technische universität
dortmund

Diplomarbeit

Datenreduktion und Merkmalsextraktion bei Ionen-Mobilitäts-Spektrometrie-Messungen

Dominik Kopczynski

12. Mai 2010

Betreuer:
Prof. Dr. rer. nat Sven Rahmann
PD. Dr. rer. nat Jörg Ingo Baumbach

Fakultät für Informatik
Algorithm Engineering (Ls11)
Technische Universität Dortmund
http://ls11-www.cs.tu-dortmund.de

Inhaltsverzeichnis

Symbolverzeichnis

Ionen-Mobilitäts-Spektrometrie

R	Menge aller Retentionszeiten
r	bestimmte Retentionszeit
T	Menge aller Driftzeiten
t	bestimmte Driftzeit
S	vollständige IMS-Messung
$S_{r,t}$	Messwert zu Retentionszeit r und Driftzeit t

Vorverarbeitung

G	Gaußmatrix für Weichzeichnungsfilter
u	Summe aller Konzentrationen in S
a	Summe der Verteilungsdichten in G
d	Median aller Messwerte
d_t	Median einer Retentionsmessung zum Driftzeitpunkt t
h	Höchster Wert in einer Retentionsmessung
f_r	Faktor für Streckung einer Driftzeitmessung zum Retentionszeitpunkt r

EM-Algorithmus

n	Anzahl aller Messwerte in einer IMS-Messung
i	Index der Messwerte
c	Anzahl aller Modelle
j	Index der Modelle
x	Mehrdimensionale Eingabe in Wahrscheinlichkeitsdichtefunktion, hier (r, t)
ω	Gewicht eines Modells i in einer Messung j
$Z_{i,j}$	Zugehörigkeit von Messpunkt zu Modell
Θ	Parametertupel
Θ_j	Parametertupel vom Modell j
P_{Θ_j}	Verteilungsdichtefunktion mit Parametern des Modells j
β	Lagrange Multiplikator

Modell

σ	1. Parameter in Θ (Standardabweichung)
μ_t	2. Parameter in Θ (Erwartungswert / Verschiebung in Driftzeit)
λ	3. Parameter in Θ (Ereignisrate)
α	4. Parameter in Θ (Mittelwert)
μ_r	5. Parameter in Θ (Verschiebung in Retentionszeit)
v	6. Parameter in Θ (Volumen)

Modellmessung

W	Wahrscheinlichkeitsmatrix
K	Konzentrationsmatrix
A	Array für aufusmmierte Wahrscheinlichkeiten aller Modellpunkte

Kapitel 1

Einführung

1.1 Ionen-Mobilitäts-Spektrometrie

Ionen-Mobilitäts-Spektrometrie (IMS) ist ein Verfahren zur Konzentrationsmessung von flüchtigen Stoffen in Gasgemischen [10]. Bei dieser Methode wird der Analyt ionisiert und mit Hilfe eines elektrischen Feldes durch ein Gas geschickt. Zur Ionisierung dienen radioaktive Stoffe, die energiereiche Elektronen emitieren. Trifft so ein Elektron auf ein Molekül, wird dieses Molekül ionisiert. Wird als Trägergas Luft verwendet, entstehen Reaktionsionen. Diese zeichnen sich durch eine hohe Intensität in der Messung aus, welche als Reaktionsionen-Peak (RIP) bezeichnet wird und sich an einer bestimmten Driftzeit über den gesamten Retentionszeitraum erstreckt.

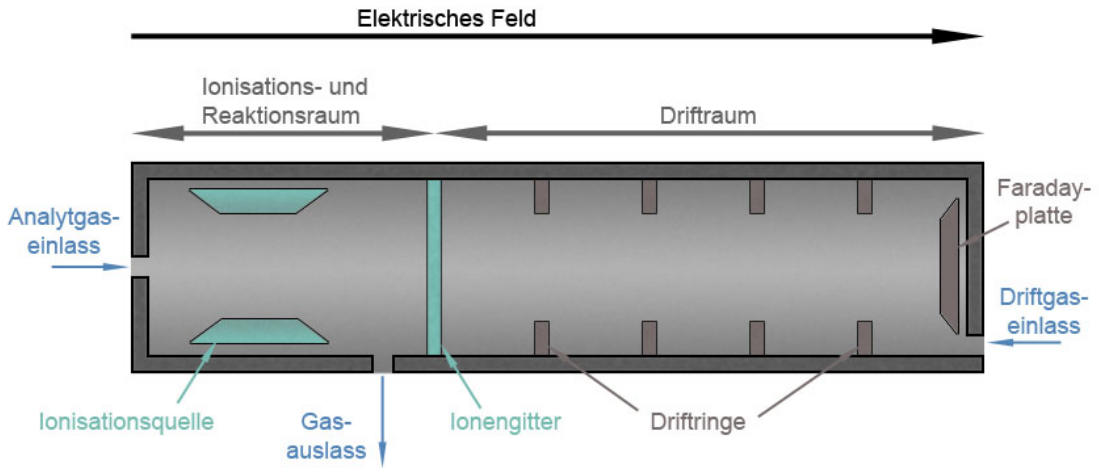


Abbildung 1.1: Querschnitt eines IMS-Messgerätes

Durch Reaktionen der Reaktionsionen mit oben genannten Molekülen entstehen Produkt-Ionen, welche sich bei der Messung als Peak manifestieren. Die Anzahl ionisierter Moleküle hängt von der Stärke der Ionisationsquelle ab, was nachteilig ist, da dadurch nicht die gesamte Menge des Analyten ionisiert wird.

Die Zeit, die die Ionen zum Durchqueren des Gases brauchen, wird Driftzeit genannt. Durch Kollisionen, die ein Abbremsen und Beschleunigen des Analytes bewirken, entsteht eine spezifische Driftzeit für ein Molekül. Da die Dirftzeit vor allem von der Masse und der Polarisierbarkeit des Analytes abhängt, wird häufig bei verschiedenen Stoffen die gleiche Driftzeit gemessen. Der Querschnitt eines IMS-Messgerätes ist in Abbildung 1.1 zu sehen.

Um eine genauere Messung zu bekommen, wird das Gas vorher durch eine Multikapillarsäule (Multi Capillary Column, MCC) geführt. Durch den Einsatz einer Multikapillarsäule wird das Gas in seine Bestandteile zerlegt, welche anschließend einzeln in die IMS-Messung gelangen. Beim Durchqueren des MCC gibt es zwei Phasen. Die erste Phase ist die stationäre Phase. An der Innenseite der Multikapillarsäule wird ein Gel angebracht, welches den Analyten an sich binden kann. Die zweite Phase ist die mobile Phase, dabei wird der Analyt mit Hilfe eines Trägergases durch die Säule gezogen. Durch Wechselwirkungen zwischen stationärer und mobiler Phase durchquert der Analyt die Säule in einer bestimmten Zeit, welche Retentionszeit genannt wird.

Die so gemessenen Konzentrationen können in ein Koordinatensystem übertragen werden, wobei die Driftzeit die x-Achse und die Retentionszeit die y-Achse darstellt. Punktwolken, die eine hohe Konzentraktion aufweisen, werden Peaks genannt. Diese Peaks haben eine charakteristische Form und zeichnen sich dadurch aus, dass sie mit zunehmender Retentionszeit "ausdünnen", also deren Konzentration sich langsam verringert.

Eine genaue Analyse der Messungen wird durch verschiedene Phänomene erschwert. Zum einen stehen bei jeder IMS-Messung nur eine bestimmte Anzahl von ionisierten Molekülen zur Verfügung, was zur Folge hat, dass, im Verhältnis zu den Konzentrationen der Stoffe eine Ionisierung statt findet und somit nur das Verhältnis der Stoffe zueinander wiedergeben wird und nicht deren tatsächliche Konzentration. Zum anderen können sich die gemessenen Konzentrationen aus den Summen mehrerer überlappender Peaks bilden, was eine eindeutige Zuordnung einer Konzentration zu einem Peak erschwert.

Ziel dieser Diplomarbeit ist, eine Merkmalsextraktion aus einer IMS-Messung durchzuführen, indem zuerst die Rohdaten der Messungen durch geeignete Filter eine Vorverarbeitung erfahren. Im anschließenden Verfahren werden die Konzentrationen mittels eines geeigneten Clustering-Verfahrens ihren zugehörigen Peaks zugeordnet. Anschließend findet eine Datenreduktion statt, indem die Messdaten in ein geeignetes Modell überführt werden, das die Konzentrationen am besten beschreibt.
Der Vorteil eines Modells ist, dass für ein Modell erheblich weniger Daten gebraucht werden, als bei einer Messung. Ein geeignetes Modell besteht aus einer mathematischen Formel, die eine konstante Anzahl von variablen Parametern hat. Durch ein

iteratives Verfahren sollen die Parameter für ein Modell mit Hilfe der Messdaten eines Peaks so genau wie möglich geschätzt werden.

1.2 IMS-Messung

Abbildung 1.2 zeigt ein musterhaftes Beispiel einer IMS-Messung. Dabei steht die x-Achse für die Driftzeit und die y-Achse für die Retentionszeit. Die Messung lässt sich als $\mathbb{R}^2 \rightarrow \mathbb{R}$ Funktion verstehen, da bei der Eingabe von Driftzeit und Retentionszeit eine Konzentration als Ausgabe herauskommt. Die Peaks heben sich deutlich vom Hintergrundrauschen ab und sind mit zunehmender Konzentration von blau, über rot bis gelb mit der höchsten Konzentration zu interpretieren.
Der gelbe senkrechte Streifen an der Driftzeit 980 wird als RIP bezeichnet und kommt in jeder IMS-Messung vor, in der Luft als Trägergas dient. Das Hintergrundrauschen ist blau/weiß dargestellt. Alle in der Diplomarbeit verwendeten Bilder einer IMS-Messung oder eines Modells einer Messung verstehen sich mit den gleichen Angaben.

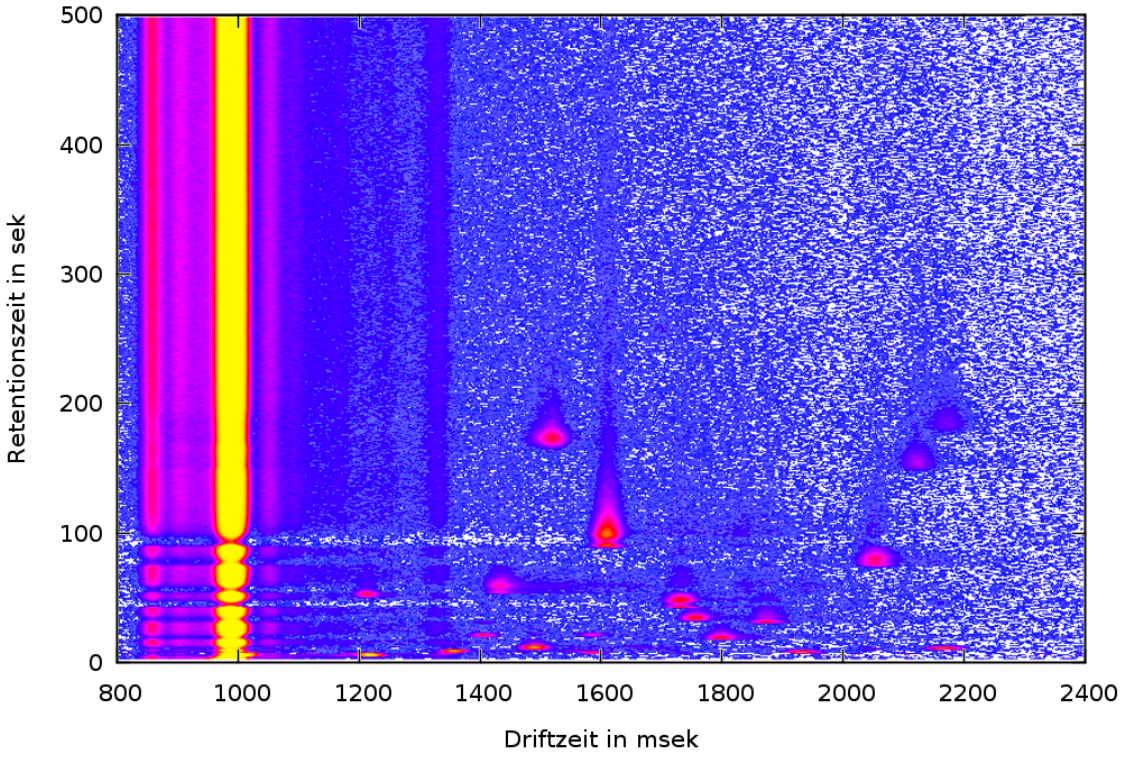


Abbildung 1.2: Musterhaftes Beispiel einer IMS-Messung

1.3 Überblick

Im kommenden Kapitel wird erläutert, welche Filter bei der Vorverarbeitung eingesetzt werden und welchem Zweck sie dienen. Im dritten Kapitel wird der zum Einsatz kommende EM-Algorithmus beschrieben, welcher sowohl das Problem der prozentualen Zugehörigkeiten von Konzentrationen beim Clustering, als auch die Parameterschätzung bei der Modellüberführung löst.
Im vierten Kapitel wird nach einem geeigneten Modell gesucht. Dabei werden verschiedene Peaks betrachtet, um Gemeinsamkeiten in ihrer Form zu finden. Anschließend wird versucht diese Gemeinsamkeiten mit mathematischen Formeln zu erfassen und zu beschreiben. Ausserdem wird ein Programm zur Erstellung von idealen Modellmessungen zur Validierung der korrekten Funktionsweise des EM-Algorithmus erstellt.
Das fünfte Kapitel beschreibt die Modifikation des EM-Algorithmus zur Parameterschätzung der Modelle für ihre Peaks. Dabei werden im ersten Schritt Funktionen für die Parameter gefunden, die im Modell verwendet werden, um die Schätzung zu verbessern. Im nächsten Schritt wird der Vorgang des EM-Algorithmus beschrieben und anschließend einheitliche Analyseergebnisse erarbeitet.

Kapitel 2

Vorverarbeitung

Die Rohdaten einer IMS-Messung sind die Konzentrationen der Moleküle im Messgas, die zur einer bestimmten Drift- und Retentionszeit gemessen wurden. Da jedoch die Rohdaten ungeeignet zur weiterführenden Analyse sind, benötigen sie eine Vorverarbeitung, um anschließend korrektere Ergebnisse bei der Merkmalsextraktion zu liefern. Zudem soll durch die Normalisierung der Daten ein einheitliches Ergebnis für mehrere Messungen entstehen.

2.1 Niveaukorrektur

Bei einer unbearbeiteten IMS-Messung liegt das durchschnittliche Niveau des Hintergrundrauschens sehr hoch. Deshalb ist es erforderlich, dass zur optimalen Hauptanalyse das Hintergrundniveau erheblich gesenkt wird und im Idealfall sogar den Nullwert erreicht.
Zuerst muss ein Mittelwert d ermittelt werden. Hierbei bietet sich nicht das arithmetische Mittel, sondern der Median an, da das arithmetische Mittel durch besonders hohe Peaks verfälscht werden kann und dadurch die Werte besonders stark reduziert werden. Zudem bestehen die meisten Werte in der Messung aus niedrigem Hintergrundrauschen und nur wenige Werte aus hohen Peaks. Tests haben gezeigt, dass immer deutlich mehr als 50% der Konzentrationen zum Hintergrundrauschen gehören. Somit sollen nun alle Werte um den Median gesenkt werden.

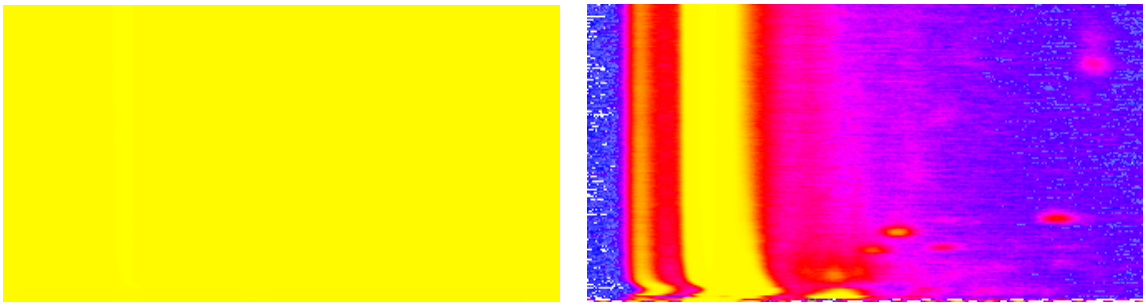
Eine schnelle Möglichkeit, den Median in einer Liste von Zahlen zu finden, ist, einen modifizierten Quicksort anzuwenden. Der originale Quicksort-Algorithmus teilt die Liste am Pivot-Element in zwei neue Listen auf und führt sich selbst rekursiv auf beiden Listen aus. Der modifizierte Algorithmus überprüft nach der Teilung am Pivot-Element, in welcher der beiden Listen sich der Rang des Medians befindet und führt nur in dieser Liste den rekursiven Aufruf durch. Heraus kommt eine unsortierte Liste, bei der jedoch der Median an seinem korrekten Rang in der Liste steht. Der Vorteil dieser Methode ist, dass der modifizierte Quicksort bei einer Liste mit n Elementen im Durchschnitt eine Laufzeit von $\mathcal{O}(n)$ hat. Der modifizierte Code (vgl. B.F.P.R.T. [3]) sieht folgendermaßen aus:

```
function QUICKSORT(Liste, links, rechts)
    if (rechts − links) > 1 then
        mitte = LISTE.ANZAHL / 2
        pivot = PARTITION(Liste, links, rechts)
        if (links <= mitte) and ((pivot − 1) >= mitte) then
            QUICKSORT(Liste, links, pivot − 1)
        else
            QUICKSORT(Liste, pivot + 1, rechts)
        end if
    end if
end function
```

Ist der Median aller Messpunkte d ermittelt, wird er von allen Werten in der Messung subtrahiert. Sollte dabei ein Wert kleiner als 0 werden, wird der Wert auf 0 gesetzt:

$$\forall r \in R, t \in T: \quad S'_{r,t} := \begin{cases} S_{r,t} - d & \text{wenn } S_{r,t} - d \geq 0, \\ 0 & \text{sonst.} \end{cases} \tag{2.1}$$

In Abbildung 2.1 ist beim Rohbild (a) zu erkennen, dass durch das viel zu hohe Grundniveau sowohl vom Betrachter als auch vom anschließenden Alayseverfahren nichts zu erkennen ist. Nach dem Korrekturfilter sind deutlich die Peaks und der RIP zu erkennen.



