *Wahrscheinlichkeitstheorie – Eine Einführung*, Akademie-Verlag Berlin 1976, Seite 76.

[DIN01] *DIN 38 402, Deutsche Einheitsverfahren zur Wasser-, Abwasser und Schlammuntersuchung*

[DYS90] Norman Dyson, *Chromatographic Integration Methods*, The Royal Society of Chemistry 1990

[ECK99] Bruce Eckel, *Thinking In C++*, Prentice Hall PTR 1999, http://www.BruceEckel.com/ThinkingInCPP2e.html

[FAR94] Gerald Farin, *Kurven und Flächen im CAGD*, 2. Auflage, Vieweg-Verlag 1994

[FRI97] Bernd Fritzke, *Some Competetive Learning Methods*, Ruhr-Universität Bochum 1997, http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/JavaPaper/

[GDI93] Kumar K. Goswami, Murthy Devarakonda, Ravishankar K. Iyer, *Prediction-Based Dynamic Load-Sharing Heuristics*, IEEE Transactions on Parallel and Distributed Systems, Vol. 4. No. 6, June 1993.

[GOT93] Siegfried Gottwald, *Fuzzy Sets and Fuzzy Logik*, Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1993

[HER95] B. Herrenberger, *Visuelle Basisliniendetektion*, Dissertation, Fakultät für Informatik und Automatisierung der Technischen Universität Ilmenau, 20.06.1995

[HIT92] Hitachi, Ltd., *Verfahren zum Durchführen einer Chromatographieanalyse von Proben und System zur Anwendung desselben*, Offenlegungsschrift DE 42 04 853 A 1, Deutsches Patentamt 1992

[HP87] *HP3350 User Reference Manual*, Hewlett Packard 1987

[HP93] Hewlett Packard, *ChemStation Handbuch*, Hewlett Packard 1993.

[HS98] M. Hahn, P. Sivers, *Eine neue Detektionsmöglichkeit zur dünnschichtchromatografischen Bestimmung von schwerflüchtigen Kohlenwasserstoffen*, Tagungsband InCom98, S. 179

[LEST84] E. Leibnitz, H.-G. Struppe, *Handbuch der Gaschromatographie*, Akademische Verlagsgeselschaft GEEST & Portig K.-G., Leipzig 1984

[PAZU89] Pierre Parent, Steven W. Zucker, *Trace Inference, Curvature Consistency, and Curve Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11. No. 8, August 1989.

[POKO85] Peter Pospisil, Bruno Kolb, *Einrichtung zur Kompensation der Basisliniendrift einer chromatographischen Trennsäule*, Offenlegungsschrift DE 33 23 744 A1, Deutsches Patentamt 1985

[SG64] A. Savitzky, M. J. E. Golay, *Smoothing and Differention of Data by Simplified Least Squares Procedures*, Journal of Chromatography, Vol 36, No 8, July 1964

[TOM92] Barret Tomlinson, *Neural net system for analyzing chromatographic peaks*, US Patent Number 5,121,443

[WAA92] *Statistik un Qualitätskontrolle in der Analytik*, Wissenschaftlicher Arbeitskreis Analytik, Arbeitskreis ,,Bewertung von Analysenergebnissen", März 1992

[WEI98] Ch. Weidling, *Praktikumsbericht*, Fak. IA, TU Ilmenau 1998

[WEI98S] Ch. Weidling, *Automatische Auswertung von Gaschromatogrammen*, Studienarbeit, Fak. IA, TU Ilmenau 1998

[ZI89] E. Ziegler, *Die COLACHROM-Kommandosprache für chromatographische Datenverarbeitung*, in: G. Gauglitz (ed.), Software-Entwicklung in der Chemie 3, Springer-Verlag Berlin Heidelberg 1989, 201-211.