(a) Unbehandeltes Messbild (b) Messbild mit Niveaukorrektur

Abbildung 2.1: Grundniveau wird auf einen niedrigen Wert gesenkt

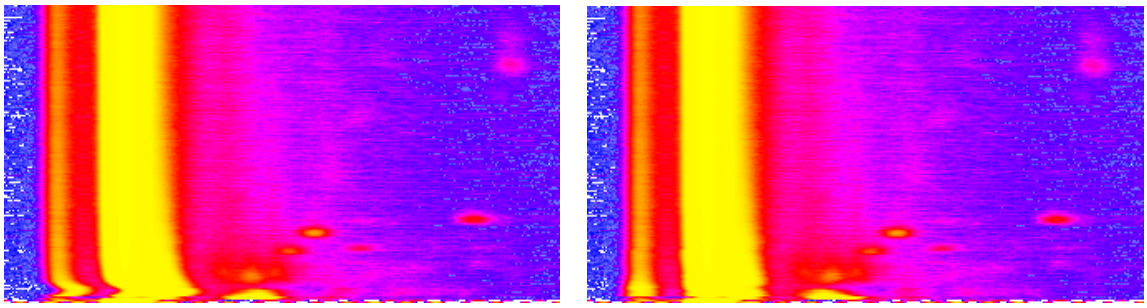
2.2 Feuchtekorrektur

Bei der Verwendung von Luft als Trägergas werden die IMS-Messungen bei niedriger Retentionszeit stakt verfälscht. Dies macht sich durch eine Verschiebung der Messdaten in der Driftzeit bemerkbar. Durch eine simple Korrektur lässt sich dieser Messfehler beheben (vgl. Boedecker [4]).

Sei t^* der Index der Spalte in der Messung, welche die höchste Summe aller aufsummierten Werte in einer Spalte hat.

$$t^* := \text{argmax}(\sum_{r=1}^{R} S_{r,t}). \tag{2.2}$$

Somit ist t^* die Referenzdriftzeit und sollte direkt auf dem RIP liegen. Nun wird für alle Retentionszeiten mit der Intervallgranze x das Intervall $[t^*-x; t^*+x]$ betrachtet. In diesem Intervall wird der höchste Wert gesucht. Befindet sich der Wert nicht an der Stelle t^*, wird die gesamte Zeile verschoben, bis der höchste Wert des Intervalls an der Referenzstelle t^* steht. Für die nächste Zeile wird das Intervall $[t' - x; t' + x]$ betrachtet, wobei t' die Stelle des höchsten Wertes aus der vorherigen Zeile vor der Verschiebung war. Dabei darf x nicht zu groß gewählt werden, da es sein kann, dass mit dem RIP überlappende Peaks eine höhere Konzentration haben, als der RIP selber und diese Zeile dann extrem verschoben wird. Das Ergebnis der Korrektur lässt sich in Abbildung 2.2 sehen.



(a) Vor der Feuchtekorrektur (b) Nach der Feuchtekorrektur

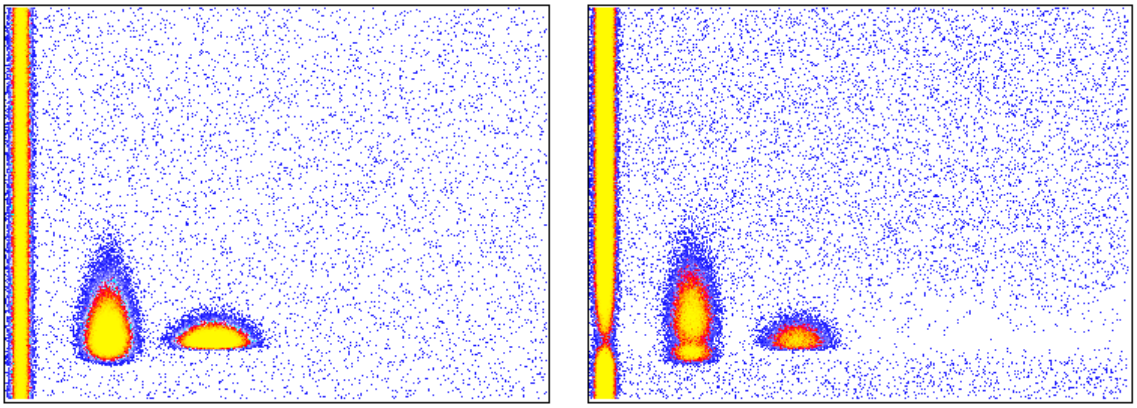
Abbildung 2.2: Feuchtekorrektur durch Geradeziehen des RIPs

2.3 Ionenkompensation

Wie bereits in der Einleitung erwähnt, werden nicht alle in der Ionisierungsphase befindlichen Moleküle ionisert, wodurch es zu einer Verfälschung der Messung kommt. Je mehr verschiedene Moleküle mit gleicher Retentionszeit es gibt, desto weniger Moleküle aus einer Stoffgruppe werden ionisiert. In den Messbildern macht sich diese Verteilung bemerkbar, indem die Peaks und der RIP wie zugeschnürt wirken.

Wie in Abbildung 2.3 zu sehen ist, sind die Konzentrationen bei einem idealen IMS-Messbild nicht auf die konstante Anzahl von Ionen in den Zeilen (Driftzeit) beschränkt. Das Hintergrundrauschen ist weitgehend gleichverteilt. Bei reelen Messbildern hingegen ist deutlich die Einschnürung der Peaks und des RIPs zu erkennen. Die Peaks werden generell vom Volumen her geringer und im ungünstigsten Fall findet eine Deformation des Peaks statt. Die Form wirkt eingedrückt und zugeschnürt. Es entsteht sogar der Eindruck, dass es sich beim eingeschnürten Peak um zwei ver-

tikal überlappende Peaks handeln könnte. Das Hintergrundrauschen ist zu diesen Retentionszeiten geringer, als im Verhältnis zu den übrigen Retentionszeiten.



(a) Ideales Messbild (b) Reales Messbild

Abbildung 2.3: Konkurrenz zwischen Ionen pro Driftzeitmessung

Eine optimale Kompensation dieses Phänomens würde den Rahmen der Diplomarbeit sprengen, da diese Ionenkonkurrenz von der Reaktionsfreudigkeit der Moleküle mit den Reaktionsionen abhängt. Zudem scheinen Peaks bei zunehmender Driftzeit weniger von dem Phänomen betroffen zu sein als zu Anfang der Messung.
Um eine korrekte Rekonstruktion der Messdaten durchführen zu können, müssten vor der Vorverarbeitung also schon die Moleküle und die Volumina der Peaks bekannt sein. Das nachfolgende Verfahren beschreibt lediglich eine vereinfachte Methode der Kompensation unter der Annahme, dass alle Werte in einer Zeile um einen bestimmten Faktor gestaucht sind.

Sei u die Summe aller Konzentrationen in der Messung und t^* der Index der Spalte in der Messung, welche die höchste Summe aller aufsummierten Werte in einer Spalte hat

$$t^* := \text{argmax}(\sum_{r=1}^{R} S_{r,t}). \tag{2.3}$$

Somit sollte t^* direkt auf dem RIP liegen. In der Spalte wird der größte Wert h gesucht. In jeder Zeile r muss nun der Faktor f_r ermittelt werden, um den alle Werte in jener Zeile gestreckt werden müssen. Dazu wird h durch den Wert an der Stelle $S_{r,t}$ geteilt:

$$f_r := \frac{h}{S_{r,t}} \qquad \forall r \in R \tag{2.4}$$

Anschließend werden alle Werte in der Zeile mit dem Faktor f_r multipliziert:

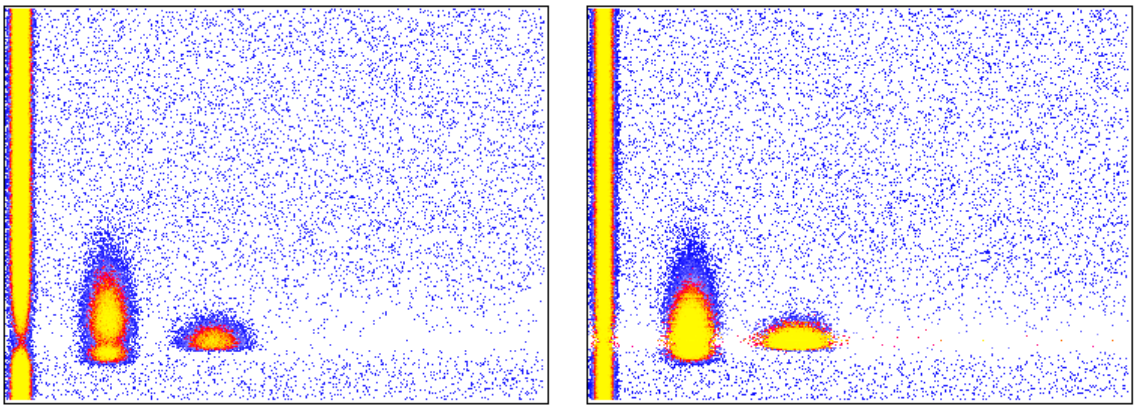
$$S'_{r,t} := S_{r,t} \cdot f_r \qquad \forall r \in R, \quad t \in T \tag{2.5}$$

Zum Schluss müssen die Werte wieder normalisiert werden, so dass die Summe aller Werte wieder dem ursprünglichen u entsprechen. Dazu sei u' die Summe aller neuen

Konzentrationen. Im letzten Schritt wird für alle Werte zur Normalisierung die neue Konzentration berechnet, indem die Werte durch u' geteilt und anschließend mit u multipliziert werden:

$$S''_{r,t} := u \cdot \frac{S'_{r,t}}{u'} \qquad \forall r \in R, \quad t \in T \tag{2.6}$$

Die Abbildung 2.4 veranschaulicht die Kompensation der Messdaten. Es ist deutlich zu erkennen, dass das Volumen der Peaks angewachsen ist und dass die Einschnürungen beseitigt sind, jedoch lässt sich mit diesem Verfahren kein homogenes Hintergrundrauschen erzeugen. Viel mehr steigen die Werte des Hintergrundrauschens zu stark an. Wenn jedoch nur an wenigen Stellen das Hintergrundrauschen zu stark ansteigt oder nicht in der Nähe der Peaks liegt, finden bei der anschließenden Merkmalsextraktion keine Verfälschungen statt.

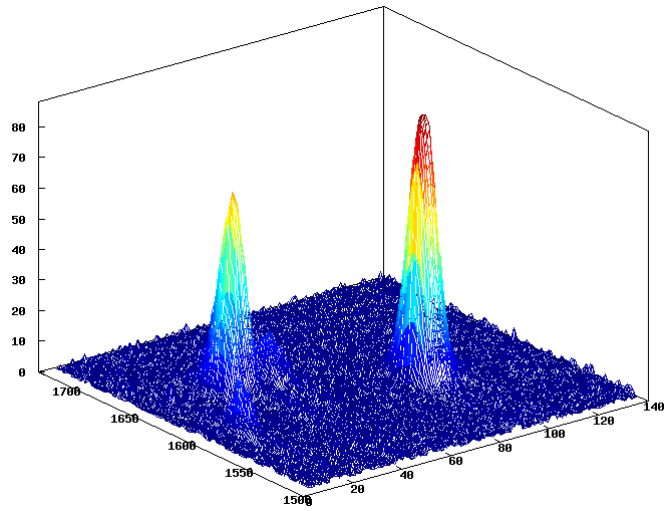


(a) Vor der Kompensation (b) Nach der Kompensation

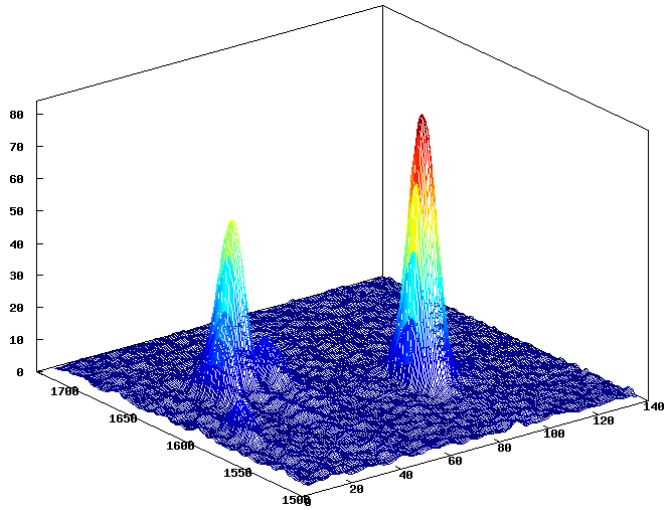
Abbildung 2.4: Kompensationsfilter

2.4 Gaußglättung

Im nächsten Schritt der Vorverarbeitung wird eine sogenannte Gaußglättung durchgeführt, wie auch in [12] vorgestellt. Die Gaußglättung ist hauptsächlich aus der Bildverarbeitung bekannt und lässt die Bilder unscharf werden. Da es in den Messungen vereinzelte Konzentrationen gibt, die im Gegensatz zu ihren benachbarten Konzentrationen verhältnismäßig gering oder hoch sind, ist der Gaußfilter ideal, um den ganzen Peak zu glätten und die abstehenden Konzentrationen ins Verhältnis zu den anderen Konzentrationen zu rücken.
Die Idee der Gaußglättung ist, dass eine Konzentration an einer Stelle abängig von ihren benachbarten Konzentrationen berechnet wird. Dabei fließt das Verhältnis der eigenen Konzentration am meisten in die Berechnung ein, wo hingegen die Konzentrationen in einem viel geringerem Verhältnis in die Rechnung einfließen, je weiter sie sich von der betrachteten Konzentration befinden. Der Name Gaußglättung kommt daher, weil als Verteilungsfunktion für die Verhältnisse eine zweidimensionale Normal-, bzw. Gaußverteilung zum Einsatz kommt.



(a) Vor der Gaußglättung



(b) Nach der Gaußglättung

Abbildung 2.5: Gaußglättungsfilter

Gegeben sei eine $m \times n$ Matrix G, welche als Werte die Dichteverteilung einer zweidimensionalen Normalverteilung hat. Die Stärke, mit der der Gaußfilter die Peaks glättet, ist abhängig von den Parametern λ und σ der Normalverteilung, sowie der Größe der Matrix. Im folgenden Beispiel ist die Größe der Matrix $m = 5$ und $n = 5$ und die Verteilung eine Standardnormalverteilung mit $\sigma = 1$ und $\lambda = 2$, damit der mittlere Wert in der Matrix dem Modalwert der Verteilung entspricht. Mit der Wahrscheinlichkeitsdichte

$$f(x, y) = \frac{1}{2\pi\sigma} \exp\left(-\frac{(x^2 - \lambda)\,(y^2 - \lambda)}{2\sigma^2}\right) \tag{2.7}$$

entsteht nun folgende Matrix

$$G(x,y) := \begin{pmatrix} 0.003 & 0.013 & 0.022 & 0.013 & 0.003 \\ 0.013 & 0.059 & 0.097 & 0.059 & 0.013 \\ 0.022 & 0.097 & 0.159 & 0.097 & 0.022 \\ 0.013 & 0.059 & 0.097 & 0.059 & 0.013 \\ 0.003 & 0.013 & 0.022 & 0.013 & 0.003 \end{pmatrix}. \tag{2.8}$$

Sei a die Summe aller Werte in der Matrix G und entspricht nun 0.982. Nun werden alle Konzentrationen neu berechnet. Die neu zu berechnende Konzentration bekommt das größte Verhältnis aus der Matrix, was dem Mittelpunkt in der Matrix entspricht. Die umliegenden Konzentrationen werden in topologischer Relation zum Mittelpunkt in der Matrix mit ihren jeweiligen Verhältnissen multipliziert, zuaddiert und zur Normierung durch a geteilt. So entsteht eine neue geglättete Konzentration mit folgender Berechnung:

$$S'_{r,t} := \frac{1}{a} \cdot \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} S_{r+x-\lambda,t+y-\lambda} \cdot G(x,y) \qquad \forall r \in R, t \in T \tag{2.9}$$

Anschließend wird eine Normalisierung durchgeführt, wie bereits in Kapitel 2.3 beschrieben. In Abbildung 2.5 ist zu erkennen, dass die Peaks deutlich glatter wirken und das Niveau des Hintergrundrauschens nochmal gesenkt wurde.

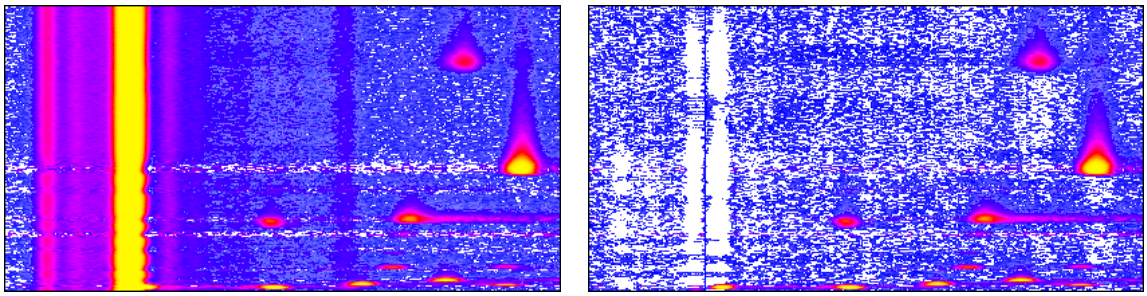
2.5 Basislinienkorrektur

Als letzter Filter bei der Vorverarbeitung der Daten wird eine Basislinienkurrektur durchgeführt. In realen Messbildern ist zu erkennen, dass ab dem RIP mit zunehmender Driftzeit der Pegel des Hintergrundrauschens abnimmt. Auch der RIP wird für die eigentliche Analyse der Daten nicht verwendet. Beides kann mit Hilfe der Basislinienkorrektur behoben werden.

Genau wie bei der Niveaukorrektur, wird hierbei auch der Median zur Korrektur benutzt. Für alle Spalten wird zunächst der Median d_t gesucht. Ist d_t für eine Spalte ermittelt, wird er von allen Werten in einer Spalte subtrahiert. Sollte dabei ein Wert kleiner als 0 werden, wird der Wert auf 0 gesetzt:

$$\forall r \in R, t \in T: \quad S'_{r,t} := \begin{cases} S_{r,t} - d_t & \text{wenn } S_{r,t} - d_t \geq 0, \\ 0 & \text{sonst.} \end{cases} \tag{2.10}$$

Abbildung 2.6 veranschaulicht die Basislinienkorrektur. Dabei ist ersichtlich, dass der RIP fast komplett verschwindet und sich das Niveau des Hintergrundrauschens deutlich gesenkt hat. Nach der Korrektur ist das Niveau des Hintergrundrauschens auch weitgehend homogen über die Anzahl der Spalten geworden. Die Peaks sind darüber hinaus erhalten geblieben und haben sich kaum geändert. Da der RIP über die komplette Retentionszeit ziemlich konstant verlief, wurde er komplett herausgefiltert, was die vertikalen Streifen erklärt.



(a) Vor der Korrektur (b) Nach der Korrektur

Abbildung 2.6: Basislinienkorrektur

Kapitel 3

EM-Algorithmus

Nach wie vor gilt es nun, nach der Vorverarbeitung, die Messung zu analysieren und die Peaks zu parametrisieren. Bei Betrachtung einer Messung liegt es nahe, ein Clusteringverfahren anzuwenden, um die Konzentrationen eindeutig einem Peak zuzuordnen. Anschließend können die einzelnen Cluster in ein Modell überführt werden. Das große Problem, das sich hierbei ergibt, ist, dass schon die Messungen solch ein Vorgehen nicht zulassen. Die Messungen haben die Eigenschaft, dass Peaks sich überschneiden können und somit eine eindeutige Zuordnung der Konzentrationen zu einem Peak überhaupt nicht möglich ist.
Viele Clustering Verfahren, wie zum Beispiel K-means, QT-means oder Hauptkomponentenanalyse, haben für dieses Vorhaben den Nachteil, dass sie hart sind, also einen Punkt ganz oder gar nicht einem Cluster zuordnen (vgl. [7]). Diese Eigenschaft würde es bei sich überlappenden Peaks unmöglich machen, ein (oder mehrere) Modell(e) zu extrahieren. Eines der Verfahren, das eine prozentuale Zuordnung zu Clustern beherrscht, ist der EM-Algorithmus.

3.1 Einführung in den EM-Algorithmus

Der EM-Algorithmus ist in erster Linie ein probabilistisches Verfahren zur Parameterschätzung von Modellen [2]. Lediglich als Nebenprodukt entsteht dabei das Clustering. EM steht hierbei für Expectation-Maximization und ist ein iteratives Verfahren. Für eine erfolgreiche Benutzung des EM-Algorithmus muss vor allen Dingen ein Modell mit einer Wahrscheinlichkeitsdichte $p(x)$ vorhanden sein, welches detailliert in Kapitel 4 beschrieben wird. X ist dabei eine mehrdimensionale Zufallsvariable. Der EM-Algorithmus unterscheidet drei Klassen von Daten.
Die erste Datenklasse sind die beobachteten Daten, in diesem Fall die Konzentration $S_{r,t}$ zu einer bestimmten Retentionszeit r und Driftzeit t in der IMS-Messung. Bei der zweiten Klasse handelt es sich um versteckte Parameter. Jede Messung/Konzentration hat eine bestimmte Anzahl von versteckten Parametern und zwar so viele, wie es Modelle in der Analyse gibt. Diese geben den prozentualen Anteil der Konzentration zum zugehörigen Modell an. Die dritte Klasse von Daten sind die End- oder Modellparameter. Die Anzahl der Modellparameter hängt von der Art und der Anzahl der Modelle ab.

Bevor der Algorithmus starten kann, müssen sowohl die versteckten Parameter, als auch die Modellparameter vorinitialisiert werden. Bei den versteckten Parametern wird lediglich eine Gleichverteilung pro Konzentration angewendet, da vorher nicht bekannt ist, zu welchem Anteil eine Konzentration zu einem Modell gehört. Eine gute Wahl der Modellparameter ist ausschlaggebend für eine erfolgreiche Analyse. Je besser die Modellparameter bei der Initialisierung geschätzt werden, desto präziser verläuft anschließend die Analyse, da sonst bei ungünstiger Wahl der Startkoordinaten der Modelle nur das Hintergrundrauschen vom Algorithmus erkannt wird und daher keine korrekten Ergebnisse erzielt werden.
Im ersten Schritt, dem Expectation-Schritt, werden zuerst alle versteckten Parameter geschätzt. Dies geschieht, indem mit Hilfe der Wahrscheinlichkeitsdichtefunktion des Modells und den versteckten Parametern des jeweiligen Modells zuerst eine bedingte Wahrscheinlichkeit ausgerechnet wird, mit der eine Konzentration von einem Modell erzeugt wird und anschließend durch die Summe aller Wahrscheinlichkeiten einer Konzentration geteilt wird.
Im zweiten Schritt, dem Maximization-Schritt, werden die Modellparameter geschätzt. Für gewöhnlich wird hierbei die Maximum-Likelihood-Methode angewendet. Mit Hilfe dieser Methode ist es möglich, geschlossene Formeln für die Modellparameter zu finden, die jeweils von den Beobachtungen und den versteckten Parametern abhängig sind.
Die beiden Schritte werden so lange wiederholt, bis entweder alle Modellparameter konvergieren oder die Likelihood-Funktion lokal maximiert ist.

Die Vorteile des EM-Algorithmus sind, dass es mit Hilfe der Methode möglich ist, eine Parameterschätzung und ein Clustering gleichzeitig durchzuführen. Der Algorithmus verbessert nach jedem Schritt die Likelihood und terminiert auf jeden Fall. Nachteil des EM-Algorithmus ist, dass nicht immer das globale Maximum, sondern auch nur ein lokales Maximum erreicht wird und es dementsprechend eine (viel) bessere Lösung geben kann. Zudem muss die Anzahl der Modelle bekannt sein und die Modellparameter müssen zu Beginn gut geschätzt werden, damit der Algorithmus auch die Daten der Peaks und nicht des Hintergrundrauschens analysiert.
Da die Vorteile, die der Algorithmus hat, deutlich für die Verwendung sprechen, wurde für die Datenreduktion und Merkmalsextraktion der EM-Algorithmus benutzt.

3.2 EM-Algorithmus für Mixture Modelle

Üblicherweise wird für die Verwendung des EM-Algorithmus zuerst eine Maximum-Likelihood-Funktion aufgestellt. Gegeben ist folgende Ausgangssituation: Es gibt $n = R \cdot T$ Messwerte x_i, wobei $1 \leq i \leq n$ ist. Darüber hinaus sei j der Index des Modells aus insgesamt c Modellen. Jedes Modell hat ein Gewicht ω_j, mit dem es zur Messung gehört. Für die Summe der Gewichte aller Modelle muss gelten $\sum_{j=1}^{c} \omega_j = 1$.

Die versteckten Parameter beschreiben die Zugehörigkeit $Z_{i,j}$ eines Messwertes x_i

zum Modell j, wobei $Z_{i,j} = 1$ ist, wenn x_i eindeutig zum Modell j gehört, ansonsten ist $Z_{i,j} = 0$. Es gilt also $\mathbb{P}_{\Theta,\omega}(Z_{i,j} = 1) = \omega_j$.
Desweiteren besitzt jedes Modell eine Verteilungsfunktion $P_{\Theta_j}(x_i)$, wobei Θ_j alle Parameter eines Modells einschließt. Für eine erfolgreiche Rechnung müssen alle Modelle und somit ihre Verteilungsfunktionen bekannt sein. Die Wahrscheinlichkeit einen Messwert x_i mit dem Modell j zu erzeugen ist folglich

$$\mathbb{P}_{\Theta,\omega}(x_i, Z_{i,j} = 1) = \omega_j \cdot P_{\Theta_j}(x_i). \tag{3.1}$$

Um nun das Mischungsverhältnis zu beschreiben, wird die Wahrscheinlichtkeit für einen Messwert x_i mit Hilfe eines beliebigen c-dimensionalen Vektors Z_i über alle Modelle gebildet

$$\mathbb{P}_{\Theta,\omega}(x_i, Z_i) = \prod_{j=1}^{c} \left[\omega_j \cdot P_{\Theta_j}(x_i)\right]^{Z_{i,j}}. \tag{3.2}$$

Für alle Messpunkte ergibt sich nun folgende Gesamt-Likelihood-Funktion

$$L_{x,Z}(\Theta, \omega) = \prod_{i=1}^{n} \prod_{j=1}^{c} \left[\omega_j \cdot P_{\Theta_j}(x_i)\right]^{Z_{i,j}}. \tag{3.3}$$

Da der Logarithmus eine streng monoton wachsende Funktion ist, kann folglich zur Vereinfachung der Formel auch der Logarithmus der Likelihood-Funktion maximiert werden,

$$\mathcal{L}_{x,Z}(\Theta, \omega) = \sum_{i=1}^{n} \sum_{j=1}^{c} Z_{i,j} \cdot \left[\log \omega_j + \log P_{\Theta_j}(x_i)\right]. \tag{3.4}$$

3.3 Berechnungen im E-Schritt

Wie beschrieben, werden im E-Schritt die versteckten Parameter geschätzt. Hierzu muss zuerst eine Zielfunktion aufgestellt werden, mit der die versteckten Parameter unter Berücksichtigung der Parameter Θ^0 und der Gewichte der Modelle ω^0 aus dem vorherigen Iterationsschritt geschätzt werden.

3.3.1 Bildung einer Zielfunktion

Die Zielfunktion hat die Aufgabe die Parameter Θ^* und ω^* mit Hilfe der neu berechneten versteckten Parameter zu maximieren. Dabei ist die Zielfunktion der Erwartungswert der Log-Likelihood-Funktion $\mathcal{L}_{x,Z}(\Theta, \omega)$ über $Z_{i,j}$, welcher als bedingte

Verteilung der $Z_{i,j}$ gegeben ist, unter der Vorraussetzung, dass x_i mit den aktuellen Parameterwerten Θ^0 und ω^0 gelten

$$\begin{aligned} f_{\Theta^0,\omega^0,x}(\Theta,\omega) &= \mathbb{E}_{(Z|(x;\Theta^0,\omega^0))}\mathcal{L}_{x,Z}(\Theta,\omega) \\ &= \mathbb{E}_{(Z|(x;\Theta^0,\omega^0))}\left[\sum_{i=1}^{n}\sum_{j=1}^{c} Z_{i,j}\cdot\left[\log\omega_j + \log P_{\Theta_j}(x_i)\right]\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c}\mathbb{E}_{\Theta^0,\omega^0}\left[Z_{i,j}|x_i\right]\cdot\left[\log\omega_j + \log P_{\Theta_j}(x_i)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c} Z^0_{i,j}\cdot\left[\log\omega_j + \log P_{\Theta_j}(x_i)\right]. \end{aligned} \tag{3.5}$$

3.3.2 Berechnung der versteckten Parameter

Unter der Berücksichtigung der Parameter Θ^0 und ω^0, die aus der vorherigen Iteration berechnet wurden, wird $Z^0_{i,j}$ berechnet, welches für die Zielfunktion benötigt wird

$$Z^0_{i,j} = \mathbb{E}_{\Theta^0,\omega^0}\left[Z_{i,j}|x_i\right]. \tag{3.6}$$

Dieser Erwartungswert ist nichts anderes als die Wahrscheinlichkeit, dass $Z_{i,j} = 1$ ist, unter der Voraussetzung, dass x_i gegeben ist und die aktuellen Parameterwerten Θ^0 und ω^0 gegeben sind

$$Z^0_{i,j} = \mathbb{P}_{\Theta^0,\omega^0}\left[Z_{i,j} = 1|x_i\right]. \tag{3.7}$$

Unter Anwendung des Satzes von Bayes für bedingte Wahrscheinlichkeiten kann $Z^0_{i,j}$ in eine Rechnung mit bekannten Wahrscheinlichkeitsfunktionen überführt werden

$$Z^0_{i,j} = \frac{\mathbb{P}_{\Theta^0,\omega^0}\left[x_i|Z_{i,j}=1\right]\cdot\mathbb{P}_{\Theta^0,\omega^0}\left[Z_{i,j}=1\right].}{\mathbb{P}_{\Theta^0,\omega^0}\left[x_i\right]} \tag{3.8}$$

Die Wahrscheinlichkeit für $\mathbb{P}_{\Theta^0,\omega^0}\left[Z_{i,j}=1\right]$ ist genau ω^0_j, wie schon im vorherigen Abschnitt definiert. Auch $\mathbb{P}_{\Theta^0,\omega^0}\left[x_i|Z_{i,j}=1\right]$ wurde bereits definiert und ist die Verteilungsfunktion eines Modells $P_{\Theta^0_j}(x_i)$. $\mathbb{P}_{\Theta^0,\omega^0}\left(x_i\right)$ ist die Wahrscheinlichkeit für einen Punkt x_i, welche als gewichtete Summe über die Wahrscheinlichkeiten aller Modelle gebildet wird: $\sum_{k=1}^{c}\omega^0_k\cdot P_{\Theta^0_k}(x_i)$. Die Gesamtformel für $Z^0_{i,j}$ sieht nun folglich so aus

$$Z^0_{i,j} = \frac{\omega^0_j\cdot P_{\Theta^0_j}(x_i)}{\sum_{k=1}^{c}\omega^0_k\cdot P_{\Theta^0_k}(x_i)}. \tag{3.9}$$

$Z^0_{i,j}$ ist also der gewichtete Anteil der Wahrscheinlichkeit von Punkt x_i zur Komponente j geteilt durch die Gesamtwahrscheinlichkeit von x_i mit den Parametern Θ^0 und ω^0. Durch diese Rechnung bleibt auch $\sum_{j=1}^{c} Z_{i,j} = 1$ gewahrt.

3.4 Berechnungen im M-Schritt

Im M-Schritt werden mit Hilfe der im E-Schritt geschätzten versteckten Parameter die neuen Modellparameter Θ^* und ω^* berechnet, die die Zielunktion $f_{\Theta^0,\omega^0,x}(\Theta,\omega)$ maximieren. Für ω^* kann eine allgemeine Formel aufgestellt werden, die modellunabhängig ist, wo hingegen Θ^* modellabhängig ist und daher erst in Kapitel 5 beschrieben wird.

Beginnend mit ω^* muss zuerst sichergestellt werden, dass durch das Maximieren stets $\sum_{j=1}^{c} \omega_j = 1$ gilt, was mit Hilfe eines Lagrange-Multiplikators gelöst werden kann. Die neue Zielfunktion mit Lagrange-Multiplikator β sieht nun folgerndermaßen aus

$$\mathbb{L}(\omega) = \beta \left(-1 + \sum_{j=1}^{c} \omega \right) + \sum_{i=1}^{n} \sum_{j=1}^{c} Z_{i,j}^{0} \cdot \left[\log \omega_j + \log P_{\Theta_j}(x_i) \right]. \tag{3.10}$$

Nun wird die Funktion abgeleitet

$$\frac{\partial \mathbb{L}(\omega)}{\partial \omega} = \beta + \sum_{i=1}^{n} \frac{1}{\omega_j} Z_{i,j}^{0}, \tag{3.11}$$

die Ableitung gleich 0 gesetzt und nach ω_j aufgelöst

$$\omega_j = \frac{\sum_{i=1}^{n} Z_{i,j}^{0}}{-\beta}. \tag{3.12}$$

Zur Berechung von β wird die Nebenbedingung $\sum_{j=1}^{c} \omega_j = 1$ benutzt, indem die Gleichung um die Summe $\sum_{j=1}^{c}$ erweitert wird

$$1 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{c} Z_{i,j}^{0}}{-\beta}. \tag{3.13}$$

Da $\sum_{j=1}^{c} Z_{i,j}^{0} = 1$ (Summe aller Zugehörigkeiten eines Messwertes) ist, ist $\beta = -n$. Durch Einsetzen von β in (3.12) entsteht als Maximierer folgende Formel

$$\omega_j^* = \frac{1}{n} \sum_{i=1}^{n} Z_{i,j}^{0}. \tag{3.14}$$

Dies lässt sich als durchschnittliches Gewicht aller Messwerte zum Modell j verstehen.

Die Bestimmung der Parameter von Θ^* ist zwar modellabhängig, jedoch lässt sich ein allgemein geltender Ansatz für die Maximierung des Parameters l in Θ_j beschreiben

$$\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial \Theta_{j,l}} = \sum_{i=1}^{n} Z_{i,j}^0 \cdot \frac{\partial \left(\log P_{\Theta_j}(x_i)\right)}{\partial \Theta_{j,l}} = 0. \tag{3.15}$$

Kapitel 4

Modell

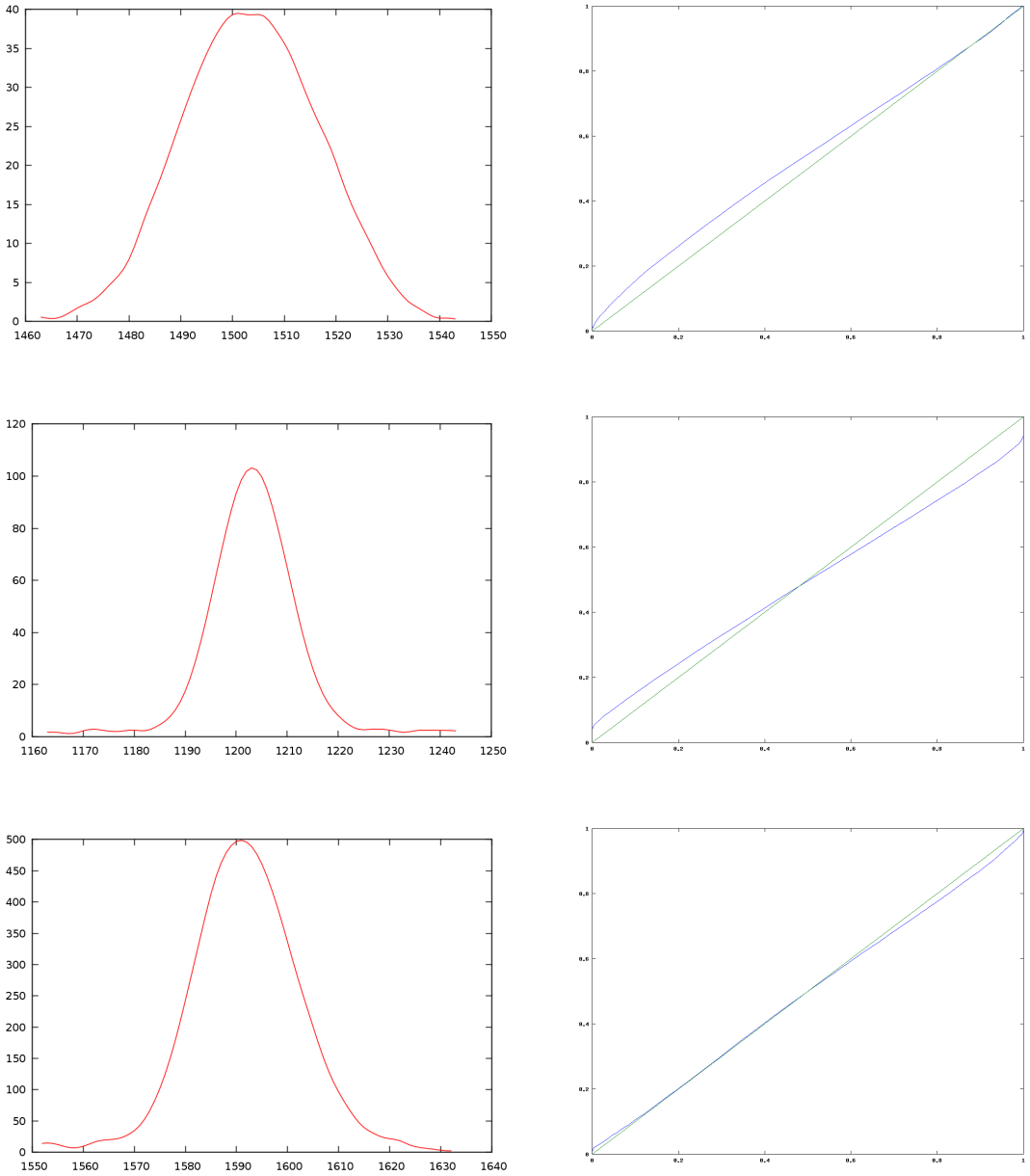
Das gesuchte Modell ist eine Abstrahierung der Peaks in einer IMS-Messung. Genau wie eine Messung ist das Modell eine $\mathbb{R}^2 \to \mathbb{R}$ Funktion, die als Eingabe eine Retentionszeit r und eine Driftzeit t hat und als Ausgabe eine Konzentration wiedergibt. Im Idealfall ist das Modell eine zweidimensionale Verteilungsfunktion, was notwendig zur korrekten Rechnung mit dem EM-Algorithmus ist und hinterher den Vorteil hat, dass das Volumen unter der Verteilungsfläche den Wert 1 hat.

4.1 Erzeugen eines Modells

Die Anzahl der für das Modell verwendeten Parameter hängt natürlich von den Funktionen ab, die für das Modell verwendet werden. Um geeignete Funktionen zu finden, wird zuerst von einigen Peaks ein horizontaler Querschnitt in Höhe des Modalwerts durchgeführt. Abbildung 4.1.a zeigt verschiedene dieser Querschnitte. Die Form der Kurven erinnert sehr stark an eine gaußsche Normalverteilung. Zur Überprüfung wird ein Quantil-Quantil-Plot eingesetzt. Ein QQ-Plot ist ein statistisches Verfahren zum Vergleichen von zwei Verteilungen. Dabei werden von beiden Verteilungsfunktionen Wertepaare gebildet und in ein Koordinatensystem übertragen. Ergeben die eingetragenen Punkte eine Gerade, kann angenommen werden, dass die beiden Verteilungen gleichverteilt sind. Hierzu wird ein QQ-Plot der Verteilungsfunktion der Normalverteilung und der Daten gebiltet. Abbildung 4.1.b zeigt, dass sich die Normalverteilung eignet, da überwiegend Geraden entstehen.

Der horizontale Schnitt verläuft in Richtung der Driftzeit und hat somit als Eingabeparameter die Driftzeit t. Die gaußsche Normalverteilung (entnommen aus [9]), hat zwei Parameter σ (Standardabweichung) und μ_t (Erwartungswert, der zugleich auch als Verschiebung des Modells in der Driftzeit-Achse verstanden werden kann), ist eine Wahrscheinlichkeitsdichte mit $f : \mathbb{R} \to \mathbb{R}_{>0}$ und $\int_{-\infty}^{\infty} f(t) = 1$ und

$$p(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t-\mu_t}{\sigma}\right)^2\right). \tag{4.1}$$



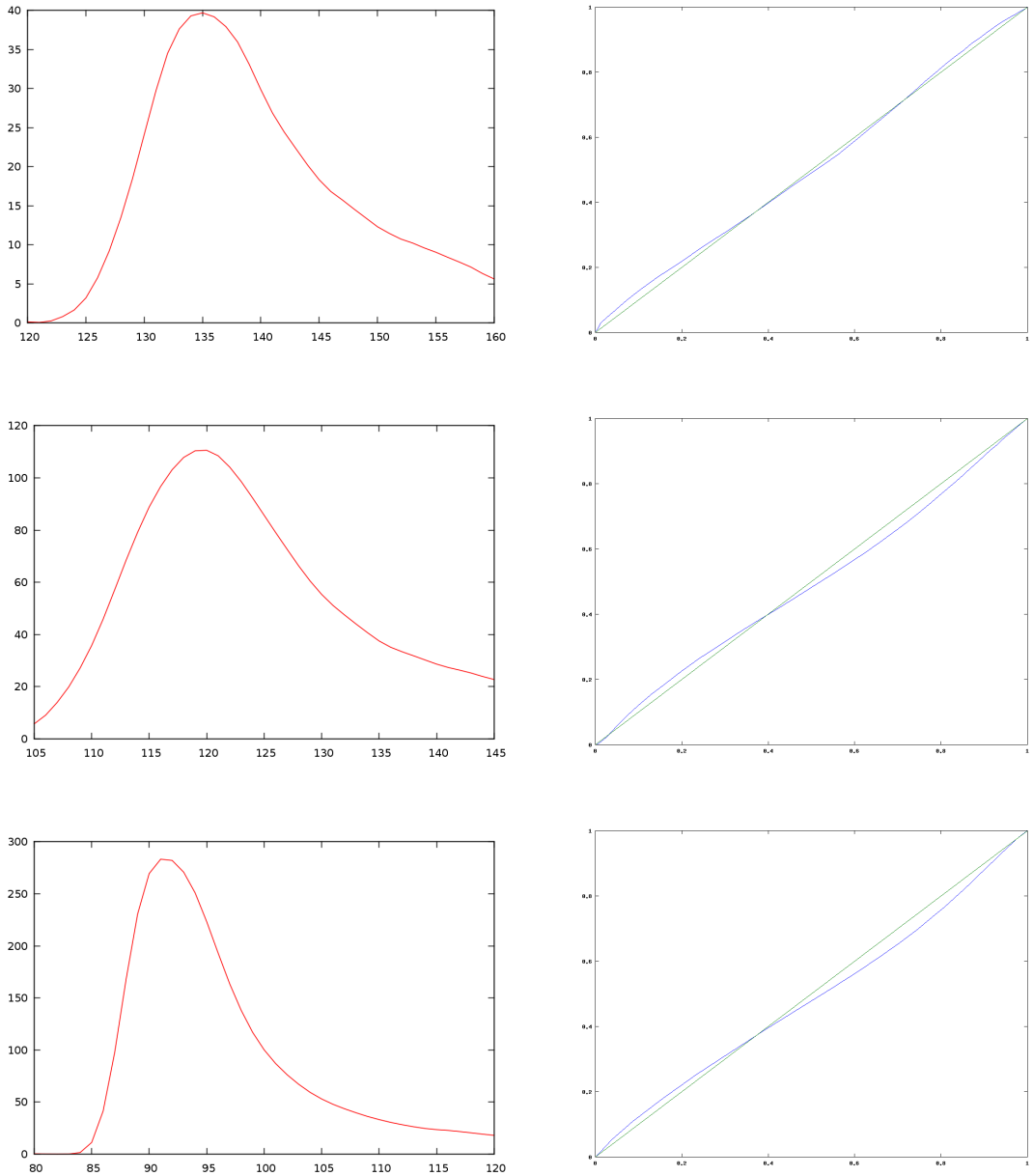
(a) Daten (b) QQ-Plot

Abbildung 4.1: Querschnitt der Messdaten in Driftzeit t und Quantil-Quantil-Plot; x-Achse: Datenquantil, y-Achse: Normalverteilungsquantil

Nun wird der vertikale Querschnitt der Peaks betrachtet. In Abbildung 4.2.a ist sofort zu erkennen, dass die Kurven nicht symmetrisch sind, dennoch alle Kurven ein bestimmtes Muster aufweisen. Die Kurven nehmen mit zunehmender Retentionszeit sehr schnell zu und ab dem Modalwert langsamer ab. Beachtlich ist hierbei, dass das Abnehmen der Kurven unterschiedlich lange dauert, von einer sehr langsamen Abnahme, bis hin zu einer sehr schnellen Abnahme. Hierbei kommen verschiedene Verteilungen in Frage:

- Chi-Quadrat-Verteilung

- Erlang-Verteilung
- F-Verteilung
- Gammaverteilung
- Inverse-Normalverteilung



(a) Daten (b) QQ-Plot

Abbildung 4.2: Querschnitt der Messdaten in Retentionszeit r und Quantil-Quantil-Plot; x-Achse: Datenquantil, y-Achse: Inverse-Normalverteilungsquantil

In verschiedenen QQ-Plots mit den vorgestellten Verteilungen erwies sich die inverse Normalverteilung als ideale Verteilung, da mit ihr in den QQ-Plots in weitesten

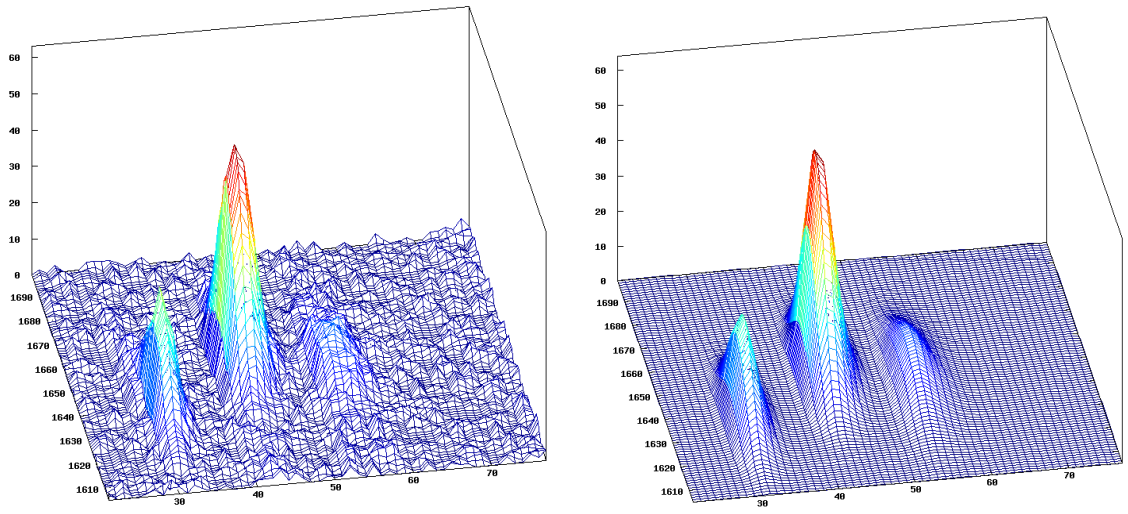
Teilen gerade Kurven erzeugt wurden. Aus diesem Grund wurde die inverse Normalverteilung ausgewählt.
Die inverse Normalverteilung ist eine Wahrscheinlichkeitsverteilung über einem halbseitig unendlichen Intervall $(0, \infty]$ und hat in ihrer Grundform zwei Parameter: λ (Ereignisrate) und α (Mittelwert). Da hierbei jedoch alle Verteilungskurven am Ursprung beginnen, wurde noch ein zusätzlicher dritter Parameter μ_r eingeführt, welcher eine Verschiebung des Modells in der Retentionszeit-Achse ermöglicht. Als Eingabeparameter wird eine Retentionszeit r eingesetzt. Die Verteilung sieht nun wie folgt aus [6]

$$p(r) = \begin{cases} \left(\frac{\lambda}{2\pi(r-\mu_r)^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda((r-\mu_r)-\alpha)^2}{2\alpha^2(r-\mu_r)}\right) & \text{für } r > \mu_r, \\ 0 & \text{sonst} \end{cases}. \tag{4.2}$$

Um nun ein zweidimensionales Modell zu bekommen, müssen die Normalverteilung und die inverse Normalverteilung miteinander multipliziert werden. Das endgültige Modell ist nun von r und t abängig, hat fünf Parameter $\Theta = (\sigma, \mu_t, \lambda, \alpha, \mu_r)$ und die Funktion

$$P_{\Theta_j}(r,t) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t-\mu_t}{\sigma}\right)^2\right) \cdot \left(\frac{\lambda}{2\pi(r-\mu_r)^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda((r-\mu_r)-\alpha)^2}{2\alpha^2(r-\mu_r)}\right) & \text{für } r > \mu_r, \\ 0 & \text{sonst} \end{cases}. \tag{4.3}$$

Im Anhang B sind die Ergebnisse sechs weiterer analysierter Peaks zu finden. In der folgenden Abbildung 4.3 ist links im Bild die reale Messung und rechts ein Modell mit geschätzten Parametern zu sehen. Es ist zu erkennen, dass das Modell den realen Peaks sehr ähnlich sieht. Eine genauere Auswertung wird in Kapitel 5 durchgeführt.



(a) Reale Peaks (b) Modell mit geschätzten Parametern

Abbildung 4.3: Vergleich: reale Messung und Modell

4.2 Erzeugen einer Modellmessung

Zur Verifikation des Modells und des anschließend modifizierten EM-Algorithmus ist es vorteilhaft, auf zumindest idealen Daten eine korrekte Merkmalsextraktion durchführen zu können. Zu diesem Zweck wurde ein Modellerzeugungsprogramm implementiert, das sowohl einen perfekten Datensatz erzeugen kann, als auch einen Datensatz mit Ionenkonkurrenz, bei dem "Einschnürungen" auftreten. Diese Datensätze waren ideal um u.a. die verschiedenen Vorverarbeitungsfilter implementieren und testen zu können. Auf der zur Arbeit beigefügten CD ist sowohl der Quellcode unter *src/* als auch eine Binary-Datei für Linux unter *bin/* gespeichert.

Für die Modellmessung mit $R \times T$ Konzentrationen wird sowohl das Modell des Peaks als auch ein Modell des RIPs und des Hintergrundrauschens gebraucht. Das Hintergrundrauschen hat das einfachste Modell, da es sich hierbei lediglich um eine Gleichverteilung mit einer Dichte von $\frac{1}{R \cdot T}$ handelt. Der RIP wird als ein vertikaler Streifen modelliert, der im horizontalen Querschnitt eine Normalverteilung besitzt. Da das Volumen der Peakmodelle in Relation zu der Summe aller Konzentrationen steht, wurde entschieden, dass nicht das Volumen, sondern die Konzentrationssumme in das Programm eingegeben wird. Alle Modelle erhalten ein Gewicht ω, mit der sie zum Gesamtmodell gehören, die Summe aller Gewichte muss 1 ergeben. Für eine Beispiel-Modellmessung mir $R = 200$ und $T = 350$ und der Summe aller Konzentrationen $u = 100000$ wurden folgende Modelle eingesetzt:

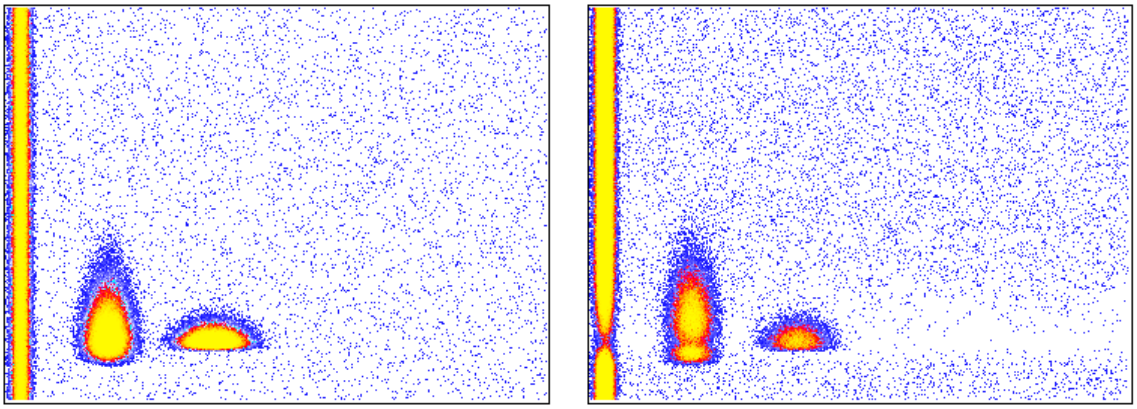
- $P_1 : \sigma = 6.84, \quad \mu_t = 68, \quad \lambda = 30.63, \quad \alpha = 192.30, \quad \mu_r = 15, \quad \omega = 0.2475$
- $P_1 : \sigma = 9.88, \quad \mu_t = 138, \quad \lambda = 11.59, \quad \alpha = 100.22, \quad \mu_r = 28, \quad \omega = 0.18$
- $P_{RIP} : \sigma = 3, \quad \mu_t = 10, \quad \omega = 0.55$
- $P_{HG} : \frac{1}{R \cdot T}, \quad \omega = 0.0225$

Im ersten Schritt werden eine $R \times T$ Matrix W für die Wahrscheinlichkeiten, eine $R \times T$ Matrix K für die Konzentrationen und ein Array A der Länge $n = R \cdot T + 1$ für die aufsummierten Wahrscheinlichkeiten erzeugt. W, K und A werden mit 0 vorinitialisiert. Nun werden für alle Punkte in W die Wahrscheinlichkeiten aus allen Modellen, multipliziert mit ihrer Dichte, zusammenaddiert:

$$W_{r,t} = \sum_{j=1}^{j} \omega_j \cdot P_{\Theta_j}(r,t)_j \qquad \forall r \in R, t \in T \tag{4.4}$$

Obwohl es sich um ein perfektes Modell handeln soll, sollen die Konzentrationen randomisiert entstehen. Zu diesem Zweck wird A nun mit Werten befüllt. Es erhält die erste Stelle im Array den Wert 0 und alle nachfolgenden Stellen den Wert ihres Vorgängers plus der aktuellen Wahrscheinlichkeit aus W, wobei hierbei die Matrix wie ein Array behandelt werden muss. Der letzte Wert aus A muss demzufolge den Wert 1 haben und alle anderen Werte $i \in n$ im Array $0 \leq A_i \leq 1$.

Nun wird “gelost”: Es wird eine randomisierte Zahl zwischen 0 und 1 gewählt. Anschließend wird mittels binärer Suche ermittelt, an welcher Position in A sich diese Zahl befindet. Ist die korrekte Position ermittelt, muss diese in die Werte r und t umgerechnet werden. Zum Schluss wird in der Konzentrationsmatrix an der Stelle $K_{r,t}$ der Wert um 1 inkrementiert. Dieser Vorgang wird insgesamt u mal wiederholt. So entsteht eine randomisierte Messung mit idealen Peaks, einem idealen RIP und Hintergrundrauschen, gut in Abbildung 4.4.a zu erkennen.
Soll Ionenkonkurrenz simuliert werden, muss das “Losen” zeilenweise erfolgen. Dabei muss u der Summe aller Konzentrationen in einer Zeile entsprechen und A die Größe $n = T + 1$ haben. Das Resultat lässt sich in Abbildung 4.4.b betrachten.



(a) Perfekte Modellmessung (b) Modellmessung mit Ionenkonkurrenz

Abbildung 4.4: Simulierte IMS-Messungen

Kapitel 5

Modifikation des EM-Algorithmus

5.1 Maximierung der Parameter

Der EM-Algorithmus wurde bereits vorgestellt und ein passendes Modell für die Modellierung der Peaks wurde ebenfalls eingeführt. Nun gilt es mit Hilfe des Modells die passenden Maximierer für die Parameter aus Θ zu bestimmen, die die Zielfunktion $f_{\Theta^0,\omega^0,x}(\Theta,\omega)$ maximieren, wie von Bilmes für die Normalverteilung [2] und von Cheng und Amin für die inverse Normalverteilung [5] beschrieben. Der allgemeine Ansatz zur Berechnung der Parameter sieht nach wie vor folgendermaßen aus

$$\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial \Theta_{j,l}} = \sum_{i=1}^{n} Z_{i,j}^0 \cdot \frac{\partial \left(\log P_{\Theta_j}(x_i)\right)}{\partial \Theta_{j,l}} = 0. \tag{5.1}$$

Für die Rechnung wird der Logarithmus des Modells berechnet

$$\log P_{\Theta_j}(r,t) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \cdot \left(-\frac{1}{2}\left(\frac{t-\mu_t}{\sigma}\right)^2\right) + \frac{1}{2}\log\left(\frac{\lambda}{2\pi(r-\mu_r)^3}\right) \cdot \left(-\frac{\lambda((r-\mu_r)-\alpha)^2}{2\alpha^2(r-\mu_r)}\right). \tag{5.2}$$

Im letzten Schritt vor dem Maximieren der Parameter muss noch sichergestellt werden, dass die Konzentrationen bei der Maximierung mitberücksichtigt werden. In den Ableitsungsfunktion (5.1) wird jeder Punkt so behandelt, als habe er die Konzentration 1. Es wird also in die Variable k_i eingeführt, die das Vorkommen (Konzentration) von Punkt i in der Messung beschreibt.

Berechnung von σ^*
Es gilt nun den ersten Parameter aus Θ zu maximieren. Hierzu wird die Verteilungsfunktion des Modells in (5.1) eingesetzt, abgeleitet, gleich 0 gesetzt und nach σ aufgelöst. In diesen Rechnungen wird der Modellindex j bei den Parametern bis auf $Z_{i,j}^0$ weggelassen

$$\begin{aligned}
\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial\sigma} &= \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(-\frac{1}{\sigma} + \frac{(t_i-\mu_t)^2}{\sigma^3}\right) \\
0 &= -\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \frac{1}{\sigma} + \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \frac{(t_i-\mu_t)^2}{\sigma^3} \\
\frac{1}{\sigma} \cdot \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i &= \frac{1}{\sigma^3} \cdot \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot (t_i-\mu_t)^2 \\
\sigma^2 &= \frac{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot (t_i-\mu_t)^2}{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i} \\
\sigma^* &= \sqrt{\frac{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot (t_i-\mu_t)^2}{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i}}. \qquad (5.3)
\end{aligned}$$

Berechnung von μ_t^*
Nun wird μ_t maximiert, indem die Verteilungsfunktion des Modells in (5.1) eingesetzt, abgeleitet, gleich 0 gesetzt und nach μ_t aufgelöst wird

$$\begin{aligned}
\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial\mu_t} &= \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{1}{\sigma^2}(t_i-\mu_t)\right) \\
0 &= \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot ((t_i-\mu_t)) \\
0 &= \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot t_i - \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \mu_t \\
\mu_t^* &= \frac{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot t_i}{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i}. \qquad (5.4)
\end{aligned}$$

Berechnung von α^*

$$\begin{aligned}
\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial\alpha} &= \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(-\frac{\lambda}{2(r-\mu_r)} \cdot \frac{-2\left((r_i-\mu_r)-\alpha\right)\alpha^2 - 2\alpha\left((r_i-\mu_r)-\alpha\right)^2}{\alpha^4} \right) \\
0 &= \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \frac{\lambda}{2(r-\mu_r)} \cdot \frac{2\alpha(r_i-\mu_r)(r_i-\mu_r-\alpha)}{\alpha^3} \\
0 &= \frac{\lambda}{\alpha^3} \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot (r_i-\mu_r-\alpha) \\
\alpha \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i &= \left(\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot r_i \right) - \left(\mu_r \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \right) \\
\alpha^* &= \frac{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot r_i}{\sum\limits_{i=1}^{n} Z_{i,j}^0 \cdot k_i} - \mu_r
\end{aligned} \tag{5.5}$$

Bei der Berechnung α^* gibt gibt es das Problem, dass in der Rechnung μ_r benutzt wird, aber de maximierte Wert μ_r^* verlangt wird. Für die Berechnung von μ_r^* wird jedoch α^* verwendet. Die Lösung des Problems ist zuerst α^* mit μ_r aus der vorherigen Iteration zu berechnen, anschließend μ_r^* zu berechnen und danach erneut α^* zu berechnen.

Berechnung von λ^*

Für die nachfolgende Berechnung wird α^* durch α ersetzt. Zur einfacheren Rechnung wird der Term $\frac{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot r_i}{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i}$ durch $\overline{r}$ substituiert wird:

$$\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial\lambda} = \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{1}{2\lambda} - \frac{((r_i-\mu_r)-\alpha)^2}{2\alpha^2(r_i-\mu_r)}\right)$$

$$0 = \frac{1}{2\lambda}\left(\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i\right) - \frac{1}{2}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \frac{(r_i-\mu_r)^2 - 2\alpha(r_i-\mu_r)+\alpha^2}{\alpha^2(r_i-\mu_r)}$$

$$\frac{1}{\lambda}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i = \sum_{i=1}^{n}\left(\frac{r_i-\mu_r-\alpha}{\alpha^2} - \frac{1}{\alpha} + \frac{1}{r_i-\mu_r}\right)$$

$$\frac{1}{\lambda}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i = \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{r_i-\mu_r-(\overline{r}-\mu_r)}{\alpha^2}\right) + \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(-\frac{1}{\alpha} + \frac{1}{r_i-\mu_r}\right)$$

$$\frac{1}{\lambda}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i = \frac{1}{\alpha^2}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot r_i - \frac{1}{\alpha^2}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \overline{r} + \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(-\frac{1}{\alpha} + \frac{1}{r_i-\mu_r}\right)$$

$$\frac{1}{\lambda}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i = \frac{1}{\alpha^2}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot r_i - \frac{1}{\alpha^2}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \frac{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot r_i}{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i} + \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{1}{r_i-\mu_r} - \frac{1}{\alpha}\right)$$

$$\frac{1}{\lambda}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i = \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{1}{r_i-\mu_r} - \frac{1}{\alpha}\right)$$

$$\lambda^* = \frac{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i}{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{1}{r_i-\mu_r} - \frac{1}{\alpha}\right)} \tag{5.6}$$

Berechnung von μ_r^*

$$\frac{\partial f_{\Theta^0,\omega^0,x}(\Theta,\omega)}{\partial\lambda} = \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{3}{2(r_i-\mu_r)} - \frac{-4\alpha^2\lambda(r_i-\mu_r-\mu_r)(r_i-\mu_r)+2\alpha^2\lambda(r_i-\mu_r-\alpha)^2}{4\alpha^2(r_i-\mu_r)}\right)$$

$$0 = \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{3}{2(r_i-\mu_r)} + \frac{\lambda(r_i-\mu_r-\alpha)(r_i-\mu_r+\alpha)}{2\alpha^2(r_i-\mu_r)^2}\right)$$

$$0 = \frac{1}{2}\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{3}{r_i-\mu_r} + \frac{\lambda(r_i-\mu_r)^2}{\alpha^2(r_i-\mu_r)^2} - \frac{\lambda\alpha^2}{\alpha^2(r_i-\mu_r)^2}\right)$$

$$0 = \sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{3}{\lambda(r_i-\mu_r)} + \frac{1}{\alpha^2} - \frac{1}{(r_i-\mu_r)^2}\right) \tag{5.7}$$

Da es an dieser Stelle nicht möglich ist (5.7) nach μ_r aufzulösen, wird das Newton-Verfahren eingesetzt, um μ_r zu berechnen. Das Newton-Verfahren ist ein Näherungsverfahren, wobei zur Bestimmung der Nullstelle einer Funktion an einem Ausgangspunkt die Nullstelle der Tangente als Näherung für die Nullstelle bestimmt wird.

Diese berechnete Näherung wird in der nächsten Iteration als neuer Ausgangspunkt verwendet. Der Vorgang wird so oft wiederholt, bis das Verfahren konvergiert (siehe [11]). Für das Verfahren mit der Näherungsfunktion $x_{neu} = x_{alt} - \frac{f(x_{alt})}{f'(x_{alt})}$ muss von (5.7) die Ableitung nach μ_r berechnet werden und in die Näherungsfunktion eingesetzt werden, sei n dabei der Iterationsschritt n

$$\mu_r^{n+1} = \mu_r^n - \frac{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{3}{\lambda(r_i-\mu_r^n)} + \frac{1}{\alpha^2} - \frac{1}{(r_i-\mu_r^n)^2}\right)}{\sum_{i=1}^{n} Z_{i,j}^0 \cdot k_i \cdot \left(\frac{3}{\lambda(r_i-\mu_r^n)^2} - \frac{1}{(r_i-\mu_r^n)^3}\right)}. \tag{5.8}$$

Berechnung des Volumens
Da das Volumen unter der Wahrscheinlichkeitsdichte 1 ist, muss als letzter Parameter ins Modell ein Volumen v eingefügt werden. Da v nicht in der Likelihood-Funktion enthalten ist, kann auch keine maximierende Funktion nach dem Prinzip der ersten fünf Parameter erstellt werden. Das Volumen eines Peaks entspricht aber dem Gewicht ω eines Modells in der Messung. Zur Bestimmung des Volumens muss lediglich die Summe aller Konzentrationen u mit dem Gewicht eines Modells multipliziert werden, es ergibt sich also

$$v = u * \omega_j. \tag{5.9}$$

5.2 Algorithmus

Im ersten Schritt müssen die Messwerte, bestehend aus x/y Koordinate und Konzentration, eingelesen werden. Des Weiteren muss die Anzahl und die grobe Position der zu parametrisierenden Peaks bekannt sein. Es wird eine Liste D erstellt, die sowohl die eingelesen Werte, als auch die Zugehörigkeiten $Z_{i,j}$ pro Messpunkt speichert. In einer zweiten Liste M werden die Parameter aller Modelle Θ und deren Gewicht im Gesamtmodell ω gespeichert. Zusätzlich kommt in M das Modell des Hintergrundrauschens als letztes Modell hinzu.

Für die Vorinitialisierung in D bekommen alle $Z_{i,j}$ pro Messwert eine gleichverteilte Wahrscheinlichkeit $\frac{1}{c}$. In M bekommen ebenfalls alle Werte ω ein gleichverteiltes Gewicht über alle Modelle. Für eine optimale Analyse ist zu beachten, dass die Startposition für einen Peak (μ_r, μ_t) mittig unterhalb des Peaks liegen sollte. Dies ist deshalb wichtig, da bei der inversen Normalverteilung die Konzentrationen, deren Retentionszeit $r \leq \mu_r$ erfüllt, auf 0 gesetzt werden und somit die Analyse u.U. nicht den ganzen Peak einnimmt und so die Modellparameter verfälscht werden. Da das Volumen während der EM-Phase nicht benutzt wird, braucht es nicht vorinitialisiert zu werden. Die restlichen drei Parameter können beliebig vorinitialisiert werden. Tests haben ergeben, dass sie mit einem vorinitialisierten Wert zwischen 1 und 5 die besten Ergebnisse erzielen.

Da die Vorverarbeitung der Messdaten keinen Einfluss nimmt auf die Vorinitialisierung, kann sie parallel dazu vorgenommen werden. Die Reihenfolge der Vorverarbeitungsfilter ist wie folgt festgelegt: Medianfilter - Feuchtekorrektur - Rekonstruktion der Ionen - Basislinienkorrekturfilter - Gaußglättungsfilter. Dadurch werden die besten Ergebnisse der Datenaufbereitung erzielt. Sind beide Vorgänge durchgeführt, beginnt der iterative EM-Prozess. Zuerst werden die aktuellen Parameter aller Modelle zwischengespeichert, um nach jedem Iterationsschritt zu überprüfen, ob die Werte konvergieren. Im E-Schritt werden die neuen versteckten Parameter $Z^0_{i,j}$ berechnet, wobei bei der Berechnung zu beachten ist, dass die Wahrscheinlichkeit des Hintergrundmodells nicht mit der Wahrscheinlichkeitsdichtefunktion des Peaks zu berechnen ist, sondern immer die Wahrscheinlichkeit $\frac{1}{n}$ hat.

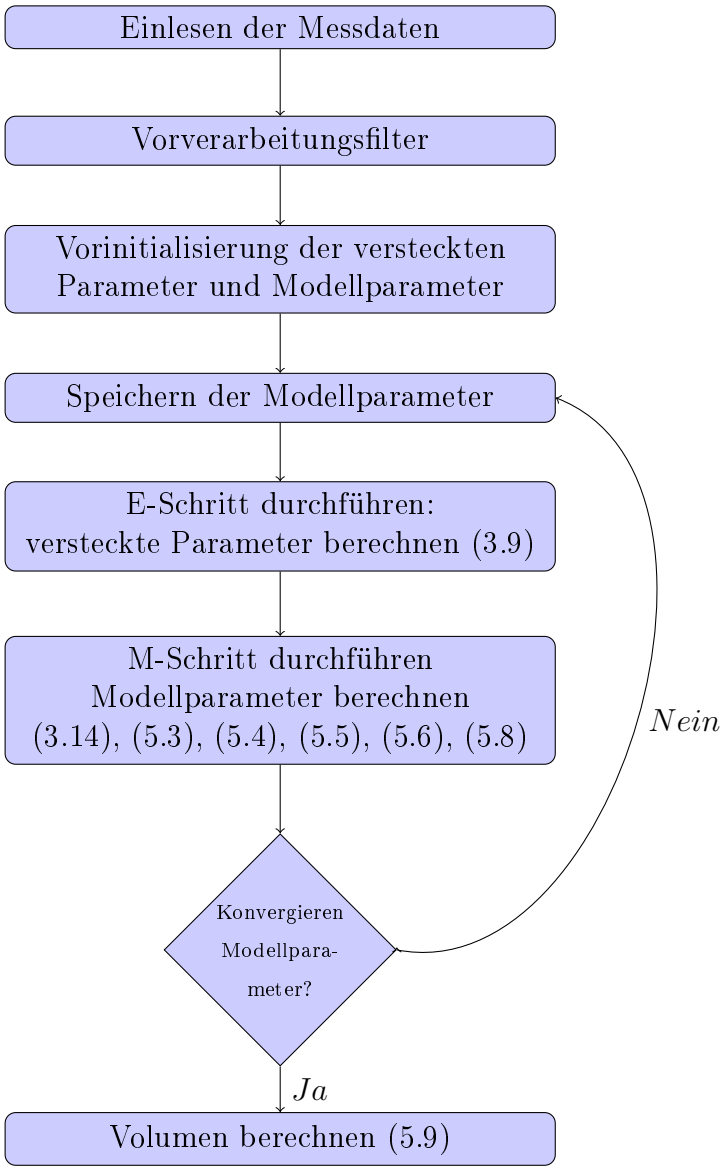
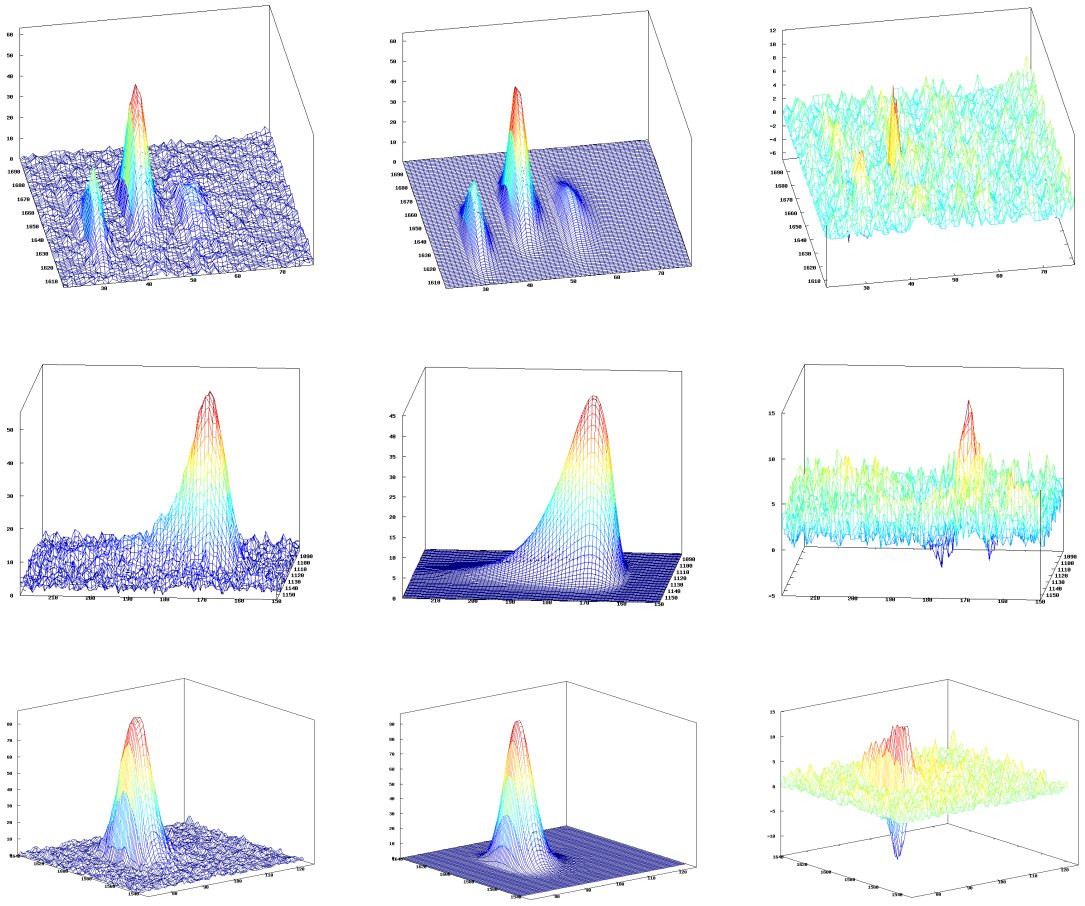


Abbildung 5.1: Ablauf des EM-Algorithmus

Im M-Schritt werden nun alle Modellparameter aller Modelle mit den Maximierer-funktionen berechnet. Da das Hintergrundmodell keine Θ Parameter hat, müssen sie auch nicht für den Hintergrund berechnet werden. Zum Ende einer Iteration wird die Konvergenz kontrolliert. Sei L die Anzahl der Parameter, l der Parameter l aus Θ und s ein Schwellenwert. Um die Konvergenz aller Parameter zu überprüfen, muss gelten

$$\sum_{l=1}^{L} \left(\frac{\Theta_l^0}{\Theta_l^*} - 1 \right)^2 \leq s. \tag{5.10}$$

Ist diese Bedingung erfüllt, bricht die EM-Phase ab. Üblicherweise sollte $s = 10^{-4}$ betragen. Je geringer der Schwellenwert wird, desto genauer werden die Modelle, aber desto länger dauert auch die Berechnung. Im letzten Schritt werden die Volumina der Modelle (mit Ausnahme des Hintergrundmodells) berechnet. Abbildung 5.1 veranschaulicht noch einmal den kompletten Vorgang.



(a) Reale Messung (b) EM-Modell (c) Differenz Messung/Modell

Abbildung 5.2: Vom EM-Algorithmus geschätzte Modelle. Es wird jeweils die Realmessung, das mit dem EM-Algorithmus entstandene Modell und eine Differenz zwischen Realmessung und Modell gezeigt.

Da die Parameter auf jeden Fall konvergieren, bricht die EM-Phase immer ab. Jedoch lässt es sich unmöglich vorhersagen, wie viele Iterationsschritte benötigt werden. Tests haben ergeben, dass bei der Anayse mit einem Modell üblicherweise 8 bis 12 Iterationen durchgeführt werden. Die Anzahl der Iterationen und die Laufzeit nehmen linear mit der Anzahl der Modelle zu. Abbildung 5.2 zeigt verschiedene Analysen.

5.3 Beispiel einer Analyse

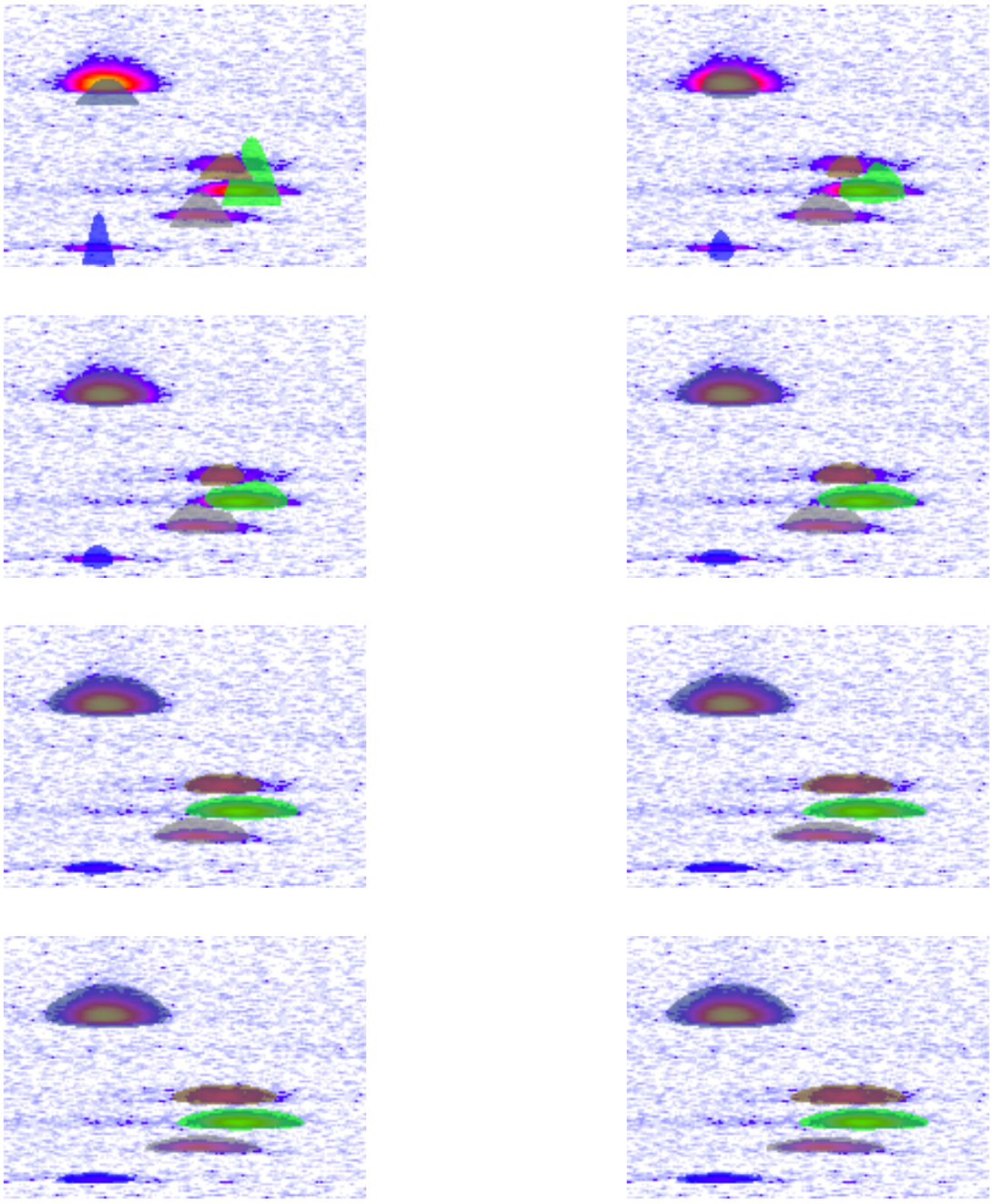


Abbildung 5.3: Maximierungsvorgang des EM-Algorithmus, gehört ein Punkt zu mind. 50% zu einem Modell, wird der Punkt farblich dem Modell zugehörend markiert.

In Abbildung 5.3 wird beispielhaft eine Analyse eines Ausschnitts einer IMS-Messung durchgeführt. Im ersten Bild ist zu erkennen, dass zuerst die Modelle, die durch die zufällig gewählten Parameter entstanden sind, überhaupt nicht zu den Peaks passen. Lediglich die Startposition der Modelle wurde mittig unterhalb der Peaks gewählt.

Zum Anfang ist sogar das grüne Modell so ungünstig gewählt, dass es sich über zwei Peaks erstreckt.

Mit jedem Iterationsschritt verbessern sich die Modelle. Es ist deutlich zu erkennen, dass die Modelle sich durch die verbessernden Parameter den Peaks deutlich anpassen. Durch die Fähigkeit mit gemischten Modellen umzugehen, wird das grüne Modell vom zweiten Peak verdrängt und betrachtet nur noch die Konzentrationen seines ersten Peaks.

Im letzten Bild ist ein lokales Optimum erreicht. Die Parameter konvergieren und der Algorithmus bricht ab. Es ist deutlich zu erkennen, dass die Modelle komplett ihre Peaks umhüllen. Alle übrigen Konzentrationen werden dementsprechend als Hintergrundrauschen klassifiziert.

5.4 Einheitliche Daten aus den Messungen

Es kommt vor, dass für eine wiederholte Messung eines Peaks, verschiedene Werte für die Parameter λ, α und μ_r berechnet werden. Das hängt damit zusammen, dass die inverse Normalverteilung die Eigenschaft hat, ähnliche Kurven zu erzeugen, die in der x-Achse verschoben sind. Durch das Einsetzen des μ_r-Parameters wird die Verschiebung kompensiert und es werden trotz verschiedener Werte zwei ähnliche Kurven erzeugt. In folgendem Beispiel wird dieses Phänomen verdeutlicht:

- $P_1 : \lambda = 10.93, \quad \alpha = 106.42, \quad \mu_r = 9.63$
- $P_2 : \lambda = 20.41, \quad \alpha = 786.08, \quad \mu_r = 0$

Trotz der unterschiedlichen Parameter sind beide Kurven in Abbildung 5.4 sehr ähnlich. Es müssen also Parameter gefunden werden, die trotz mehrerer Messungen eines Peaks immer ähnliche Werte annehmen. Da die beiden Kurven ähnlich sind, liegt sowohl deren Modalwert als auch deren Dichte am Modalwert sehr nahe beieinander. Da die Fläche der Verteilungen immer 1 ist, hängt das Volumen der Modelle nur von deren Gewichtung ab. Tests haben ergeben, dass bei mehrmaliger Analyse immer sehr ähnliche Gewichtungen berechnet werden.
Die Formel zur Berechnung der Position des Modalwerts bei der inversen Normalverteilung ist folgende (vgl. [6])

$$r_{Mod} = \alpha \left[\left(1 + \frac{9\alpha^2}{4\lambda^2} \right)^{\frac{1}{2}} - \frac{3\alpha}{2\lambda} \right] + \mu_r. \tag{5.11}$$

Im Beispiel beträgt bei P_1 die Position des Modalwertes $r_1 = 19.01$. Bei P_2 liegt die Position bei $r_2 = 19.62$. Die Positionen liegen also nur um eine halbe Einheit auseinander, wo hingegen deren Parameter μ_r um ganze 9.63 Einheiten auseinander liegen. Zur anschließlenden Auswertung der Parameter aus der Analyse sollte die Position des Modalwerts der Peaks betrachtet werden.
Da auch die Dichte am Modalwert ähnlich ist, sollte sie bei einer anschließenden

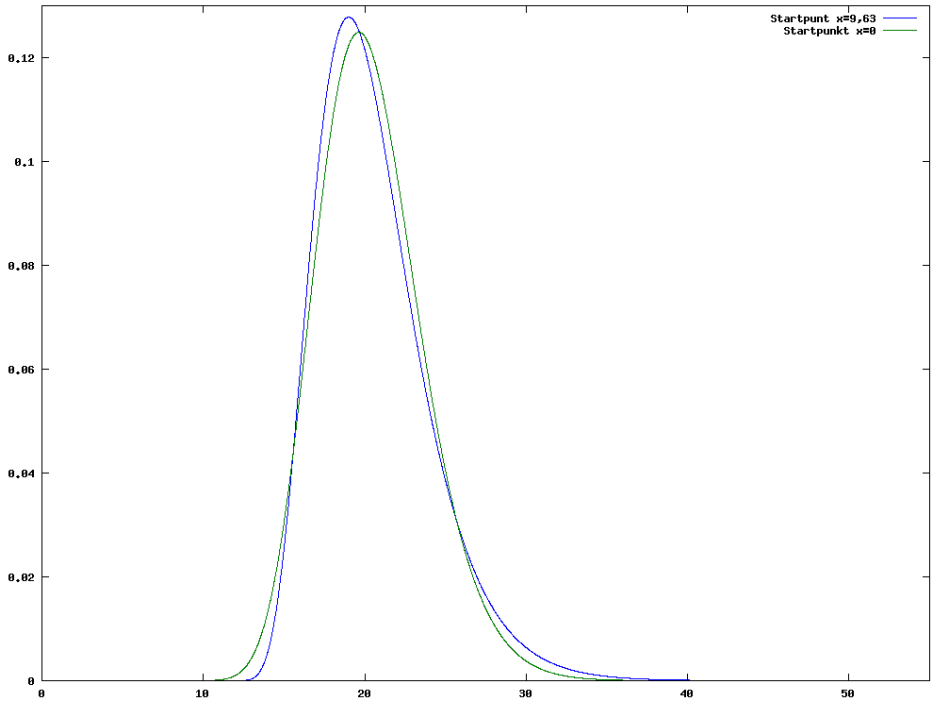


Abbildung 5.4: Ähnliche Kurven trotz unterschiedlicher Parameter

Auswertung der Parameter betrachtet werden. Da der Mittelwert α der inversen Normalverteilung um den Wert μ_r verschoben ist, wird ebenfalls der unkorrigierte Mittelwert $e = \alpha + \mu_r$ betrachtet.

5.5 Beispiel für einheitliche Parameter

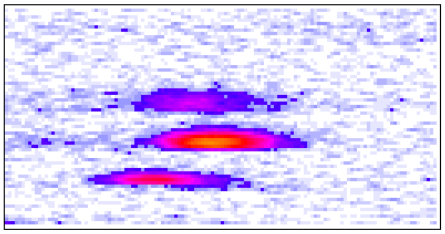


Abbildung 5.5: Beispielhafte Daten für mehrere Analysen für einheitliche Parameter, der Ausschnitt ist im Intervall von $r = [1600, 1700]$ und $t = [20, 70]$

Insgesamt bieten sich vier Parameter $(r_{Mod}, p(r_{Mod}), e, v)$ als einheitliche Parameter an. Anhand der Daten von Abbildung 5.5 wurden mehrere Analysen durchgeführt. Tabelle 5.5 zeigt bei verschiedenen Startwerten der Analyse, wie unterschiedlich die Modellparameter sind. Hingegen sind die im vorherigen Abschnitt erwähnten Parameter sehr einheitlich.

Modelle	Startwerte		Modellparameter			Einheitliche Parameter			
	μ_r	μ_t	α	λ	μ_r	r_{Mod}	$p(r_{Mod})$	e	v
Modell 1	22	1636	4.67	56.25	26.53	30.65	0.33	31.20	2566.31
	24	1638	7.48	259.84	23.68	30.85	0.32	31.16	2552.95
	16	1635	16.74	3501.24	14.38	31.00	0.34	31.12	2267.82
	25	1636	6.36	161.72	24.81	30.82	0.32	31.18	2530.59
Modell 2	36	1656	7.08	120.53	35.77	42.26	0.25	42.85	6804.55
	32	1658	11.71	588.70	31.11	42.27	0.25	42.82	6748.53
	35	1650	6.36	93.77	36.43	42.19	0.26	42.81	6374.88
	39	1649	4.69	28.81	38.17	41.85	0.25	42.86	6799.81
Modell 3	47	1656	8.97	150.50	45.54	53.75	0.19	54.51	2310.60
	48	1652	7.23	75.08	47.28	53.54	0.18	54.51	2246.62
	46	1648	8.46	162.14	45.90	53.72	0.21	54.36	2427.29
	50	1641	5.62	32.19	48.91	53.25	0.20	54.54	2184.46

Tabelle 5.1: Vergleich der Modellparameter mit den einheitlichen Paramertern

Kapitel 6

Zusammenfassung

Die Ionen-Mobilitäts-Spektrometrie ist, in Bezug auf die Atemluft, ein junges Gebiet, in dem sicherlich noch viel Forschung betrieben werden kann. IMS findet schon Einsatz in Bereichen, wie Drogen/Sprengstoffkontrolle in Flughäfen oder Luftanalyse in Städten, wo das Spekrum der zu analysierenden Gase beschränkt ist und somit die Merkmalsextraktion nicht allzu komplex ist. Doch die IMS hat Potenzial ihr Einsatzgebiet auf alle Bereiche, in denen die Beschaffenheit von Gasen eine Rolle spielt, zu erweitern.

Stationäre Geräte könnten in der Medizin zur Analyse der Ausatemluft oder in der Industrie als Sicherheitsmodul beim Arbeiten mit gefährlichen Gasen eingesetzt werden. Mit mobilen Geräten gäbe es die Möglichkeit sämtliche Gebiete einzunehmen, in denen noch Spürhunde eingesetzt werden, wie z.B. bei mobilen Kontrollen oder bei der Suche nach Lawinenopfern.

6.1 Ergebnisse

Diese Diplomarbeit beschreibt ein Verfahren zur präzisen Merkmalsextraktion bei IMS-Messungen. Die in den Messdaten vorhandenen Peaks werden mit Hilfe des EM-Algorithmus analysiert und in ein passendes Modell überführt. Dabei werden die Parameter des Modells so lange angepasst, bis ein Modell einen Peak optimal beschreibt. Da diese Methode zum ersten Mal zur Analyse von IMS-Messungen eingesetzt wurde, eröffnet sie völlig neue Möglichkeiten. Zum ersten Mal ist es möglich präzise das Volumen eines Peaks zu bestimmen. Frühere Verfahren zur Volumenschätzung haben lediglich den Modalwert eines Peaks ermittelt und mit dessen Konzentration in Relation zur räumlichen Ausprägung des Peaks ein Volumen bestimmt.
Das Modell kann darüber hinaus alle (hauptsächlich in den Messungen vorkommenden) symmetrischen Peaks beschreiben. Es ist möglich sogar große Peaks mit mehreren Tausend Messpunkten auf nur sechs Parameter zu reduzieren. Eine komplette Messung mit einer Million Messpunkten kann somit (natürlich abhängig von der Anzahl der Peaks) auf unter 100 Parameter reduziert werden, was eine Reduktion der Daten um den Faktor 10000 entspricht.

Die Vorteile dieser Methode sind die hohe Reduktion der Daten, das erstmalige Berechnen des Volumens eines Peaks und der Umgang mit sich überlappenden Peaks. Natürlich hat diese Methode auch Nachteile. Zum einen muss schon vor der Analyse die Anzahl und die Position der Peaks bekannt sein, um die Parameter zur Analyse optimal vorzuinitialisieren. Zum anderen muss ein hohes Maß an Vorverarbeitung geleistet werden, bei dem das Hintergrundrauschen auf ein homogenes Niveau über die ganze Messung gesenkt wird, da es sonst vorkommt, dass das Verfahren hohe Stellen in der Nähe von Peaks als Bestandteil des Peaks erkennt.

6.2 Merkmalsextraktion mit knappen Ressourcen

Bei der Realisierung eines mobilen IMS-Messgerätes müssen viele Einschränkungen berücksichtigt werden. Ein mobiles IMS-Messgerät sollte zum Sparen von Energie wenig Rechenleistung benötigen und einen geringen Speicher haben. Darüber hinaus sollte der Analysevorgang noch während einer Messung stattfinden, was eine Analyse mit unvollständigen Messdaten vorraussetzt. Tests mit dem in der Arbeit vorgestellten Verfahren haben ergeben, dass unvollständige Peaks nicht korrekt erkannt und somit falsch beschrieben werden. Es gibt jedoch Verfahren zur Maximum-Likelihood-Parameterschätzung mit unvollständigen Daten [8], was aber im Rahmen der Diplomarbeit nicht berücksichtigt wurde. Das in der Arbeit vorgestellte Verfahren arbeitet auf einem vollständigen Datensatz mit einer bekannter Peakanzahl und mit bekannten Peakpositionen.

Ein Verfahren, das zur Detektion von Peaks benutzt werden kann ist, das Watershed Verfahren [1], bei dem ein abstrakter Wasserspiegel allmälig gesenkt wird und die so entstehenden “Inseln” als Peaks detektiert werden. Somit kann dann auch sofort die Position des Peaks zur Analyse mitdetektiert werden.

Anhang A

IMS-Analyseprogramm

Für die Zwecke der Diplomarbeit wurde ein Programm zur Visualisierung und Analyse von IMS-Messungen entwickelt. Es wurde in der Programmiersprache C/C++ unter einer Linux Distribution implementiert. Für die Verwendung von C/C++ spricht, dass es die am weitesten verbreitete Programmiersprache ist und es die gleichen Bibliotheken für alle gängigen Plattformen gibt. Somit lässt sich das Programm auch problemlos unter Microsoft Windows kompilieren. Auf der zur Arbeit beigefügten CD ist sowohl der Quellcode unter *src/* als auch eine Binary-Datei für Linux unter *bin/* gespeichert.

A.1 Programmstart

Da die Ausgabe der Modellparameter auf einer Konsole erfolgt, ist es zwingend notwendig das Programm mittels einer Konsole zu starten. Die Messdaten einer IMS-Messung werden in einem speziellen Format in einer *.csv* Datei gespeichert (vgl. Boedecker [4]). Als notwendiger Parameter wird zusätzlich der Pfad zur *.csv* Datei angegeben.

```
./anayse path/to/file.csv
```

Es öffnet sich ein Fenster, wie in Abbildung A.1 dargestellt. Der Benutzer hat nun die Möglichkeit mit Hilfe der Pfeil-Tasten das Messbild horizontal oder vertikal zu verschieben. Mit der "+", bzw. "-" Taste ist es möglich in das Messbild rein, bzw. raus zu zoomen. Das Programm hat zwei Zustände, den Datenaufnahmezustand und den Bearbeitungszustand. Zuerst kann der Benutzer einen Ausschnitt aus dem kompletten Messbild durch Drücken und Halten der rechten Maustaste ausschneiden. Durch Klicken auf das Messbild können anschließend die Startwerte der einzelnen Modelle festgelegt werden.

Durch Drücken der Leerstaste wird die Analyse gestartet, das Programm wechselt in den Bearbeitungszustand. Es ist dem Benutzer möglich den Verlauf der Analyse zu verfolgen. Durch wiederholtes Drücken der Leertaste, wird die Analyse neu gestartet. Mit der "ESC" Taste kann der Anaylsevorgang abgebrochen werden. Gehört ein Datenpunkt zu mind. 50% zu einem Modell wird es farblich zum Modell markiert.

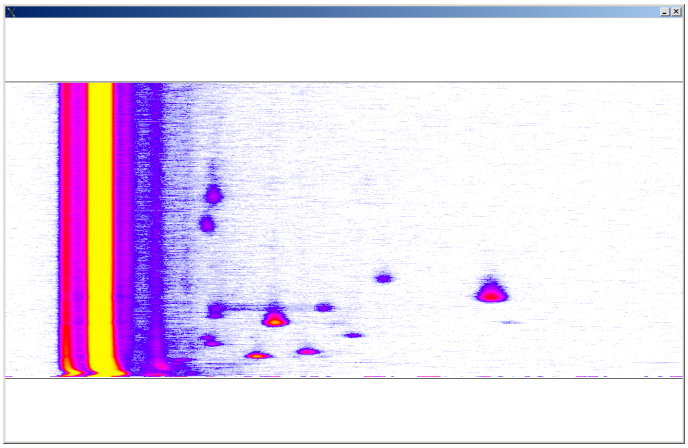


Abbildung A.1: Hauptfenster des IMS-Analyseprogramms

Diese Markierung wird halbtransparent über das Messbild gelegt.

Ist die Analyse abgeschlossen, werden die Modelle von den Konzentrationen der Daten im Messbild abgezogen. Die Modellparameter erscheinen nun in der Konsole. Durch Drücken der "s" Taste können die Markierungen der Modelle aus- bzw. eingeblendet werden. Der Benutzer hat jetzt die Möglichkeit das Programm mittels der "ESC" Taste zu beenden, oder mittels der "r" Taste wieder in den Datenaufnahmezusatend zu wechseln. Alle gespeicherten Modelle werden gelöscht und die Ansicht des Ausschnitts wird aufgehoben.

A.2 Optionen

Die in der Diplomarbeit vorgestellten Vorverarbeitungsfilter wurden ebenfalls implemeniert. Diese sind optional aktivierbar. Aus Gründen des Speicherbedarfs, können diese Filter nicht beliebig ein- und ausgeschaltet werden, sondern müssen beim Aufruf des Programms als optionale Parameter eingegeben werden. Folgende Optionen sind möglich:

-z	Analysesequenzen werden in Bildern im Ordner *./Bilder* gespeichert
-m	aktiviert den Medianfilter
-f	aktiviert die Feuchtekorrektur
-r	aktiviert die Rekonstruktion der Ionen
-b	aktiviert den Basislinienkorrekturfilter
-g	aktiviert den Gaußglättungsfilter

Die Reihenfolge der Filter ist fest implementiert und kann nicht verändert werden, da sonst die Daten nicht optimal verarbeitet werden. Soll ein Filter im bereits laufendem Programm eingesetzt werden, muss das Programm geschlossen und mit den entsprechenden Parametern neu gestartet werden.

Anhang B

Analyse verschiedener Peaks

Die nachfolgenden Analysen der Peaks zeigen pro Seite den untersuchten Peak, die geschätzten Parameter, einen Querschnitt in der Driftzeit und den dazu gehörigen QQ-Plot mit einer Normalverteilung, sowie einen Querschnitt in der Retentionszeit und den dazu gehörigen QQ-Plot mit einer inversen Normalverteilung. Die ersten beiden Beispiele zeigen sehr gute Ergebnisse und bestätigen, dass das Modell zu den Peaks passt. Die zweiten beiden Beipiele zeigen ein mittelmäßiges Ergebnis und die letzten beiden Beispiele zeigen eine unzureichende Analyse. Ein Dokument mit 57 Analysen ist auf der zur Arbeit beigefügten CD in der Datei *Peak-Analysen.pdf* zu finden.

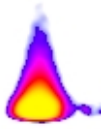
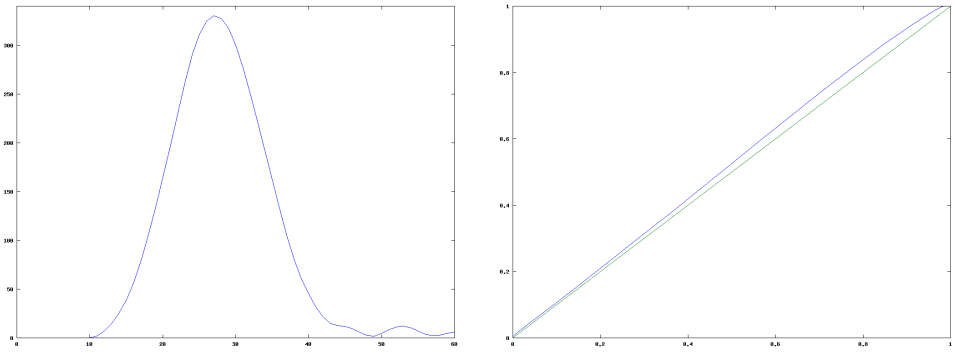


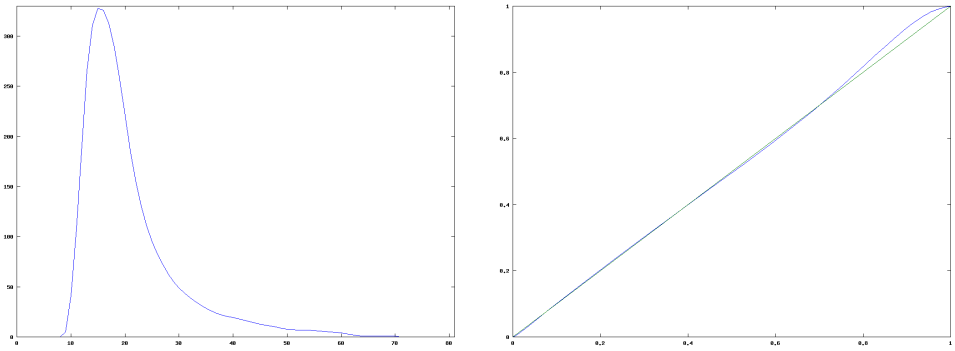
Abbildung B.1: Datenpeak einer IMS-Messung

$$\sigma = 5.9284, \quad \mu_t = 28.2498, \quad \lambda = 38.5360, \quad \alpha = 12.7247, \quad \mu_r = 7.2334$$



(a) Querschnitt in Dirftzeit t (b) QQ-Plot

Abbildung B.2: QQ-Plot zur Überprüfung der Daten mit der Normalverteilung



(a) Querschnitt in Retentionszeit r (b) QQ-Plot

Abbildung B.3: QQ-Plot zur Überprüfung der Daten mit der inversen Normalverteilung

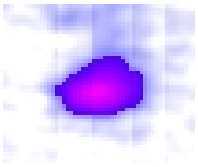
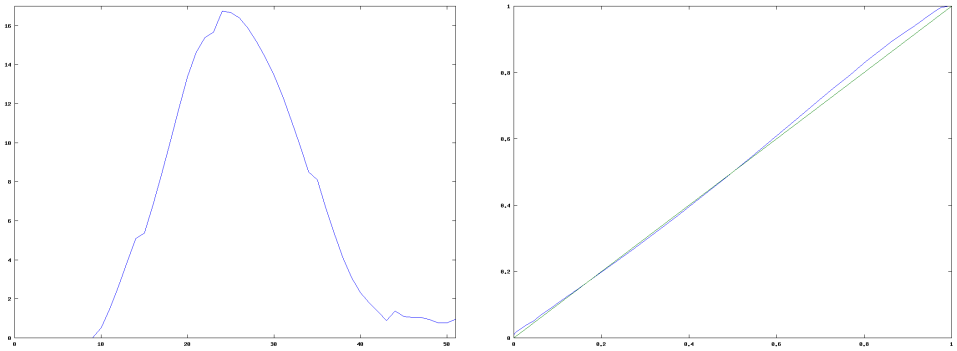


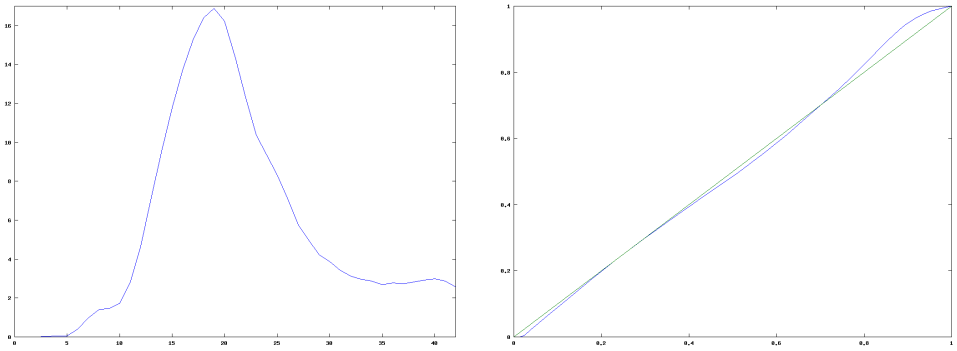
Abbildung B.4: Datenpeak einer IMS-Messung

$$\sigma = 6.9366, \quad \mu_t = 26.8028, \quad \lambda = 122.5430, \quad \alpha = 18.6032, \quad \mu_r = 3.3925$$



(a) Querschnitt in Dirftzeit t (b) QQ-Plot

Abbildung B.5: QQ-Plot zur Überprüfung der Daten mit der Normalverteilung



(a) Querschnitt in Retentionszeit r (b) QQ-Plot

Abbildung B.6: QQ-Plot zur Überprüfung der Daten mit der inversen Normalverteilung

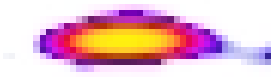
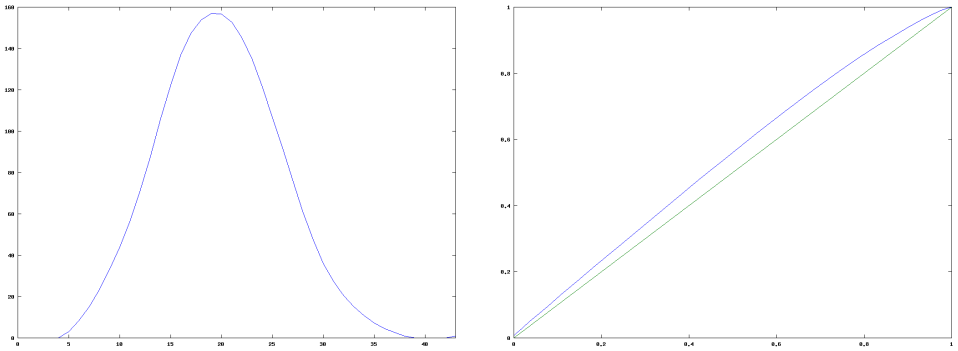


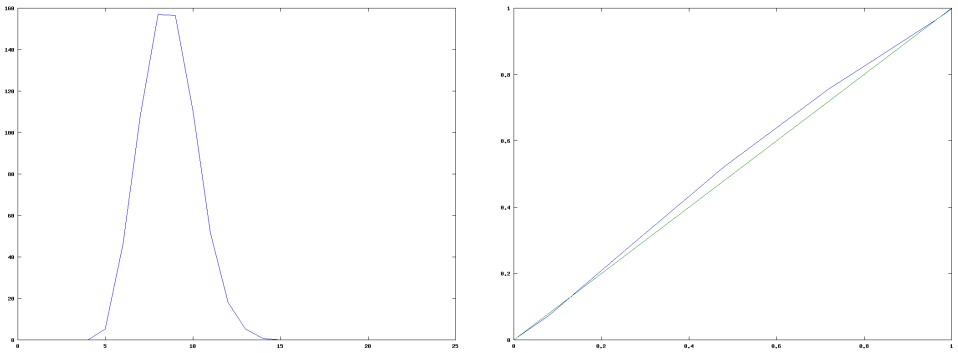
Abbildung B.7: Datenpeak einer IMS-Messung

$$\sigma = 5.5949, \quad \mu_t = 19.8146, \quad \lambda = 374.8340, \quad \alpha = 9.5723, \quad \mu_r = -1.0311$$



(a) Querschnitt in Dirftzeit t (b) QQ-Plot

Abbildung B.8: QQ-Plot zur Überprüfung der Daten mit der Normalverteilung



(a) Querschnitt in Retentionszeit r (b) QQ-Plot

Abbildung B.9: QQ-Plot zur Überprüfung der Daten mit der inversen Normalverteilung

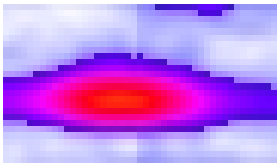
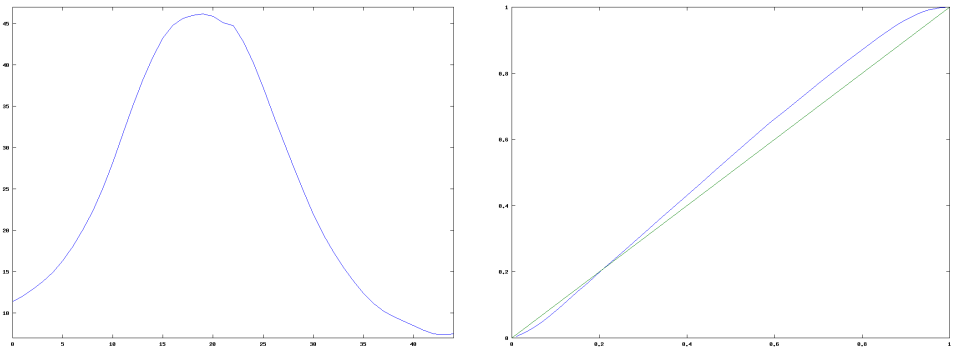


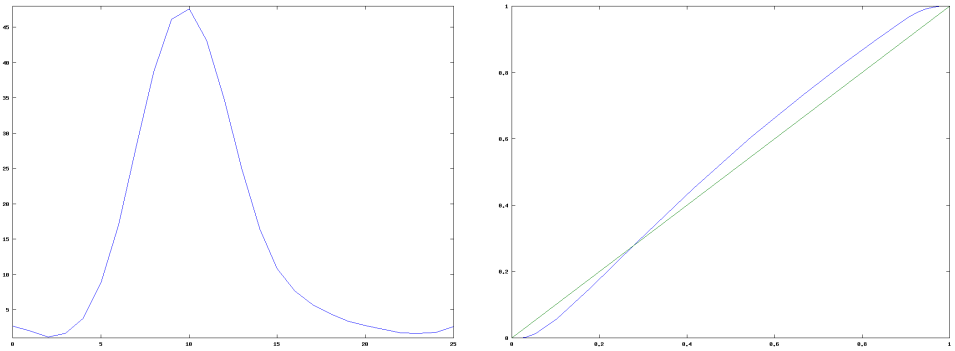
Abbildung B.10: Datenpeak einer IMS-Messung

$$\sigma = 8.3971, \quad \mu_t = 19.2208, \quad \lambda = 220.4680, \quad \alpha = 11.4740, \quad \mu_r = -1.3739$$



(a) Querschnitt in Dirftzeit t (b) QQ-Plot

Abbildung B.11: QQ-Plot zur Überprüfung der Daten mit der Normalverteilung



(a) Querschnitt in Retentionszeit r (b) QQ-Plot

Abbildung B.12: QQ-Plot zur Überprüfung der Daten mit der inversen Normalverteilung

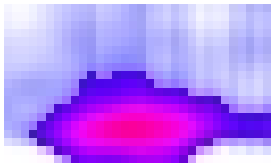
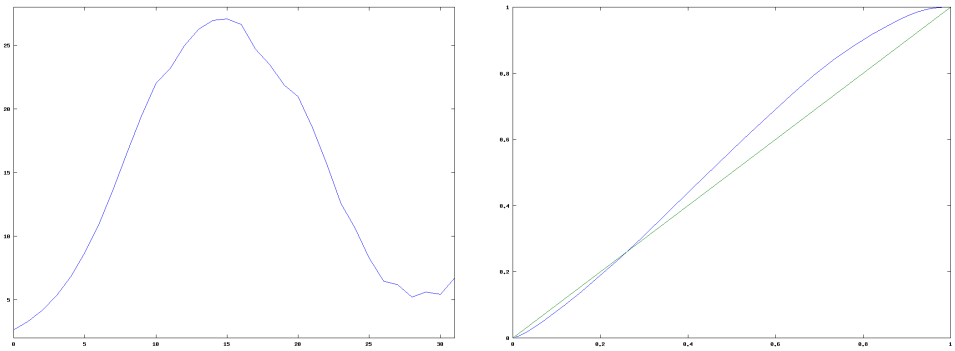


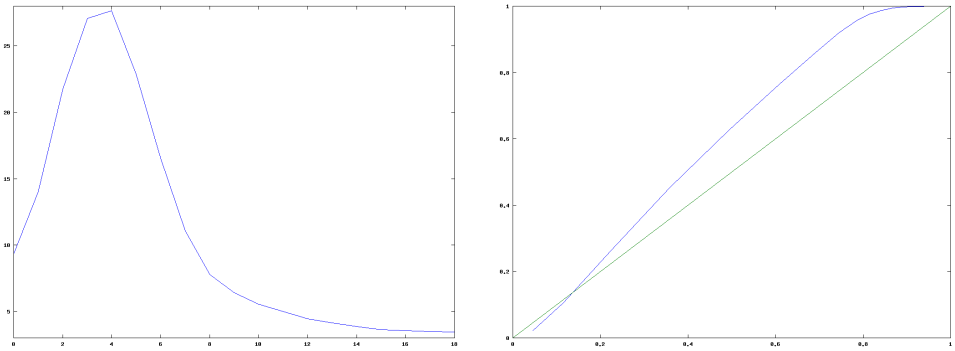
Abbildung B.13: Datenpeak einer IMS-Messung

$$\sigma = 5.3103, \quad \mu_t = 15.1553, \quad \lambda = 104.0300, \quad \alpha = 8.1931, \quad \mu_r = -4.1252$$



(a) Querschnitt in Dirftzeit t (b) QQ-Plot

Abbildung B.14: QQ-Plot zur Überprüfung der Daten mit der Normalverteilung



(a) Querschnitt in Retentionszeit r (b) QQ-Plot

Abbildung B.15: QQ-Plot zur Überprüfung der Daten mit der inversen Normalverteilung

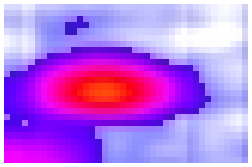
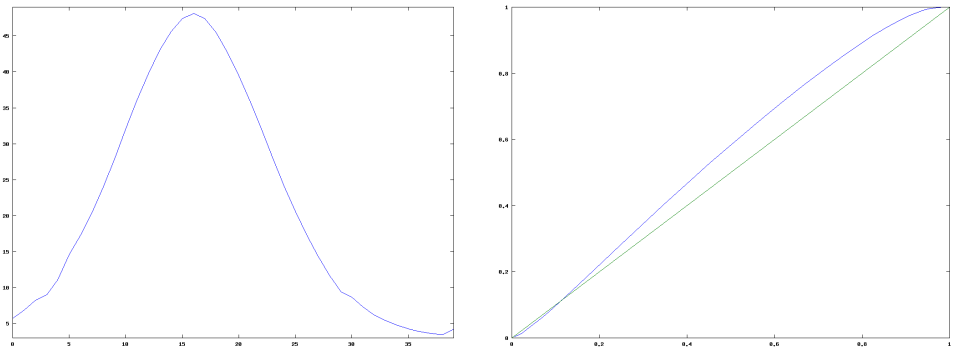


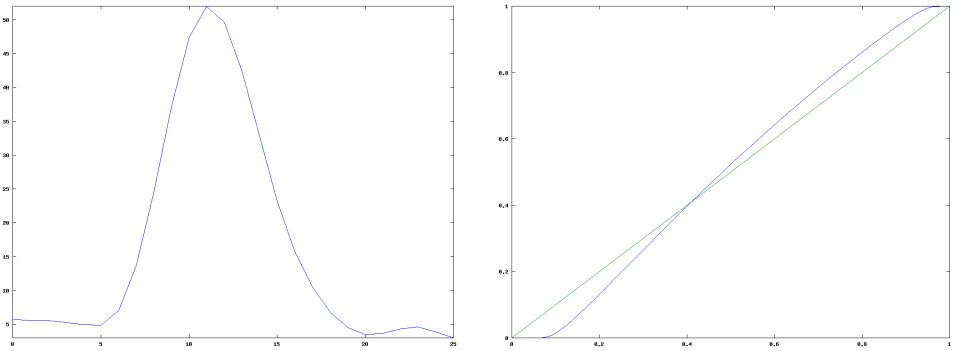
Abbildung B.16: Datenpeak einer IMS-Messung

$$\sigma = 6.2026, \quad \mu_t = 15.9943, \quad \lambda = 254.3390, \quad \alpha = 12.2503, \quad \mu_r = -0.6645$$



(a) Querschnitt in Dirftzeit t (b) QQ-Plot

Abbildung B.17: QQ-Plot zur Überprüfung der Daten mit der Normalverteilung



(a) Querschnitt in Retentionszeit r (b) QQ-Plot

Abbildung B.18: QQ-Plot zur Überprüfung der Daten mit der inversen Normalverteilung

Abbildungsverzeichnis

Literaturverzeichnis

[1] Beucher, S. und C. Lantuéjoul: *Use of watersheds in contour detection.* Proceedings of IEEE International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation, 1979.

[2] Bilmes, J.: *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* Technischer Bericht, 1998.

[3] Blum, M., R.W. Floyd, V.R. Pratt, R.L. Rivest und R.E. Tarjan: *Time Bounds for Selection.* J. Comput. Syst. Sci., 7(4):448–461, 1973.

[4] Bödecker, B.: *Entwicklung eines Verfahrens zur Klassifikation von Ionenmobilitätsspektrometerdaten.* Diplomarbeit, TU Dortmund, 2007.

[5] Cheng, R.C.H. und N.A.K. Amin: *Maximum Likelihood Estimation of Parameters in the Inverse Gaussian Distribution, with Unknown Origin.* American Statistical Association and American Society for Quality, 1981.

[6] Chhikara, R.S. und J.L. Folks: *The inverse gaussian distribution: theory, methodology, and applications.* Marcel Dekker, Inc., New York, NY, USA, 1989.

[7] Gan, G., C. Ma und J. Wu: *Data clustering. Theory, algorithms, and applications.* ASA-SIAM Series on Statistics and Applied Probability 20. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). xxii, 466 p., 2007.

[8] Gupta, R.: *Maximum likelihood estimate of the parameters of a truncated inverse gaussian distribution.* Metrika, 20(1):51–53, December 1973.

[9] Hogg, R.V.: *Introduction to mathematical statistics.* Macmillan, 1970.

[10] Ruzsányi, V.: *Analyse flüchtiger Metaboliten von der Ausatemluft mittels Ionenmobilitätsspektrometer.* Doktorarbeit, TU Dortmund, 2005.

[11] Tjalling, J.Y.: *Historical Development of the Newton-Raphson Method.* Society for Industrial and Applied Mathematics, 37(4):531–551, 1995.

[12] Weule, J.: *Iteration nichtlinearer Gauß-Filter in der Bildverarbeitung.* Doktorarbeit, Uni Düsseldorf, 1994.